ENHANCING DIAGNOSTIC ACCURACY WITH ENSEMBLE TECHNIQUES: DETECTING COVID-19 AND PNEUMONIA ON CHEST X-RAY IMAGES

Fatma A. Mostafa[1], Lamiaa A. Elrefaei[1], Mostafa M. Fouda[2] and Aya Hossam[1]

(Received: 27-May-2024, Revised: 17-Jul.-2024, Accepted: 12-Aug.-2024)

ABSTRACT

Lung diseases such as COVID-19 and pneumonia can lead to severe complications, including breathing difficulties, decreased lung function and respiratory failure, which can be life-threatening if not promptly treated. Chest X-ray imaging techniques have proven to be quick, effective and cost-efficient in diagnosing and monitoring these diseases. Additionally, artificial intelligence, particularly through deep learning and machine learning, has shown promising results in detecting various lung diseases, including COVID-19 and pneumonia. This technology's ability to analyze large datasets rapidly has contributed to reducing the spread of these diseases and has significantly advanced biomedical research in various medical disciplines. In this research paper, we introduced various advanced ensemble techniques as bagging, boosting, stacking and blending with different algorithms, to enhance the performance of our classification models in detecting coronavirus and pneumonia. We specifically focused on combining convolutional neural network (CNN) and vision transformer (ViT) models to create powerful ensemble models. Our objective was to determine the most accurate ensemble technique for diagnosing lung diseases. We assessed their ability to correctly classify chest X-ray images as either COVID-19, pneumonia or normal. The CatBoost model achieved the highest accuracy, F1-score and ROC-AUC score of 99.753%, 99.51% and 99.99%, respectively using the COVID-19 Radiography dataset. The bagging ensemble model achieved the highest accuracy, F1-score and ROC-AUC score of 95.08%, 95.2% and 99.69%, respectively using COVIDx CXR-4. The results indicate that the advanced ensemble techniques can significantly improve the performance of machine-learning models.

KEYWORDS

COVID-19, CXR, DenseNet-169, ViT-l32, Stacking, Boosting, Bagging, Blending.

1. INTRODUCTION

Lung diseases are a group of conditions that affect the health of the respiratory system and include many different diseases, such as pneumonia, COVID-19, pulmonary fibrosis and asthma [1]. One of the diseases that currently poses a major challenge is the new Coronavirus (COVID-19), where the number of deaths exceeded 7 million cases until April 2024 [2]. Coronavirus cases are currently a common lung disease. The new coronavirus mutates rapidly, leading to an increase in infection cases, with the number of infections reaching more than 700 million cases [2]. Pneumonia is not less dangerous than COVID-19. Therefore, early detection of these diseases reduces their risk [3]. One of the methods that is widely used in the detection of lung diseases is X-rays, as it is fast and inexpensive compared to other methods [4]. While X-rays are valuable in the detection of lung diseases, they may pose potential health risks with repeated exposure to ionizing radiation. In addition, while X-rays provide valuable information about the structure of the lungs, they may not always provide detailed insights into specific lung conditions, such as distinguishing between different types of pneumonia or identifying lung cancer at an early stage. In such cases, additional imaging techniques, such as computed tomography (CT), magnetic resonance imaging (MRI) or diagnostic tests, may be necessary for a comprehensive evaluation.

Artificial intelligence (AI) has revolutionized the medical field by searching medical data and revealing insights to enhance patient experiences and health outcomes [5]. It does this by utilizing machine-learning and deep-learning models. AI is frequently used in medical-imaging settings, analyzing CT scans, X-rays, MRIs and other images to look for lesions or other findings that a human radiologist might overlook, in addition to dealing with a huge volume of data quickly. Chest X-rays

1. F. A. Mostafa, L. A. Elrefaei and A. Hossam are with Department of Electrical Engineering, Faculty of Engineering at Shoubra, Benha University, Cairo, Egypt. Emails: fatma.mostafa13@feng.bu.edu.eg, lamia.alrefaai@feng.bu.edu.eg and aya.ahmed@feng.bu.edu.eg

2. Mostafa M. Fouda is with Department of Electrical and Computer Engineering, Idaho State University, Pocatello, ID, USA. Email: mfouda@ieee.org

(CXR) help detect lung abnormalities early and are also less expensive than other tests [6].

There are many types of deep-learning algorithms, such as CNNs, Multilayer Perceptrons (MLPs), Recurrent Neural Networks (RNNs) and Autoencoders, which are mainly used for image classification and object detection [7]-[9].

Ensemble methods are techniques the goal of which is to improve the performance of machine-learning models by creating one more reliable model by combining multiple models. Through the use of these techniques, a more reliable and more accurate model is produced by merging the predictions of several separate models. The main types of ensemble techniques are: bagging, boosting, stacking and blending techniques [10]. The bagging ensemble technique is mainly applied in classification which can increase accuracy and eliminate overfitting [11]. Boosting ensemble technique combines several weak learners to form a strong one; it takes many algorithms including gradient boosting, adaptive boosting (AdaBoost), light gradient-boosting machine (LightGBM) and extreme gradient boosting (XGBoost) [12]. The stacking ensemble technique uses predictions for multiple models to build a new strong model [13]. The blending ensemble technique also combines predictions from multiple models as stacking and can improve the overall performance of the model [14]-[15].

The integration of AI and deep-learning algorithms with ensemble methods holds significant promise for advancing the capabilities of medical imaging, leading to more accurate and more efficient detection and diagnosis of lung diseases and other medical conditions. We must take several considerations associated with the integration of AI and deep-learning algorithms with ensemble methods in medical imaging as data-management considerations, including storage, retrieval and processing, as well as algorithmic methods for disease classification and segmentation to ensure the successful implementation of these technologies in clinical practice.

The key contributions of this paper are:

1. Involving a comprehensive evaluation of the optimized CNN and ViT models.
2. Optimizing CNN and ViT hyper-parameters: The paper focuses on optimizing the hyper-parameters of Convolutional Neural Network (CNN) and Vision Transformer (ViT) models to reduce model losses and achieve the best accuracy in diagnosing lung diseases, particularly pneumonia and COVID-19.
3. Application of different ensemble techniques between CNN and ViT: The study explores the application of different ensemble techniques between CNN and ViT models. This involves leveraging ensemble methods to combine the strengths of these two architectures for improved diagnostic accuracy in lung-disease classification.
4. Comparison between the types of advanced ensemble techniques and fusion: The paper makes a comparison between different types of advanced ensemble techniques and fusion methods. This comparison provides insights into the effectiveness of various ensemble approaches in enhancing the accuracy of lung-disease diagnosis.
5. Use of the upgraded ensemble model to classify and recognize lung diseases: The upgraded ensemble model is utilized to classify and recognize lung diseases, such as pneumonia and COVID-19. This involves leveraging the optimized CNN and ViT models, along with ensemble techniques, to achieve accurate classification and recognition of lung diseases based on medical-imaging data.

This paper is organized as follows: Section 2 highlights related work. Next, Section 3 discusses the proposed work, Section 4 discusses the experimental results and lastly, conclusions and future-research directions are highlighted in Section 5.

2. Related Work

We divided the previous studies according to the types of ensemble techniques that can be used to diagnose lung diseases based on X-ray images, including methods, such as bagging, stacking, boosting, blending and weighted-average techniques.

Hasan et al. [16] created a model for automatically detecting pneumonia using CXR images. The weighted-average ensemble model was applied to three models; namely, VGG16, MobileNetV2 and DenseNet169. The ensemble model achieved an accuracy of 92%. Tang et al. [17] also presented a weighted-average ensemble approach to detect COVID-19 with an accuracy of 95%. They used COVIDx dataset to evaluate their experiment. They used COVID-Net as the candidate to generate multiple model snapshots in terms of its promising performance for its CXR image-based COVID-19 case detection.

Govindarajan et al. [18] presented an Extreme Gradient Boosting (XGBoost) classifier that can detect tuberculosis (TB) disease using CXR images from NIAID TB Portals repository. All images were 401 at a resolution of 1024×1024 pixels. The accuracy resulting from this model was 93%. Also, Kalaivani et al. [19] presented an ensemble boosted model using CNN and four different classifiers (decision tree, random forest, AdaBoost and support vector machine) to detect COVID-19. The suggested model achieved an accuracy of 99.35%. The images used were 5178 abnormal CXR images and 4310 normal CXR images collected from different sources.

Soundrapandiyan et al. [20] introduced a stacked ensemble model for detecting coronavirus (COVID-19) from chest X-ray images. The stacked ensemble technique between the models ResNet50, VGG19, Xception and DarkNet19 is named WavStaCovNet-19. The model achieved an accuracy of 94.24% on 4 classes (COVID-19, viral pneumonia, bacterial pneumonia and normal). They used two datasets, the COVID-19 image data-collection repository and chest X-ray images for normal cases and pneumonia [21]. Huang et al. [22] also presented a stacking ensemble model on the classification of multiple chest diseases including COVID-19 using the COVID-19 Radiography dataset. The model achieved an accuracy of 99.21%, Precision of 99.23%, Recall of 99.25%, F1-score of 99.20% and (area under the curve) AUC of 99.51% on the chest X-ray dataset. They verified that the combined model had better performance than individual pre-trained models. Six EfficientNetV2 models including EfficientNetV2-B0, EfficientNetV2- B1, EfficientNetV2-B2, EfficientNetV2-B3, EfficientNetV2-S and EfficientNetV2-M were stacked.

EROL et al. [23] made a comparison between three types of ensemble techniques; namely, bagging, AdaBoost and random forest, to detect COVID-19. Adaboost classifier achieved the highest accuracy of 97.25%. Bagging and random-forest classifiers achieved an accuracy of 96.69% and 96.89%, respectively. The BIGDATA-COVID19 dataset was used that includes age, sex and routine blood-test results of 1218 patients.

Banerjee et al. [24] presented the blending ensemble technique of DenseNet-201 snapshots, providing a variety of information regarding the features that were taken out of CXRs to detect COVID-19. To merge the decision scores, they employed the decision-level fusion approach, which involves a Random Forest (RF) meta-learner and the blending method. On the large COVID-X dataset, the model achieved an accuracy score of 94.55% and on the smaller dataset by Chowdhury et al., the model achieved an accuracy score of 98.13%.

There are some common difficulties/issues in the papers that we presented, because of which the reliability of the proposed deep-learning models can be questioned as data imbalance, image-size handling, dataset availability and high correlation of errors when employing ensemble techniques. The potential ways to overcome these issues include further experimentation, data augmentation, ensemble-model refinement and feature engineering to extract more relevant features from the CXR images.

3. Proposed Work

3.1 Dataset

The COVID-19 radiography database of chest X-ray images is one of the available datasets utilized to develop and evaluate deep-learning models for the detection and classification of lung diseases, particularly COVID-19, as shown in Figure 1. The database consists of a collection of images from multiple sources, such as the COVID-19 Image Dataset, the COVID-19 Database of the Italian Society of Medical and Interventional Radiology (SIRM) and images from several different publications [25]-[26]. It includes 3616 COVID-19-positive cases, 10,192 Normal, 6012 Lung Opacity (Non-COVID lung infection) and 1345 Viral Pneumonia images in PNG format [27]. All images have the same resolution of 299×299. The dataset was split into training, validation and test sets. The training set was used to train the model, the validation set was used to validate and tune the hyper-parameters of the model and the test set was used to evaluate the overall performance of the model. The training set included 1075 normal images, 1075 images of COVID-19 and 1075 images of pneumonia. The validation set and test set included 135 normal images, 135 COVID-19 images and 135 pneumonia images.

The COVIDx CXR-4 dataset is an open-source benchmark dataset that combines 5 different publicly available datasets which include: COVID-19 Image data collection, COVID-19 Chest X-Ray Dataset Initiative, COVID-19 radiography database, RSNA Pneumonia Detection challenge dataset and ActualMed COVID-19 Chest X-Ray Dataset Initiative [28]. The dataset includes 84,818 images from 45,342 patients in PNG format with the same resolution of 1024×1024. The dataset is classified into positive and negative COVID-19 samples. We split the data into training, validation and test sets. The training set includes 9600 images, the validation set includes 1200 images and the test set includes 1200 images. Each set contains 3 classes: COVID-19, normal and pneumonia.

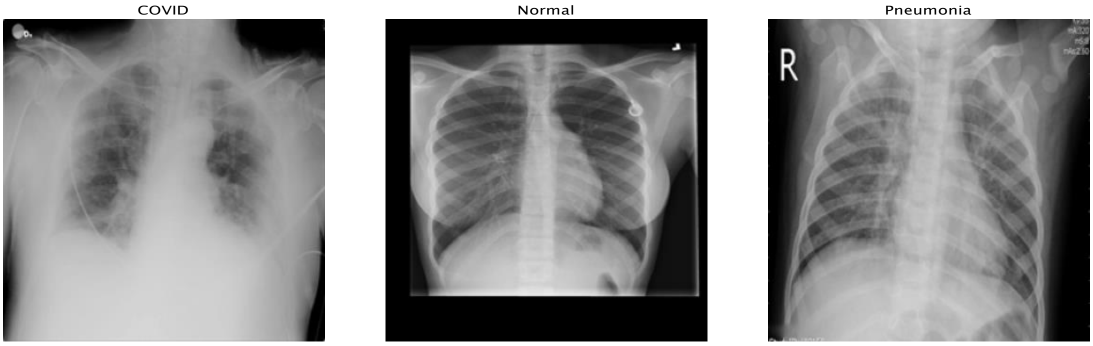


Figure 1. A batch of images from the training dataset.

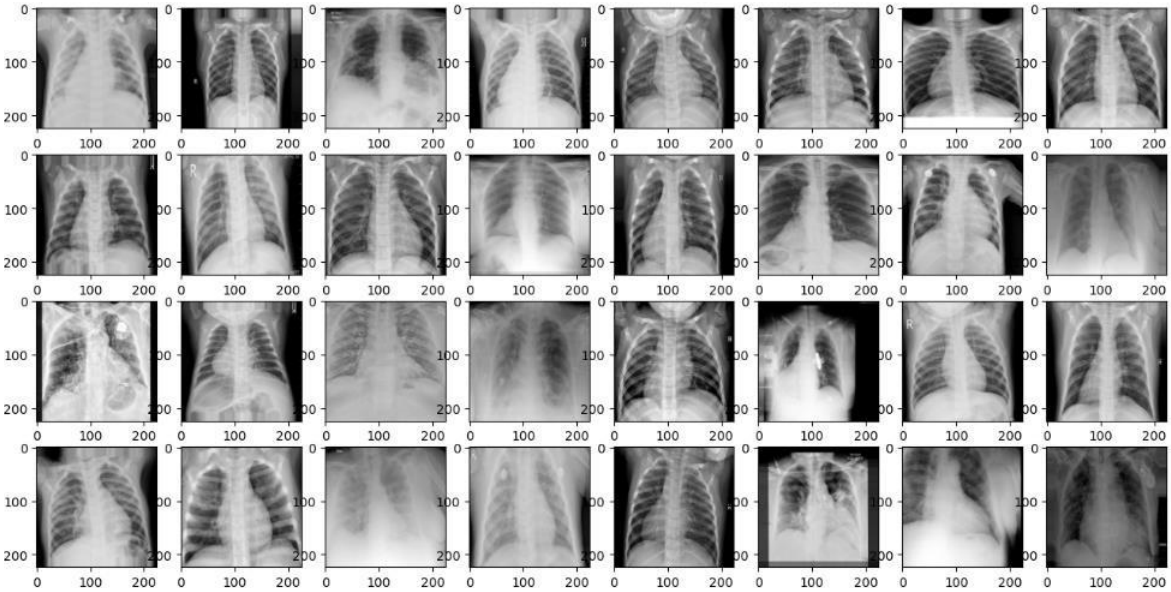


Figure 2. A batch of training images after data augmentation.

3.2 Data Preparation

A data-augmentation technique is used to increase data diversity, improve generalization and reduce overfitting in deep-learning models. Random transformations, such as rescale, shear, zoom and horizontal flip, were applied to the training images as shown in Figure 2. Both the shear-range and zoom-range parameters were set to 0.2 in the ImageDataGenerator class from the Keras library. Images have been resized to dimensions of 224×224.

3.3 Model

This research is a continuation of previous work where we used both DenseNet-169 and vision transformer (ViT-l32) models to detect COVID-19 and pneumonia lung diseases. The results were as follows: the accuracy of both models was 92.31% and 92.56%, respectively. To improve the performance of the two models, they were combined using two different types of ensemble techniques: decision-level fusion and feature-level fusion. Indeed, the accuracy improved to 93.3% and 94.54%, respectively [3]. Now we use other ensemble methods on the same two models and the same dataset to achieve a better-performance model.

3.3.1 DenseNet-169

A CNN architecture called DenseNet-169 was created specifically for image-classification applications. It belongs to the DenseNet model family, which is famous for having dense connectivity between layers. It is composed of several dense blocks as shown in Figure 3, each dense block consisting of several convolutional layers. The primary advantage of DenseNet is its dense connection architecture, where each layer is feed-forward connected to every other layer. Enhanced information flow and feature reuse across the network are made possible by this dense connectivity. To reduce the number of parameters in the network and reduce the spatial dimensions of feature maps, transition layers between dense blocks are used. Transition layers include a combination of convolutional layers, pooling layers, batch-normalization and nonlinear-activation functions. When compared to the DenseNet-121 model, the DenseNet-169 is larger and more accurate. It is about 55MB in size and contains about 169 layers [29]. The increased depth allows DenseNet-169 to capture more complex features and potentially achieve improved accuracy in image-classification tasks [30]. The DenseNet-169 model takes an image with a size of 224×224 pixels as input [31]. Compared with other CNN architectures, it is relatively low in parameters.

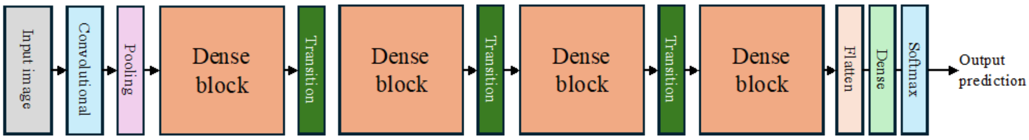


Figure 3. DenseNet-169 architecture.

3.3.2 Vision Transformer

The vision transformer (ViT) has shown promising effectiveness for global feature extraction in many tasks of computer vision, such as image recognition, image classification, object detection and image segmentation [32]. It has gained attention due to its ability to capture long-term dependencies and its generalizability across different data modalities. The ViT model is an effective tool for image classification, because it uses self-attention processes to obtain global information from an image. By capturing global and local representations from shallow and deep layers, the ViT model differs from traditional CNN, which concentrates mostly on local features using convolutional filters [33], but ViT processes images using patches and the self-attention method by converting input images into a sequence of tokens. The input image is split up into fixed-size patches in the first stage. Every patch is viewed as a token and is put through a linear embedding process. Subsequently, position embeddings are appended to the patch embeddings to furnish spatial details regarding the patches' placement within the image. The sequences of patch embeddings and position embeddings are then put into a typical transformer encoder. The transformer encoder is composed of feed-forward neural networks and numerous layers of self-attention. By creating attention maps from the provided embedded visual tokens, the multi-head self-attention (MSA) enables the model to focus on several regions of the input image concurrently. Batch normalization is used to improve training stability and reduce training time. The residual connections improve the overall performance of the network [34]. Figure 4 illustrates the ViT architecture.

There are several variations related to ViT as ViT-l16, ViT-l32, ViT-b16, ViT-b32 and data-efficient image transformers (DeiT). The ViT-l32 is considered more significant and more powerful than some of the other variants [35]-[36] due to its ability to achieve superior results in image-recognition tasks, so ViT-l32 was chosen. l means 'large' and 32 refers to batch size.

3.3.3 The Used Ensemble Techniques Background

The ensemble techniques refer to the use of several base models and combining their predictions to improve the overall performance and accuracy of the system [37]. Instead of relying on a single model, ensemble techniques leverage the diversity and collective intelligence of multiple models to produce one optimal predictive model. It can reduce bias, variance and overfitting [38]-[39]. There are several types of ensemble techniques as bagging, stacking, blending and boosting [40].

In this paper, we applied a comprehensive range of ensemble techniques, including Bagging, Gradient Boosting, AdaBoost, XGBoost, LightGBM, CatBoost, Stacking and Blending.

Bagging, also known as bootstrap aggregation, is an ensemble method that involves training multiple models independently on random sub-sets of the data [41]-[43]. The bagging involves the following steps: (1) Generating predictions from the base models using the holdout set (test set). (2) Combining the predictions as features for the bagging model using concatenate function. (3) Training a bagging model using the combined features and true labels. (4) Generating predictions from the trained bagging model on the combined features. (5) Evaluating the bagged model performance using the accuracy_score function with the combined predictions.

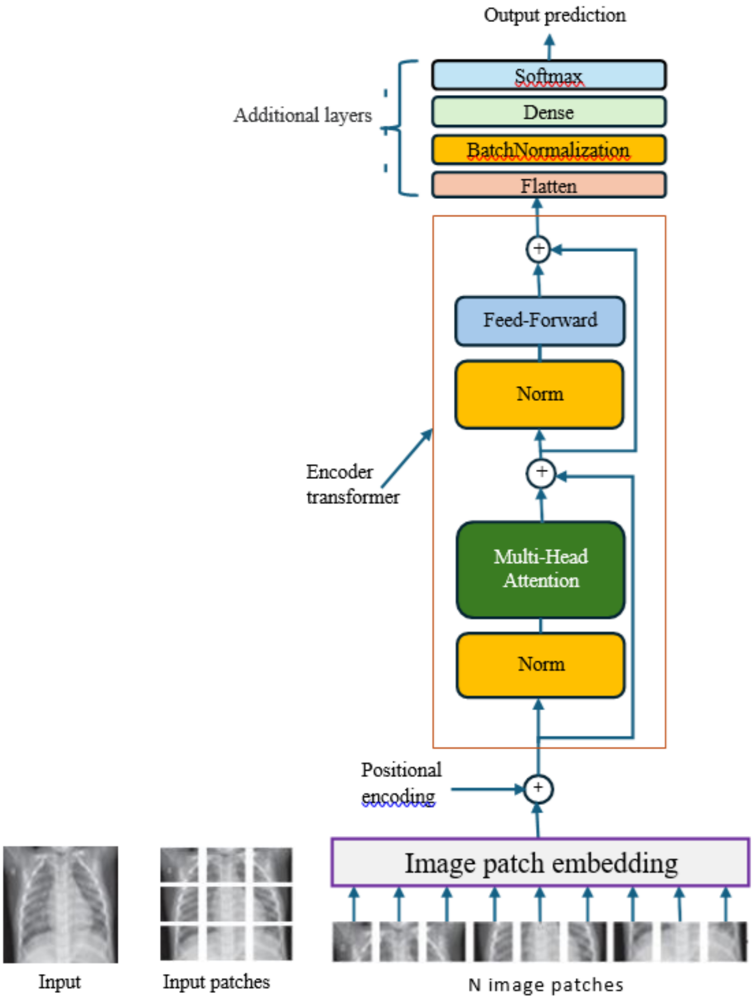


Figure 4. ViT-l32 structure.

Stacking uses a meta-learning algorithm to learn how to best combine the predictions from two or more models [44]. It can combine the capabilities of a group of models that perform well on a classification or regression task and make predictions that perform better than any single model in the group [45]-[46]. The stacking involves the following steps: (1) Compiling the models and making predictions. (2) Combining the predictions from the base models to create meta-features. (3) Useing the combined meta-features as input features for the meta-learner. (4) Training the meta-learner model using the meta-features and the true labels. (5) Evaluating the performance of the stacked model using appropriate evaluation metrics. This implementation follows the stacking ensemble technique, where predictions from base models are combined using a meta-model to improve predictive performance. The use of a meta-model allows for the aggregation of predictions from diverse base models, aiming to reduce overfitting and improve generalization.

Boosting is designed to address potential limitations of individual models, such as generalizability and data bias. By combining the predictions of multiple models, ensemble methods aim to improve the robustness and generalization of machine-learning models by combining a group of weak learners into a strong learner. It trains the models sequentially, where each subsequent model focuses on correcting the mistakes made by the previous models, so it can minimize training errors [47]-[48]. There are many types of boosting algorithms, such as AdaBoost, CatBoost, XGBoost and LGBoost [49]-[50].

The short form for adaptive boosting (AdaBoost) is a powerful ensemble learning algorithm used in machine learning for addressing binary-classification challenges.

The catBoost is a high-performance open-source library for gradient boosting on decision trees, designed for use in classification, regression and ranking tasks.

The eXtreme Gradient Boosting (XGBoost) is an open-source software library that provides a regularizing gradient boosting framework for various programming languages, such as C++, Java and Python.

The light gradient boosting machine (LGBoost) is an open-source gradient boosting framework that is designed for efficiency, scalability and high performance in machine-learning tasks.

The boosting steps: (1) Training each model separately. (2) Generating predictions from each model. (3) Combining the predictions using the concatenate function. (4) Training a boosting model (GradientBoost- ingClassifier) using the stacked features. (5) Generating predictions from the trained boosting model on the stacked features. (6) Evaluate the performance of the boosted model using appropriate evaluation metrics.

Blending is an ensemble approach that can improve the model performance to be more accurate. It uses a particular method to merge predictions from various models contributing to the ensemble. The steps used in the blending process are: (1) Generating the predictions from the base models. (2) Building the model from the test set and the predictions. (3) The building model serves as the meta-model. (4) Generating predictions from the meta-model.

4. RESULTS AND DISCUSSION

The proposed method was implemented using Python 3.8 with additional libraries, such as Pandas, Tensor Flow and Keras. Windows 10 Operating System powered the System with configuration, Intel(R) Xeon(R) CPU E5-2687W v3 @ 3.10 GHz, NVIDIA GeForce GTX 970 GPU and 64 GB RAM.

The first dataset "COVID-19 Radiography dataset" obtained from Kaggle [27], was utilized to train and test DenseNet-169 and ViT-l32 models for multi-level classification aimed at detecting COVID-19 patients. The training and validation sets comprised 90% of the dataset, while the testing set utilized the remaining 10%, as outlined in Table 1. Python and machine-learning libraries were employed for implementation, with the Python programming language utilized to train and evaluate the proposed models, which were pre-trained using TensorFlow. The training data underwent modification through data-augmentation techniques, as illustrated in Figure 2. The second dataset COVIDx CXR-4 was split into the training, validation and test set in the ratio 8:1:1, as shown in Table 2.

The pre-trained DenseNet 169 model was trained on the initialization weights illustrated in Table 3 using the first dataset and the Adam optimizer. Subsequently, the ViT-l32 was separately trained on the same dataset. Predictions were generated from the two models using the test set and combined using ensemble techniques, as shown in Figure 5.

Table 1. Class-wise distribution of CXR samples in the COVID-19 Radiography database.

Phase	COVID-19	Normal	Pneumonia	Total
Training	1075	1075	1075	3225
Validation	135	135	135	405
Test	135	135	135	405
Total	1345	1345	1345	4035

Table 2. Class-wise distribution of CXR samples in the COVIDx CXR-4 dataset.

Phase	COVID-19	Normal	Pneumonia	Total
Training	3200	3200	3200	9600
Validation	400	400	400	1200
Test	400	400	400	1200
Total	4000	4000	4000	12000

Table 3. Training parameters.

Training Parameters	**Values/Types**
Number of epochs	100
Batch size	32
Optimizer	Adam
Learning rate	0.001
Zoom and shear range	20%
Fill mode	Nearest
Rescale	1./255
Horizontal flip	True
Shuffle	True
Class mode	Categorical

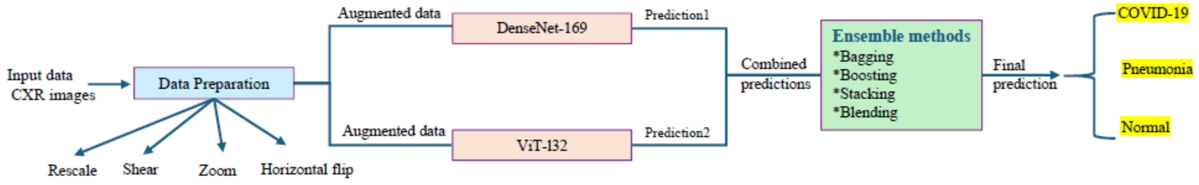


Figure 5. Proposed workflow model.

Ensemble techniques typically refer to Bagging (bootstrap-aggregating), Boosting or Stacking/Blending techniques to induce high variability among the base models. These techniques aim to combine the predictions from multiple models to improve predictive performance. To evaluate the bagged-model performance, we combined the predictions from the two models using a voting mechanism (majority voting). Finally, we evaluated the combined predictions and obtained an accuracy of 98.27% from the random-forest bagging model, as shown in Figure 6a.

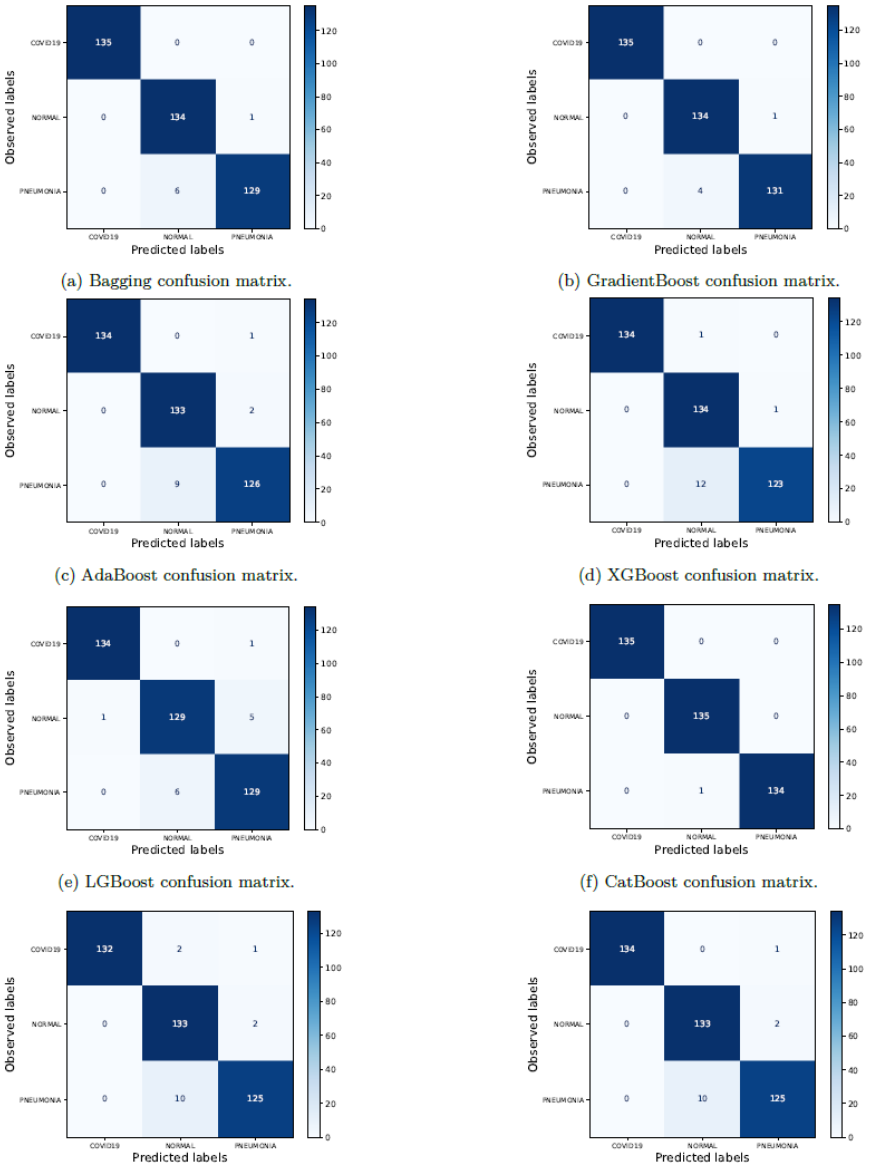
When implementing boosting ensemble techniques using the DenseNet169 and ViTl32 models, the weak learners (base models) are combined sequentially to form a strong learner (ensemble model). The gradient- boost, AdaBoost, XGBoost, LGBoost and CatBoost ensemble techniques achieved an accuracy of 98.765%, 97.04%, 96.54%, 97.79% and 99.753%, respectively as shown in Figure 6b, Figure 6c, Figure 6d, Figure 6e and Figure 6f.

Stacking, also known as stacked generalization, allows a training algorithm to ensemble several similar learning-algorithm predictions. A stacking model is implemented using a holdout set to generate predictions from base models (CNN and ViTl32 models). These predictions are then concatenated to create a stacked dataset. The true labels for the holdout set are one-hot encoded using the OneHotEncoder class. A meta-model for multi-class classification is defined and trained using the stacked dataset. The meta-model consists of three dense layers with ReLU and Softmax activations. The accuracy and F1-score of the stacked ensemble model are then calculated. The accuracy result was 96.296%, as shown in Figure 6g.

The blending ensemble technique combines the predictions of several base models to enhance overall predictive performance, minimize overfitting and leverage the advantages of different methods. It achieved an accuracy of 96.79%, as shown in Figure 6h.

So, from the previous results, CatBoost achieved the highest accuracy using one dataset. The optimization for the ensemble implementation involved: 1) A combination of hyper-parameter tuning, such as the learning rate, depth and regularization parameters. 2) Model selection. 3) Validation and benchmarking against other algorithms to achieve the highest accuracy for the specific dataset.

Other performance metrics, such as precision, recall, f1-score, sensitivity, specificity and ROC-AUC score, can be calculated using the COVID-19 Radiography dataset and the results are shown in Table 4.



(a) Bagging confusion matrix.

(b) GradientBoost confusion matrix.

(c) AdaBoost confusion matrix.

(d) XGBoost confusion matrix.

(e) LGBoost confusion matrix.

(f) CatBoost confusion matrix.

(g) Stacking confusion matrix.

(h) Blending confusion matrix.

Figure 6. Confusion matrices for all ensemble models using COVID-19 Radiography dataset.

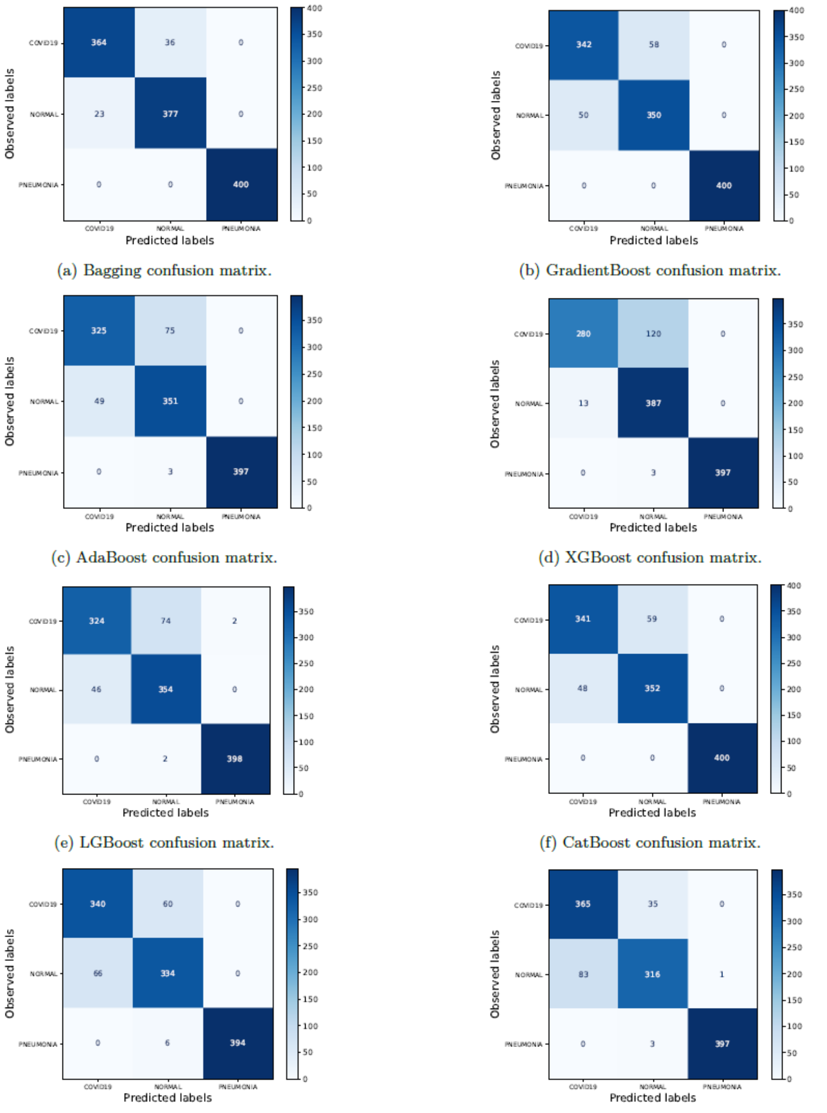
Table 4. Performance of the proposed ensemble models using the COVID-19 Radiography dataset.

Ensemble Techniques	Precision	Recall	F1-score	Sensitivity	Specificity	ROC-AUC score
Bagging	98.53%	98.52%	98.52%	100%	100%	99.69%
GradientBoost	98.55%	98.52%	98.52%	100%	100%	99.94%
AdaBoost	97.79%	97.78%	97.78%	100%	100%	99.54%
XGBoost	97%	96.79%	96.80%	100%	99.26%	99%
LGBoost	97.55%	97.53%	97.53%	100%	100%	99.84%
CatBoost	99.51%	99.51%	99.51%	100%	100%	99.99%
Stacking	99.25%	100%	96.06%	100%	99.25%	99.69%
Blending	99.25%	100%	96.06%	100%	99.25%	99.72%

When using more than one dataset (COVIDx CXR-4) the results became as follows: RF bagging model achieved an accuracy of 95.08%, as shown in Figure 7a. The gradient-boost, AdaBoost, XGBoost, LGBoost and CatBoost ensemble techniques achieved an accuracy of 91%, 89.42%, 88.67%, 89.67% and 91.08%, respectively, as shown in Figure 7b, Figure 7c, Figure 7d, Figure 7e and Figure 7f. The stacking and blending models achieved an accuracy of 89% and 89.83%, as shown in Figure 7g and Figure 7h, respectively. From these results, the RF bagging ensemble model achieved the highest accuracy of 95.42%. The other performance metrics using COVIDx CXR-4 can be calculated and the results are shown in Table 5.

Table 5. Performance of the proposed ensemble models using COVIDx CXR-4.

Ensemble Techniques	Precision	Recall	F1-score	Sensitivity	Specificity	ROC-AUC score
Bagging	95.17%	95.17%	95.2%	92%	93.5%	99.69%
GradientBoost	92.45%	92.42%	92.41%	86.75%	90.5%	98.6%
AdaBoost	91.2%	91.1%	91.2%	83.5%	90.68%	96.37%
XGBoost	91.15%	91.17%	90.1%	95.25%	75.25%	96.45%
LGBoost	91.67%	91.67%	91.65%	86.25%	89.89%	98.53%
CatBoost	93.1%	93.1%	93.1%	89%	90.25%	98.79%
Stacking	91.94%	91.92%	91.92%	86.5%	89.5%	98.35%
Blending	91.3%	91.25%	91.25%	89%	85%	98.34%



(a) Bagging confusion matrix.

(b) GradientBoost confusion matrix.

(c) AdaBoost confusion matrix.

(d) XGBoost confusion matrix.

(e) LGBoost confusion matrix.

(f) CatBoost confusion matrix.

(g) Stacking confusion matrix.

(h) Blending confusion matrix.

Figure 7. Confusion matrices for all ensemble models using COVIDx CXR-4 dataset.

Table 6. Comparison between the proposed model and previous ensemble studies.

Model	Dataset	Number of images	Accuracy
Weighted average ensemble model [16]	Chest X-ray dataset	The dataset is divided into 2 classes: Normal and Pneumonia. 4192 for the training set, 1040 for the validation set, and 624 for the test set.	92%
Weighted average ensemble model [17]	COVIDx dataset	The dataset is divided into 3 classes: Normal, Pneumonia, and COVID-19. 13898 for the training set and 1579 for the test set.	95%
XGBoost classifier [18]	Chest X-ray image database	The dataset is divided into 2 classes: Multi-Drug Resistant Tuberculosis (MDR–TB) and Drug Sensitive Tuberculosis (DS-TB). 199 DS-TB and 202 MDR-TB.	93%
Boosted model [19]	New dataset collected from different sources	The dataset is divided into 3 classes: COVID-19, Pneumonia, and Normal. 6755 for the training set, 1822 for the validation set, and 911 for the test set.	99.35%
Stacked model [20]	COVID-19 image data collection repository and chest X-ray images	The dataset is divided into 2 classes: Normal and Pneumonia. COVID-19 image data collection repository: 286 images for training and 192 for testing. chest X-ray images dataset: 3900 images for training and 624 for testing.	94.24%
Stacked model [22]	COVID-19 Radiography Dataset, Chest X-Ray Images (Pneumonia), and Tuberculosis Chest X-ray Dataset	The dataset is divided into 5 classes: COVID-19, Bacterial pneumonia, Tuberculosis, Viral pneumonia, and Normal. In training, validation, and test subsets: 900, 100, and 200, respectively.	99.21%
Adaboost classifier [23]	BIGDATA-COVID19	The dataset is divided into 2 classes: Severity and Dead. 3996 for the training set and 999 for the test set.	97.25%
Blending model [24]	(COVID-X and Chowdhury datasets	The dataset is divided into 3 classes: Normal, Pneumonia, and COVID-19. The COVID-X dataset: 11115 for the training set, 1579 for the validation set, and 2777 for the test set. The COVID dataset by Chowdhury et al.: 1807 for the training set, 321 for the validation set, and 777 for the test set.	98.13%
Decision-level fusion model [3]	COVID-19 Radiography database	The dataset is divided into 3 classes: Normal, Pneumonia, and COVID-19. 3225 images for the training set, 405 for the validation set, and 403 for the test set.	93.3%
Feature-level fusion model [3]	COVID-19 Radiography database	The dataset is divided into 3 classes: Normal, Pneumonia, and COVID-19. 3225 images for the training set, 405 for the validation set, and 403 for the test set.	94.54%
Proposed RF bagging model	COVID-19 Radiography dataset	3225 images for the training set, 405 for the validation set, and 405 for the test set.	98.27%
Proposed GradientBoost model	COVID-19 Radiography dataset	3225 images for the training set, 405 for the validation set, and 405 for the test set.	98.765%
Proposed AdaBoost model	COVID-19 Radiography dataset	3225 images for the training set, 405 for the validation set, and 405 for the test set.	97.04%
Proposed XGBoost model	COVID-19 Radiography dataset	3225 images for the training set, 405 for the validation set, and 405 for the test set.	96.54%
Proposed LGBoost model	COVID-19 Radiography dataset	3225 images for the training set, 405 for the validation set, and 405 for the test set.	96.79%
Proposed CatBoost model	COVID-19 Radiography dataset	3225 images for the training set, 405 for the validation set, and 405 for the test set.	99.753%
Proposed stacking model	COVID-19 Radiography dataset	3225 images for the training set, 405 for the validation set, and 405 for the test set.	96.3%
Proposed blending model	COVID-19 Radiography dataset	3225 images for the training set, 405 for the validation set, and 405 for the test set.	96.79%
Proposed RF bagging model	COVIDx CXR-4 dataset	9600 images for the training set, 1200 for the validation set, and 1200 for the test set.	95.08%
Proposed GradientBoost model	COVIDx CXR-4 dataset	9600 images for the training set, 1200 for the validation set, and 1200 for the test set.	91%
Proposed AdaBoost model	COVIDx CXR-4 dataset	9600 images for the training set, 1200 for the validation set, and 1200 for the test set.	89.42%
Proposed XGBoost model	COVIDx CXR-4 dataset	9600 images for the training set, 1200 for the validation set, and 1200 for the test set.	88.67%
Proposed LGBoost model	COVIDx CXR-4 dataset	9600 images for the training set, 1200 for the validation set, and 1200 for the test set.	89.67%
Proposed CatBoost model	COVIDx CXR-4 dataset	9600 images for the training set, 1200 for the validation set, and 1200 for the test set.	91.08%
Proposed stacking model	COVIDx CXR-4 dataset	9600 images for the training set, 1200 for the validation set, and 1200 for the test set.	89%
Proposed blending model	COVIDx CXR-4 dataset	9600 images for the training set, 1200 for the validation set, and 1200 for the test set.	89.83%

By looking at the results summarized in Figure 8, we find that the accuracy of the models decreased when using more than one dataset, which may be due to various factors, such as data complexity and diversity. As previously mentioned, this work is a continuation of our last work [3] to improve the model's performance. The results we obtained from the earlier work were as follows: decision-level fusion and feature-level fusion achieved an accuracy of 93.3% and 94.53%, respectively. However, the results from the advanced ensemble techniques reached 99.753% when using the same COVID-19 Radiograph dataset. Finally, the results confirm that the performance of these advanced ensemble models surpasses that of fusion models, as shown in Figure 8. Table 6 compares the previous studies and our proposed methods.

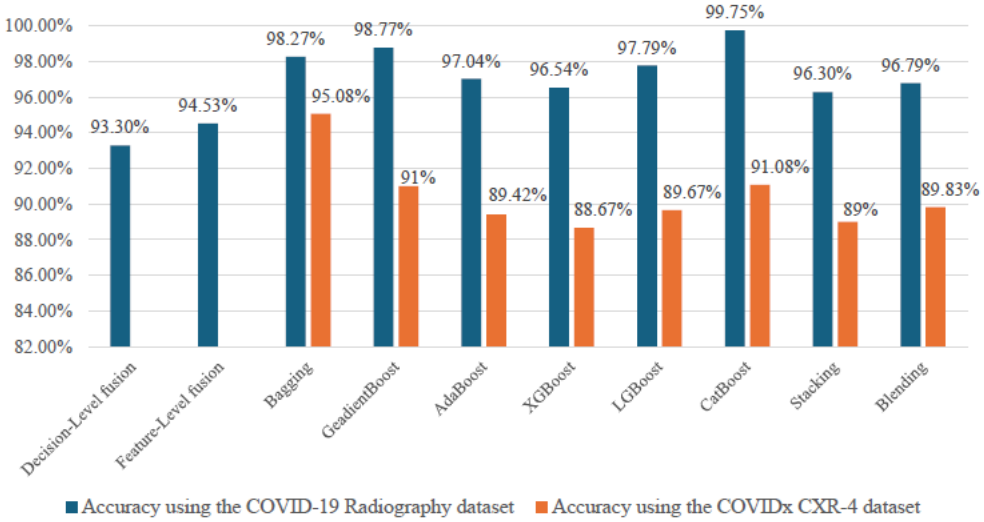


Figure 8. The accuracy of our proposed ensemble models.

5. Conclusions and Future Research Directions

This paper presented different types of advanced ensemble techniques to improve the model performance in diagnosing lung diseases using CXR images. We used pre-trained CNN models DenseNet-169 and ViT-l32. This work is related to a previous research paper, but we presented different methods that showed more accurate results. In previous work, we used simple ensemble techniques, such as feature-level fusion and decision-level fusion, achieving accuracy results as follows: 94.54% and 93.3%, respectively. However, using advanced ensemble techniques, we achieved a higher accuracy of fusion operations, reaching 99.753%. To aid in lung-disease prevention and early diagnosis, researchers continue to develop a variety of detection technologies and architectures by increasing the size and diversity of training datasets, but this can be costly and time-consuming. To avoid these issues, researchers are exploring techniques like data augmentation to address the challenge of limited datasets.

In our future work, we plan to use multiple datasets and explainable AI (XAI) models to enhance the accuracy and comprehensiveness of lung-disease diagnosis and classification. We also consider including multi-model data to expand the feature space and improve disease-classification accuracy, such as medical records and clinical metadata.

References

[1] F. A. Mostafa, L. A. Elrefaei, M. M. Fouda and A. Hossam, "A Survey on AI Techniques for Thoracic Diseases Diagnosis Using Medical Images," Diagnostics, vol. 12, no. 12, p. 3034, 2022.

[2] I. Team, "Coronavirus Cases," [Online]. Available: https://www.worldometers.info/ coronavirus/, 2024.

[3] F. A. Mostafa, L. A. Elrefaei, M. M. Fouda and A. Hossam, "Diagnosis of Lung Diseases from Chest X- Ray Images Using Different Fusion Techniques," Proc. of the 2023 11th Int. Conf. on Information and Communication Technology (ICoICT), pp. 429–435, Melaka, Malaysia, 2023.

[4] M. H. Saad, S. Hashima, W. Sayed, E. H. El-Shazly, A. H. Madian and M. M. Fouda, "Early Diagnosis of COVID-19 Images Using Optimal CNN Hyper-parameters," Diagnostics, vol. 13, no. 1, p. 76, 2023.

[5] Z. Ahmad, S. Rahim, M. Zubair and J. Abdul-Ghafar, "Artificial Intelligence (AI) in Medicine, Current Applications and Future Role with Special Emphasis on Its Potential and Promise in Pathology: Present and Future Impact, Obstacles Including Costs and Acceptance among Pathologists, Practical and Philosophical Considerations: A Comprehensive Review," Diagnostic pathology, vol. 16, pp. 1–16, 2021.

[6] H. Q. Nguyen et al., "VinDr-CXR: An Open Dataset of Chest X-Rays with Radiologist's Annotations," Scientific Data, vol. 9, no. 1, p. 429, 2022.

[7] A. Shrestha and A. Mahmood, "Review of Deep Learning Algorithms and Architectures," IEEE Access, vol. 7, pp. 53040–53065, 2019.

[8] Z. Yu, K. Wang, Z. Wan, S. Xie and Z. Lv, "Popular Deep Learning Algorithms for Disease Prediction: A Review," Cluster Computing, vol. 26, no. 2, pp. 1231–1251, 2023.

[9] A. A. Chowdhury, K. T. Hasan and K. K. S. Hoque, "Analysis and Prediction of COVID-19 Pandemic in Bangladesh by Using ANFIS and LSTM Network," Cognitive Computation, vol. 13, pp. 761-770, [Online], Available: https://link.springer.com/article/10.1007/s12559-021-09859-0, Apr. 2021.

[10] D. Kuzinkovas and S. Clement, "The Detection of COVID-19 in Chest X-Rays Using Ensemble CNN Techniques," Information, vol. 14, no. 7, p. 370, 2023.

[11] A. Zizaan and A. Idri, "Evaluating and Comparing Bagging and Boosting of Hybrid Learning for Breast Cancer Screening," Scientific African, vol. 23, p. e01989, DOI: 10.1016/j.sciaf.2023.e01989, 2024.

[12] U. Ahmed, J. C.-W. Lin and G. Srivastava, "Towards Early Diagnosis and Intervention: An Ensemble Voting Model for Precise Vital Sign Prediction in Respiratory Disease," IEEE Journal of Biomedical and Health Informatics, pp. 1-13, DOI: 10.1109/JBHI.2023.3270888, 2023.

[13] U. Bhimavarapu, N. Chintalapudi and G. Battineni, "Multi-classification of Lung Infections Using Improved Stacking Convolution Neural Network," Technologies, vol. 11, no. 5, p. 128, 2023.

[14] T. Mahesh et al., "Blended Ensemble Learning Prediction Model for Strengthening Diagnosis and Treatment of Chronic Diabetes Disease," Computational Intelligence and Neuroscience, vol. 2022, DOI: 10.1155/2022/4451792, 2022.

[15] T. Mahesh et al., "Early Predictive Model for Breast Cancer Classification Using Blended Ensemble Learning," Int. J. of System Assurance Eng. and Management, vol. 15, no. 1, pp. 188–197, 2024.

[16] R. Hasan, S. M. Azmat Ullah, A. Nandi and A. Taher, "Improving Pneumonia Diagnosis: A Deep Transfer Learning CNN Ensemble Approach for Accurate Chest X-ray Image Analysis," Proc. of the 2023 Int. Conf. on Information and Communication Technology for Sustainable Development (ICICT4SD), pp. 109–113, hal-04163800f, 2023.

[17] S. Tang, C. Wang, J. Nie, N. Kumar, Y. Zhang, Z. Xiong and A. Barnawi, "EDL-COVID: Ensemble Deep Learning for COVID-19 Case Detection from Chest X-ray Images," IEEE Transactions on Industrial Informatics, vol. 17, no. 9, pp. 6539–6549, 2021.

[18] S. Govindarajan, S. R. Manuskandan and R. Swaminathan, "Diagnostics of Multi Drug Resistant Tuberculosis in Chest Radiographs Using Local Textures & Extreme Gradient Boosting," Current Directions in Biomedical Engineering, vol. 9, no. 1, pp. 721–724, 2023.

[19] S. Kalaivani and K. Seetharaman, "A Three-stage Ensemble Boosted Convolutional Neural Network for Classification and Analysis of COVID-19 Chest X-ray Images," Int. Journal of Cognitive Computing in Engineering, vol. 3, pp. 35–45, 2022.

[20] R. Soundrapandiyan, H. Naidu, M. Karuppiah, M. Maheswari and R. C. Poonia, "AI-based Wavelet and Stacked Deep Learning Architecture for Detecting Coronavirus (COVID-19) from Chest X-ray Images," Computers and Electrical Engineering, vol. 108, Article no. 108711, 2023.

[21] P. Mooney, "Chest X-ray Images (Pneumonia)," Kaggle, [Online], Available: https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia, Mar. 2018.

[22] M.-L. Huang and Y.-C. Liao, "Stacking Ensemble and ECA-EfficientNetV2 Convolutional Neural Networks on Classification of Multiple Chest Diseases Including COVID-19," Academic Radiology, vol. 30, no. 9, pp. 1915–1935, 2023.

[23] G. Erol and B. Uzbaş, "Detection of COVID-19 Severity and Mortality from Blood Parameters by Ensemble Learning Methods," Karaelmas Fen ve Mühendislik Dergisi, vol. 13, p. 316–328, 2023.

[24] A. Banerjee, A. Sarkar, S. Roy, P. K. Singh and R. Sarkar, "Covid-19 Chest X-ray Detection through Blending Ensemble of CNN Snapshots," Biomedical Signal Processing and Control, vol. 78, p. 104000, 2022.

[25] Peshotan, "Summary of COVID-19 Radiography Database," [Online], Available: Https://github.com/sfu-db/covid19-datasets/blob/master/datasets-details/radiography-kaggle-COVID-19-dataset.md, 2020.

[26] T. Pham, "Classification of COVID-19 Chest X-rays with Deep Learning: New Models or Fine Tuning?" Health Information Science and Systems, vol. 9, no. 1, DOI: 10.1007/s13755-020-00135-3, 2020.

[27] T. Rahman, "Covid-19 Radiography Database," [Online], Available: https: /www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database, Mar. 2022.

[28] Y. Wu, H. Gunraj, C. en Amy Tai and A. Wong, "COVIDx CXR-4: An Expanded Multi-institutional Open-source Benchmark Dataset for Chest X-ray Image-based Computer-aided COVID-19 Diagnostics," arViv: 2311.17677, DOI: arxiv-2311.17677, 2023.

[29] P. P. Dalvi, D. R. Edla and B. R. Purushothama, "Diagnosis of Coronavirus Disease from Chest X-ray Images Using Densenet-169 Architecture," SN Computer Science, vol. 4, Article no. 214, [Online], Available: https://link.springer.com/article/10.1007/s42979-022-01627-7, Feb. 2023.

[30] M. K. Bohmrah and H. Kaur, "Classification of COVID-19 Patients Using Efficient Fine-tuned Deep Learning DenseNet Model," Global Transitions Proceedings, vol. 2, no. 2, pp. 476–483, 2021.

[31] A. Ala and Polat, "Detection of COVID-19 from Computed Tomography Images with DenseNet based Deep Learning Models," Proc. of the 2021 29th Signal Processing and Communications Applications Conference (SIU), pp. 1-4, Istanbul, Turkey, 2021.

[32] K. Han et al., "A Survey on Vision Transformer," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 1, pp. 87–110, 2023.

[33] S. Park et al., "Multi-task Vision Transformer Using Low-level Chest X-ray Feature Corpus for COVID-19 Diagnosis and Severity Quantification," Medical Image Analysis, vol. 75, Article no. 102299, 2022.

[34] C. C. Ukwuoma et al., "Automated Lung-related Pneumonia and COVID-19 Detection Based on Novel Feature Extraction Framework and Vision Transformer Approaches Using Chest X-ray Images," Bioengineering, vol. 9, no. 11, Article no. 709, 2022.

[35] T. Aitazaz, A. Tubaishat, F. Al-Obeidat, B. Shah, T. Zia and A. Tariq, "Transfer Learning for Histopathology Images: An Empirical Study," Neural Computing and Applications, vol. 35, pp. 7963-7974, Jul. 2022.

[36] G. Boesch, "Vision Transformers (ViT) in Image Recognition," 2024 Guide, [Online], Available: https://viso.ai/deep-learning/vision-transformer-vit/, Jun. 2024.

[37] 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, p. e0256630, 2021.

[38] S. Kaleem, A. Sohail, M. U. Tariq, M. Babar and B. Qureshi, "Ensemble Learning for Multi-class COVID-19 Detection from Big Data," Plos One, vol. 18, no. 10, p. e0292587, 2023.

[39] O. O. Abayomi-Alli et al., "An Ensemble Learning Model for COVID-19 Detection from Blood Test Samples," Sensors, vol. 22, no. 6, Article no. 2224, 2022.

[40] S. Shukla, K. Arya, B. Garg and P. Lezama, "Pneumonia Detection Using Gradient Boosting on Selected Features Extracted from DenseNet," Proc. of the 7th ASRES Int. Conf. on Intelligent Technologies, in Book Series: Lecture Notes in Networks and Systems, vol. 685, pp. 181–195, 2022.

[41] A. S. Ashour, M. M. Eissa, M. A. Wahba, R. A. Elsawy, H. F. Elgnainy, M. S. Tolba and W. S. Mohamed, "Ensemble-based Bag of Features for Automated Classification of Normal and COVID-19 CXR Images," Biomedical Signal Processing and Control, vol. 68, p. 102656, 2021.

[42] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince and F. Herrera, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting- and Hybrid-based Approaches," IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews), vol. 42, no. 4, pp. 463–484, 2011.

[43] R. Odegua, "An Empirical Study of Ensemble Techniques (Bagging, Boosting and Stacking)," Proc. of Deep Learning Conf. (IndabaXAt), DOI: 10.13140/RG.2.2.35180.10882, 2019.

[44] Z. Hu, H. Qiu, Z. Su, M. Shen and Z. Chen, "A Stacking Ensemble Model to Predict Daily Number of Hospital Admissions for Cardiovascular Diseases," IEEE Access, vol. 8, pp. 138719–138729, 2020.

[45] M. Gour and S. Jain, "Automated COVID-19 Detection from X-ray and CT Images with Stacked Ensemble Convolutional Neural Network," Biocybernetics and Biomedical Engineering, vol. 42, no. 1, pp. 27–41, 2022.

[46] A. A. Hammam, H. H. Elmousalami and A. E. Hassanien, "Stacking Deep Learning for Early COVID-19 Vision Diagnosis," in Chapter: Big Data Analytics and Artificial Intelligence Against COVID-19: Innovation Vision and Approach, vol. 78, pp. 297–307, DOI: 10.1007/978-3-030-55258-9_18, 2020.

[47] J. A. ALzubi, B. Bharathikannan, S. Tanwar, R. Manikandan, A. Khanna and C. Thaventhiran, "Boosted Neural Network Ensemble Classification for Lung Cancer Disease Diagnosis," Applied Soft Computing, vol. 80, pp. 579–591, 2019.

[48] G. Li, R. Togo, T. Ogawa and M. Haseyama, "Boosting Automatic COVID-19 Detection Performance with Self-supervised Learning and Batch Knowledge Ensembling," Computers in Biology and Medicine, vol. 158, Article no. 106877, 2023.

[49] N. Habib, M. M. Hasan, M. M. Reza and M. M. Rahman, "Ensemble of CheXNet and VGG-19 Feature Extractor with Random Forest Classifier for Pediatric Pneumonia Detection," SN Computer Science, vol. 1, no. 6, Article no. 359, 2020.

[50] A. B. Godbin and S. G. Jasmine, "A Machine Learning Based Approach for Diagnosing Pneumonia with Boosting Techniques," in Chapter: Machine Intelligence for Smart Applications, Part of the Book Series: Studies in Computational Intelligence, vol. 1105, pp. 145–160, Springer, 2023.

ملخص البحث:

إنّ أمراض الرّئة، مثل فيروس كورونا وذات الرّئة، يمكن أنْ تسبّب تداعيات شديدة، بما في ذلك صعوبات في التّنفُّس، وضعف في عمل الرّئة، وفشل في الجهاز التّنفُّسي، ممّا قد يهدّد الحياة إذا لم يُعالج على الفْور. وقد أثبتت الصُّور الشُّعاعية للصّدْر سرعتها وفاعليتها وجدواها في تشْخيص هذه الأمراض والكَشْف عنها. بالإضافة إلى ذلك، أعطى الذّكاء الاصطناعي من خلال تعلُّم الآلة والتّعلُّم العميق نتائج واعدةً في الكشْف عن العديد من أمراض الرّئة، ومنها فيروس كورونا وذات الرّئة. وقد أسهمت قدرةُ هذه التكنولوجيا على التّعامل مع مجموعات البيانات الضّخمة وتحليلها في الحدّ من انتشار هذه الأمراض وكان لها دور بارز في تقدُّم البحث العلمي في العديد من التّخصُّصات الطِّبّية والبيولوجية.

في هذا البحث، نستخدم مجموعةً من تقنيات توحيد الأداء عن طريق خوارزميات مختلفة لتحسين أداء نماذج التّصنيف في الكشْف عن الإصابة بقيروس كورونا وذات الرّئة. وبالتّحديد، فقد ركّزنا على إيجاد نماذج مجمَّعة تقوم على توحيد الأداء وتستند على الشّبكات العصبية الالتفافية ومحوّلات الرّؤية. وكان هدفنا تحديد أفضل تلك النّماذج القائمة على توحيد الأداء في تشْخيص أمراض الرّئة، وذلك من خلال تقييم قدرة النّماذج المستخدمة على التّصنيف الصّحيح للصُّور الشّعاعية للصّدر إلى صُوَرٍ تدلُّ على الإصابة بفيروس كورونا، وأخرى تدلُّ على الإصابة بذات الرّئة، وثالثة طبيعية تشير ألى الخُلوّ من أمراض الرّئة.

ويمكن القول إنّ تقنيات توحيد الأداء تعْمل على تحسين دقّة التّشخيص للنّماذج المستخدمة في الكشْف عن أمراض الرّئة، القائمة على تعلُّم الآلة.