A multi-class classification framework for disease screening and disease diagnosis of COVID-19 from chest X-ray images

Ebenezer Jangam¹,² · Chandra Sekhara Rao Annavarapu² · Aaron Antonio Dias Barreto³

Received: 14 September 2021 / Revised: 5 May 2022 / Accepted: 24 August 2022 / Published online: 21 September 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract
To accurately diagnose multiple lung diseases from chest X-rays, the critical aspect is to identify lung diseases with high sensitivity and specificity. This study proposed a novel multi-class classification framework that minimises either false positives or false negatives that is useful in computer aided diagnosis or computer aided detection respectively. To minimise false positives or false negatives, we generated respective stacked ensemble from pre-trained models and fully connected layers using selection metric and systematic method. The diversity of base classifiers was based on diverse set of false positives or false negatives generated. The proposed multi-class framework was evaluated on two chest X-ray datasets, and the performance was compared with the existing models and base classifiers. Moreover, we used LIME (Local Interpretable Model-agnostic Explanations) to locate the regions focused by the multi-class classification framework.

Keywords Stacked ensemble · Multi-class classification · COVID-19 · Deep learning · Transfer learning

Chandra Sekhara Rao Annavarapu
acsrao@iitism.ac.in

Ebenezer Jangam
jebenezer@vrsiddhartha.ac.in

Aaron Antonio Dias Barreto
aaron.18je0003@ece.iitism.ac.in

¹ Department of Information Technology, Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada, Andhra Pradesh, India
² Department of Computer Science Engineering, Indian Institute of Technology(ISM), Dhanbad, Jharkhand, India
³ Department of Electronics Engineering, Indian Institute of Technology(ISM), Dhanbad, Jharkhand, India
1 Introduction

Chest X-rays can be used to diagnose multiple diseases like effusion, pneumonia, infiltration, nodule, cardiomegaly. With the outbreak of COVID-19, the significance of a computer-aided diagnosis (CAD) system that assists radiologists in decision making and diagnosis of diseases has increased dramatically. Over the past decade, CAD has played a key role in diagnosing lung diseases. In the current scenario of COVID-19, the real-time reverse transcription-polymerase chain reaction (RT-PCR) method was used to test whether a patient is COVID-19-positive or not. However, the drawback of RT-PCR is high false negatives, which subsequently leads to the further spread of the virus, resulting in a pandemic [3, 16, 31].

The alternatives of RT-PCR for the diagnosis of COVID-19 are CT scan and chest X-ray scan. Chest X-ray scan is an affordable option that takes less time than RT-PCR and CT scan. Diagnosis using CT scan and chest X-ray scan of the symptomatic individuals yielded significantly low false negatives when compared to RT-PCR [8, 37, 43, 71]. However, the challenges with medical image analysis are as follows. The first challenge is a lack of large datasets. The second challenge is in feature extraction. A lack of large datasets can be addressed using transfer learning; feature extraction can be addressed using pre-trained deep-learning models and fully connected layers. In the design of a deep neural network, the width and depth are two factors that affect the model’s performance. The width refers to the number of neurons in the fully connected layers, whereas depth refers to trainable layers.

In addition to the challenges mentioned above, detection of multiple lung diseases by distinguishing the features related to COVID-19-pneumonia and other pneumonia is a challenging task [39, 40] because COVID-19 characteristics are similar to other kinds of pneumonia [20]. The traditional feature learning approaches may not discern the dynamic characteristics in COVID-19 chest X-ray scans, and CT scans [4]. However, promising alternative solutions like convolutional neural network (CNN) were used to automatically extract the features for complex tasks in the past [58]. Therefore, deep learning (DL) using CNN is a promising option for automatic feature extraction in a disease like COVID-19.

However, DL models need a large amount of data for the accurate prediction of the disease. Although data augmentation techniques were used to increase the size of the datasets, the performance of the DL model did not improve significantly. Therefore, transfer learning (TL) techniques can be employed to improve the performance of the DL models when the required amount of data are not available. For instance, the weights of the pre-trained CNN models of large-scale datasets, such as ImageNet, can be transferred to the task of COVID-19 diagnosis. Moreover, some studies reported that the pre-trained CNN models like Visual Geometry Group (VGG)-19 [61], the residual network (ResNet) [70] and the densely connected convolutional network (DenseNet) [21], gave high accuracy in the detection of COVID-19 from chest CT scan images and chest X-ray scan images. This study used DL, TL, and stacking concepts to design an ensemble to distinguish COVID-19 pneumonia from other pneumonia and normal chest X-ray images.

The contributions of the study are as follows:

− We proposed a novel multi-class classification framework that can generate a pair of classifiers: disease screening and detection models.
− We proposed a new pair of selection measures to select accurate and diverse base classifiers from the pool of base classifiers for disease screening and disease detection mode, respectively.
– We explored the concessions between recall and precision to select the optimal threshold that results in high recall and accuracy.
– We used explainable Artificial Intelligence (AI) to illustrate that the proposed model focuses on the right areas of the chest X-ray images to predict.

The study is organized as follows. Preliminaries needed for the study are explained in Section 2. Existing models for multi-class classification related to COVID-19 are summarised in Section 3. The systematic approach, selection metrics, and the proposed multi-class classification framework are explained in Section 4. The data and methodology of the experiment are explained in Section 5. The results obtained from the experiment are analyzed in Section 6. The performance of the proposed stacked ensemble is compared with the existing models in Section 7. Section 8 concludes the study.

2 Preliminaries

This section explains the fundamentals used in this study.

2.1 Deep learning

Deep learning (DL) [38] is a subset of machine learning that is based on artificial neural networks (ANNs). There can be multiple hidden layers depending on the complexity of the problem under consideration. One of the requirements of the DL is that it needs large datasets for training.

2.2 Transfer learning

When large datasets are not available for training, transfer learning [50] is a promising solution. Transfer learning helps by avoiding the need to initiate the learning process from scratch. The weights learned from conventional large datasets are transferred to a model used to address a problem of interest. There are two primary advantages of transfer learning: saving time during training and solving a problem for which large datasets are not available.

2.3 Convolutional neural network (CNN)

A CNN has two components: a convolutional base and a classifier. The convolutional base can capture the features from an image. The classifier is responsible for assigning labels to each example based on the features learned by the convolutional base.

In medical image analysis, CNNs have exhibited better performance than clinical experts in various tasks [56]. In the case of chest radiography, as massive datasets are available, it is feasible to use a CNN to learn the features of a disease when the features are known. However, for diseases that have ambiguous characteristics, a CNN alone cannot capture the features [69].

A CNN performed better when compared with experts under different clinical scenarios. In the case of chest imaging, diverse algorithms have been developed to process the data. Training CT images in deep learning is a complex task because of the involvement of a 3D signal.
2.4 Stacking

Stacking is an ensemble learning technique in which heterogeneous base classifiers are combined using a meta classifier to provide better performance for a given task. The key phrases in the generation of a stacking ensemble are (1) the generation of a pool of base classifiers, (2) the selection of base classifiers that form the stacked ensemble, and (3) the combination of the outputs of the base classifiers using a meta-classifier [57].

2.5 Pre-trained models

A model was pre-trained on a large benchmark dataset to solve a similar but different problem. Training a new model on a large dataset is expensive; it is common to use existing pre-trained models. Although there are many existing pre-trained models [6], the following models were selected based on the popularity and performance when employed for various classification related tasks.

2.5.1 Visual geometry group (VGG)-19

VGG-19 [61] is a variation of a CNN proposed by the Visual Geometry Group from the University of Oxford that contains 19 layers and nearly 144 million parameters. The authors [61] investigated the impact of the depth of the CNN on its accuracy. After a thorough evaluation, the optimal performance was found at a depth of 16–19 layers. VGG-19 was successfully used for localization and classification tasks.

2.5.2 Residual network (ResNet)-101

ResNet-101 [70] is a variation of CNN with 101 layers that uses the concept of a skip connection to address the problem of vanishing/exploding gradients; this problem is prevalent in plain deep networks. A skip connection feeds the input from one layer to the following layer without any modification. ResNet-101 has been successfully applied to image classification, localization, and detection tasks.

2.5.3 Densely connected convolutional networks (DenseNet)-169

DenseNet-169 [21] is a variation of a CNN that addressed the vanishing/exploding gradient problem by ensuring maximum information and gradient flow. Each layer in every dense block is connected to every other layer in a feed-forward fashion. In addition to mitigating the vanishing gradient problem, DenseNet-169 has other advantages, such as the promotion of feature reuse, a reduction in the number of parameters, and the promotion of feature propagation.

2.5.4 Wide residual network (Wide ResNet)-50-2

Wide ResNet-50-2 [73] is a modified residual network that works by decreasing the depth and increasing the width. Wide ResNet-50-2 has a depth of 50 and a width of two with approximately 69 million parameters.
3 Literature review

In this section, we focus on existing multi-class classification techniques to distinguish COVID-19-pneumonia from other types of pneumonia and pneumonia-free chest X-ray images. Existing models proposed for multi-class classification of COVID-19-pneumonia were listed in Table 1.

Most of the existing studies used TL because available datasets were small. Some of the studies used a single pre-trained computer vision model for multi-class classification. AlexNet model was used for the classification of COVID-19-pneumonia, viral pneumonia, bacterial pneumonia, and normal CXR scans [24]. VGG-16 model was used for the classification of three categories of CXR scans: COVID-19 pneumonia, non-COVID-19 pneumonia, and the healthy specimens [48]. DCNN based model Inception V3 was used for the detection of coronavirus pneumonia infected patients using chest X-ray radiographs [7]. A study presented a deep CNN model based on Xception architecture for classification of normal, pneumonia-bacterial, pneumonia-viral and COVID-19 chest X-ray images [35]. The DarkNet model was used in another study [49] as a classifier for the 'you only look once (YOLO) real time object detection system.

Studies have explored multiple pre-trained models for multi-class classification and reported the best-performing models. According to study [13], DenseNet-201 outperforms other different deep CNN networks. Another study [60] classified the chest X-rays into four categories viz. normal, pneumonia, tuberculosis (TB), and COVID-19. Further, the X-rays indicating COVID-19 were classified by severity-basis into mild, medium, and severe categories using ResNet-18. The deep learning model VGG-16 has given better results for classifying pneumonia, TB, and normal. For the segregation of normal pneumonia and COVID-19, the DenseNet-161 has given better accuracy. Pre-trained models on ImageNet and the NIH ChestX-ray14 dataset were used for fine-tuning the classification model in the study [9]. Two deep learning models following a lightweight architecture were proposed and proved to be more robust and reliable in COVID-19 detection than a baseline ResNet8 by another study [32].

A study [23] evaluated various pre-trained models for diagnosing COVID-19, pneumonia, and lung cancer from a combination of chest X-ray and CT images. The study found that the combination of VGG-19 and a CNN model outperforms three other combinations [23]. A fine-tuned DenseNet-121 achieved high accuracy for four classifications [30]. The study [33] considered 11 pre-trained convolutional neural network models. The results illustrated that VGG-16, ResNet50V2, and dense neural network (DenseNet169) models had given better accuracy for the classes of COVID-19, viral (Non-COVID-19) pneumonia, and bacterial pneumonia, respectively. A study [74] used multiple pre-trained convolutional backbones, such as VGG-16, ResNet-50, and EfficientNetB0, and the highest overall detection accuracy was achieved by EfficientNetB0. A gradient class activation mapping technique was used to highlight the regions of the input image that were important for predictions [74].

Multiple pre-trained models have been combined to achieve better performance. For instance, Insta Cov Net-19 was a combination of ResNet-101, Inception v3, Xception, MobileNetv2 and NASNet [17]. Another study [27] used pre-trained deep CNN models (ResNet-18, ResNet-50, ResNet-101, VGG-16 and VGG-19) for feature extraction and and a support vector machine (SVM) as a classifier. A study proposed concatenation of the
| Study                     | Data                                           | Method                         | Contribution                                                      |
|--------------------------|-----------------------------------------------|--------------------------------|------------------------------------------------------------------|
| Ibrahim et al. [24]      | chest X-rays combined from different sources | AlexNet                        | Two-way, three-way and four-way classification                   |
| Nishio et al. [48]       | 1248 images taken from two public chest X-ray datasets | VGG-16                        | CADx system for evaluation of COVID-19 pneumonia, non-COVID-19 pneumonia and healthy images |
| Asif et al. [7]          | Mixed dataset of CXR and CT scan images       | Deep CNN based InceptionV3     | Three class classification of COVID-19 pneumonia, non-COVID-19 pneumonia and healthy images |
| Khan et al. [35]         | Mixed dataset of CXR images                   | Modification of Xception architecture and transfer learning | Three-class classification of COVID-19 CXR images |
| Chowdhury et al. [13]    | Mixed dataset of CXR images                   | Image augmentation, transfer learning and multiple pre-trained models | Comparision of performance of different pre-trained models |
| Shelke et al. [60]       | Indian dataset of CXR images                  | Pneumonia detection using DenseNet-161, COVID-19 detection using ResNet-18 and VGG-16 | Classification of COVID-19 pneumonia from normal, pneumonia and tuberculosis |
| Bassi and Attux [9]      | Indian dataset of CXR images                  | Pneumonia detection using DenseNet-161, COVID-19 detection using ResNet-18 and VGG-16 | Classification of COVID-19 pneumonia from normal, pneumonia and tuberculosis. |
| Karakanis et al. [32]    | CXR images and synthetic images               | GAN for synthetic images; Lightweight ResNet8 for COVID-19 detection; GRAD CAM for heatmap generation | Three-class classification of COVID-19 pneumonia from normal, pneumonia |
| Ibrahim et al. [23]      | 33,676 CXR and CT images from RSNA and SIRM  | ResNet152V2 and VGG19          | For classification of COVID-19 pneumonia from lung cancer and pneumonia, VGG19 provided better accuracy. |
| Ibrahim et al. [23]      | CXR images from two different datasets        | Evaluation using ResNet, MobileNet, DenseNet and InceptionV3 | Comparision of accuracy of five pre-trained models. DenseNet121 provided better accuracy. |
| Karar et al. [33]        | CXR images                                   | VGG16, ResNet50V2 and DenseNet169 | VGG16, ResNet50V2 and DenseNet169 provided better performance for COVID19, Viral pneumonia and bacterial pneumonia respectively. |
Xception and ResNet50V2 networks for classifying X-ray images into into normal, pneumonia and COVID-19 [53]. Some of the studies have proposed variations of the existing architectures: CovXNet [44]; CoroDet [22]; and decompose, transfer and compose (DeTraC) [1] for the classification of COVID-19 chest X-ray images.

Some models propose a methodology that comprises multiple steps. For example, a three-step model is presented in [10]. The first step is to detect the presence of pneumonia in the given chest x-ray image. The second step is to distinguish between COVID-19 and pneumonia. The last step is aimed at localizing the areas in the X-ray that are symptomatic of the COVID-19 presence [10].

In [45] lung ultrasound images are used for detection of COVID 19. The authors of [45] propose a CNN network with fewer learning parameters, a multi layer fusion approach is employed to increase the efficiency of the model. In [26] the authors concluded that for diagnosis of COVID 19, chest CT scan is sensitive and moderately specific, chest X-ray is moderately sensitive and moderately specific while ultrasound is sensitive but not specific, therefore there is a higher probability of getting a false positive with ultrasound than with chest X-ray. Therefore, we chose to use datasets consisting of chest X-ray images.
Most of the deep learning models developed to detect COVID-19, usually are trained on either chest CT scans or chest X-ray images but in [46] a CNN model is proposed that can collectively be used for both chest CT scans and Chest X-ray images. In [66] the authors used a CNN to extract features to learn individual image level representations and they used a graph convolutional network to learn relation-aware features. They then used Deep feature fusion to fuse the learned image level features and the relation-aware features.

In [68] the authors first used pretrained models to learn features, they then proposed a novel transfer feature learning algorithm which used number of layers to be removed as a hyper-parameter. They then proposed a selection algorithm of pre-trained network for fusion to determine the two best models characterized by the pretrained models and the number of layers to be removed. Finally, they used deep CCT fusion by discriminant correlation analysis to fuse the two features from the two models.

Some studies [64] have focused on application of lightweight deep learning architecture for binary classification of pneumonia and normal images. Development of light weight architectures reduces response time and computational resources required for disease detection. Feature extraction was performed to reduce computational costs to detect COVID-19. Authors in [55] used CNN for feature extraction and Marine Predators algorithm to select relevant features. A hybrid approach using parallel fusion and optimization of deep learning models for COVID-19 detection in [34]. Authors in [5] proposed mask extraction method based on multi-agent Deep Reinforcement Learning (DRL) from COVID-19 CT images.

Since using AI for efficient detection of COVID-19 from chest X-ray images or chest CT scans requires patient records from hospitals or test records from testing facilities, privacy becomes a major issue therefore Federated Machine Learning is a promising area to explore in this context. The authors of [2] studied the efficacy of federated learning versus traditional learning. In this paper the authors used a descriptive dataset and chest X-ray images. The authors concluded that federated machine learning gives a better prediction accuracy but a higher performance time as compared to a traditional machine learning. In [25, 47] a light weight CNN-tailored shallow architecture was proposed to detect COVID-19 from Chest X-ray. The model was designed with fewer parameters as compared to other deep learning models and the model produced no false negatives. Researchers [36] have used structured model pruning to improve model memory usage and speed on TPUs without losing accuracy, especially for small datasets

Researchers used single datasets for the evaluation of COVID-19 detection models [12, 14, 19, 41, 42, 51, 52, 59, 62, 67, 72]. However, it is good to evaluate on multiple datasets before arriving at a conclusion about the performance of a model.

Models [7, 13, 18, 23, 24, 30, 33, 35, 48, 60, 74] were trained for multiple classifications; however, the single best performing pre-trained model was considered as the best model. A few models combined features from different pre-trained models [17, 27, 53], but there is no guarantee that the highlighted combination was the best one. Models [1, 22, 44] were modifications in existing architectures, and further modifications in architectures can improve their performance. Moreover, none of the existing models focused on the minimisation of false positives.

The reduction of false positives is crucial in diagnosing individuals as it prevents them from undergoing unnecessary treatment; financial and medical resources can also be utilized more effectively. Although there are a few papers that address the minimization of false positives [29, 42], there was insufficient coverage of the topic and, specifically, the consequences of false positives. Stacked ensembles were designed for medical image analysis to provide high accuracy [29] and high recall along with that accuracy [28].
Models [12, 14, 19, 41, 42, 51, 52, 59, 62, 67, 72] that detected COVID-19 based on binary classification did not consider the other pneumonia cases and were evaluated using a single dataset. Models [7, 9, 13, 23, 24, 30, 32, 33, 35, 48, 49, 60, 74] were trained for multi-class classification; however, the single best performing pre-trained model is considered as the best model. A few models combined features from different pre-trained models [17, 27, 53], but there is no guarantee that the highlighted combination was the best one. Models [1, 22, 44] were modifications in existing architectures and further modifications in architectures can improve the performance. Moreover, all the existing models did not focus on minimizing false positives [28, 29], which is crucial in the case of COVID-19 screening. On the other hand, minimization of false positives is crucial in the diagnosis of COVID-19. We propose a multi-class framework for disease screening and diagnosis using a systematic approach.

Although there are existing deep learning models, the focus of the paper is to select efficient models for a given task. We proposed a selection metric and systematic method to select efficient deep learning models to form the stacked ensemble.

Existing works used pre-trained models for COVID-19 detection. Some authors used selective pre-trained models and compared the performance of the pre-trained models and reported the better performing models. Some authors have modified the existing pre-trained models architecture.

In this study, we proposed a metric and method to select efficient base classifiers to form the stacked ensemble. Most of the existing works do not focus on the minimisation of false positives or false negatives which is a crucial aspect in computer aided diagnosis. We designed stacked ensembles that minimise false positives or false negatives.

4 Proposed multi-class classification framework

This section presents an overview of the method used to generate stacked ensembles in a multi-class classification framework with high accuracy, followed by the selection measures used to select base classifiers. The details of the proposed multi-class classification framework are presented in the last subsection. The overview of the proposed multi-class classification framework was presented in Fig. 1.

![Fig. 1 Overview of generation of multi-class classification framework](image-url)
4.1 Method for the generation of multi-class classification framework

The generation of a stacked ensemble comprises of three phases: generation, selection and aggregation.

1. The set of pre-trained models that provide a better performance in the detection of COVID-19 is selected. Each pre-trained model is used to generate multiple base classifiers by appending the fully connected layers.
2. The second step is to select the efficient base classifiers that generate a distinct set of false positives or false negatives depending on the selection mode. The base classifiers are arranged in descending according to their performance and ability to provide distinct false positives.
3. The meta-classifier collects the outputs from the base classifiers and combines them using a weighted average to predict the class of the given sample.

4.2 Selection metrics

A pair of selection metrics were designed for the multi-class classification of COVID-19, where accuracy, recall, and precision play a key role.

Suppose there are $N$ examples in the validation set. Out of $N$ examples, $N_1$ are positive examples and $N_0$ are negative examples. In other words, the validation set constituted of positive and negative examples $N = N_0 + N_1$. The diversity between a pair of classifiers $c_i$ and $c_j$ can be calculated following. The selection metric is designed such that diverse and efficient base classifiers are chosen to form the disease screening model and disease diagnosis model. The diversity is measured in a diverse set of false positives generated from a considered pair of base classifiers. When the false positives of a model are minimized, the precision of the corresponding model increases. The following notations were adopted for designing the selection metric:

- $N$ is the total number of samples in the dataset.
- $N_p$ is the number of positive samples.
- $N_n$ is the number of negative samples.
- $c_i$ is the $i^{th}$ base classifier.
- $N_{0}^{ab}$ is the number of samples that belong to the negative class with predictions $a$ and $b$ from the classifiers $c_i$ and $c_j$, respectively.
- $N_{1}^{ab}$ is the number of samples that belong to the positive class with predictions $a$ and $b$ from the classifiers $c_i$ and $c_j$, respectively.
- $N_{q1}^{ab}$ is the number of positive samples that belong to the disease class $q$ with predictions $a$ and $b$ from the classifiers $c_i$ and $c_j$, respectively.
- $N_{q0}^{ab}$ is the number of negative samples that belong to the disease class $q$ with predictions $a$ and $b$ from the classifiers $c_i$ and $c_j$, respectively.
- $a_i$ is the accuracy of the base classifier $c_i$.
- $r_i$ is the recall of the base classifier $c_i$.
- $p_i$ is the precision of the base classifier $c_i$.

4.2.1 Selection metric for disease screening model

Base classifiers with the following characteristics were selected to form disease screening model.
– The accuracy $a_i$ of the base classifier $c_i$ should be high.
– The recall $r_i$ of the base classifier $c_i$ should be high.
– The pair of base classifiers under consideration should give different sets of false negatives for each class. Let the number of classes be $k$. Among $k$ classes, $k - 1$ classes denote different diseases, and the remaining one class corresponds to a healthy or normal class.

The false negatives corresponding to each disease class should be reduced in disease screening mode, and the false positives corresponding to the normal class should be minimized. For a disease class $q$, the diversity of a pair of classifiers $c_i$ and $c_j$ is

$$s_{ijq} = \frac{(N_{q1}^{10} + N_{q1}^{01})}{(N_{q1}^{01} + N_{q1}^{00} + N_{q1}^{10} + N_{q1}^{00})}$$

For normal class, the diversity of a pair of classifiers $c_i$ and $c_j$ is

$$s_{ij} = \frac{(N_{01}^{10} + N_{01}^{01})}{(N_{01}^{01} + N_{00}^{01} + N_{10}^{01} + N_{00}^{00})}$$

where the value of $q$ ranges from 1 to $k - 1$.

The selection metric was calculated using the following formula:

$$a_i r_i \left( \sum_{q=1}^{k-1} \sum_{j=1}^{m} s_{ijq} + \sum_{j=1}^{m} s_{ij} \right)$$

### 4.2.2 Selection metric for disease diagnosis model

Base classifiers with the following features were selected to form disease diagnosis model.

– The accuracy $a_i$ of the base classifier $c_i$ should be high.
– The precision $p_i$ of the base classifier $c_i$ should be high.
– The pair of base classifiers under consideration should give different sets of false positives for each class. Let the number of classes be $k$. Among $k$ classes, $k - 1$ classes denote different diseases, and the remaining one class corresponds to a healthy or normal class.

The false positives corresponding to each disease class should be reduced in disease diagnosis mode, and the false negatives corresponding to the normal class should be minimized. For a disease class $q$, the diversity of a pair of classifiers $c_i$ and $c_j$ is

$$d_{ijq} = \frac{(N_{q0}^{10} + N_{q0}^{01})}{(N_{q0}^{01} + N_{q0}^{00} + N_{q0}^{10} + N_{q0}^{00})}$$

For normal class, the diversity of a pair of classifiers $c_i$ and $c_j$ is

$$d_{ij} = \frac{(N_{01}^{10} + N_{10}^{01})}{(N_{10}^{01} + N_{00}^{01} + N_{10}^{00} + N_{00}^{00})}$$

The selection metric was calculated using the following formula:

$$a_i p_i \left( \sum_{q=1}^{k-1} \sum_{j=1}^{m} d_{ijq} + \sum_{j=1}^{m} d_{ij} \right)$$
Algorithm 1

**Data:** Validation set $V$

**Input:** Pre-trained models $p_1, p_2, \ldots, p_n$

A variable $k$ which is the number of base classifiers in the stacked ensemble model $E$

**Output:** A stacking based multi-class classification model for disease screening model $DS$

$i=1, j=1$

Set of base classifiers $B = \emptyset$

Set of diversity measures $D = \emptyset$

Set of selection measures $S = \emptyset$

while $i \leq n$

| Append FC layers varying from 1 to $m$ to pre-trained model $p_i$ to generate a set of base classifiers $b_{k+m(i-1)}$ where $k \in \{1, 2, 3, \ldots, m\}$
| Append the base classifiers to the set $B$
| $i = i + 1$

end

$i=1$

while $i \leq m \ast n$

| Evaluate $b_i$ on dataset $V$
| Calculate accuracy $a_i$, precision $pr_i$, and recall $r_i$

end

Sort $B$ using $a_i$ and $r_i$ and select the base classifier $ds_1$ with highest value of $a_i$ and $r_i$

while $i \leq 2m$

| Calculate the pairwise diversity for base classifiers $ds_1$ and $b_i$ in $B$
| For a disease class $q$, the diversity of a pair of classifiers $c_i$ and $c_j$ is
| $s_{1iq}^{dis} = \frac{(N_{q1}^{10} + N_{q1}^{01})}{(N_{q1}^{01} + N_{q1}^{00} + N_{q1}^{10} + N_{q1}^{00})}$
| For normal class, the diversity of a pair of classifiers $c_i$ and $c_j$ is
| $s_{1i}^{nor} = \frac{(N_{0}^{10} + N_{0}^{01})}{(N_{0}^{01} + N_{0}^{00} + N_{0}^{10} + N_{0}^{00})}$

Append the values $s_{1i}$ to set $S$

end

Select $ds_2$ and $ds_3$ based on the selection metric values;

Combine the outputs of $ds_1$, $ds_2$ and $ds_3$ using a meta-classifier $mc$ using weighted averaging to generate the ensemble $DS$ for disease screening;

4.3 Proposed multi-class classification framework

We proposed a multi-class classification framework comprising two models: disease screening and disease diagnosis models. Based on the context, a suitable model can be selected. The architecture of the multi-class classification framework is explained in this section.

The disease screening model and disease diagnosis model were projected as two different architectures because either one of them can be selected based on the context. Disease screening model minimises the false negatives corresponding to disease samples and the disease diagnosis model minimises the false positives corresponding to disease samples.
Algorithm 2 Algorithm to generate stacked ensemble for disease diagnosis.

Data: Validation set $V$
Input: Pre-trained models $p_1, p_2, \ldots, p_n$
A variable $k$ which is the number of base classifiers in the stacked ensemble model $E$
Output: A stacking based multi-class classification model for disease diagnosis $DD$
i=1, j=1;
Set of base classifiers $B = \emptyset$;
Set of diversity measures $D = \emptyset$;
Set of selection measures $S = \emptyset$;
\begin{algorithm}
while $i \leq n$
do
Append FC layers varying from 1 to $m$ to pre-trained model $p_i$ to generate a set of base classifiers $b_{k+m(i-1)}$ where $k \in \{1, 2, 3, \ldots, m\}$;
Append the base classifiers to the set $B$;
i = i + 1;
end
i=1;
\end{algorithm}
\begin{algorithm}
while $i \leq m \ast n$
do
Evaluate $b_i$ on dataset $V$;
Calculate accuracy $a_i$, precision $pr_i$, and recall $r_i$;
end
Sort $B$ based on $a_i$ and $p_i$ and select the base classifier $dd_i$ with highest product of $a_i$ and $p_i$;
\end{algorithm}
\begin{algorithm}
while $i \leq 2m$
do
Calculate the pairwise diversity for base classifiers $dd_1$ and $b_i$ in $B$ using the following equations.
\begin{itemize}
  \item For a disease class $q$, the diversity of a pair of classifiers $c_i$ and $c_j$ is
  \[d_{ij}^{dis} = \frac{(N_{q0}^{10} + N_{q0}^{01})}{(N_{q0}^{01} + N_{q0}^{00} + N_{q0}^{01} + N_{q0}^{00})}\]
  \item For normal class, the diversity of a pair of classifiers $c_i$ and $c_j$ is
  \[d_{ij}^{nor} = \frac{(N_{1}^{10} + N_{1}^{01})}{(N_{1}^{01} + N_{1}^{00} + N_{1}^{01} + N_{1}^{00})}\]
\end{itemize}
Append the values $d_{i1}$ to set $D$;
end
Select $dd2$ and $dd3$ based on the selection metric values;
Combine the outputs of $dd1$, $dd2$, $dd3$ using a meta-classifier $mc$ using weighted averaging to generate the ensemble $DD$ for disease diagnosis;
\end{algorithm}

4.3.1 Disease screening mode

We selected three base classifiers $ds_1, ds_2, ds_3$ based on the selection measure values according to Algorithm 1. The VGG-19 model combination was selected as the first base classifier when sorted based on the selection metric.
The first fully connected layer reduces the output from the VGG-19 to a single column comprised of 1,000 rows, which is reduced further by the second fully connected layer to a
column comprised of 200 rows. The final fully connected layer reduced the output to only three rows. The activation function used with the first and second fully connected layers was rectified linear unit (ReLu), and the activation function used with the third fully connected layer is sigmoid. The first fully connected layer has an additional dropout layer, which reduces the chances of over-fitting.

The second base classifier included in the disease screening model is ResNet-101 with three fully connected layers, and the third base classifier was Wide ResNet 50 2 with three fully connected layers. The three base classifiers are connected to the meta-classifier using weighted averaging. The architecture of the disease screening model is shown in Fig. 2.

### 4.3.2 Disease diagnosis model

Disease diagnosis model was generated according to Algorithm 2. The first model in the disease diagnosis model was VGG-19, appended with three fully connected layers. The first fully connected layer reduces the column comprised of 1,000 rows to a column vector comprised of 500 rows, which is further reduced to a column vector of 200 rows by the second fully connected layer. The last fully connected layer reduces this column vector to a column vector of 3 rows. The activation function used at the first and second fully connected layer is ReLu, and the sigmoid activation function is used at the third fully connected layer. A dropout layer is added at the first fully connected layer to reduce the chances of overfitting.

The second base classifier was comprised of Wide ResNet 50 2 along with one fully connected layer. The third base classifier chosen was comprised of Wide ResNet 50 2 and three fully connected layers. The outputs of the three classifiers were provided to the single neuron using weighted averaging. The meta classifier was designed such that it takes as input the outputs of three base classifiers and provides the predicted class as the output using weighted averaging as depicted in Fig. 3. The sigmoid function is used as the activation function to predict the class of the given image.

![Fig. 2 Disease screening model architecture](image-url)
4.3.3 Combined mode

The disease screening model and disease diagnosis model were projected as two different architectures because either one of them can be selected based on the context. If both functions (diagnosis and screening) are needed simultaneously both the models can be combined into a single model with four base classifiers. Among the four base classifiers two base classifiers are common to both the functions (disease screening and disease diagnosis).

5 Data and methodology

The details of the data and methodology used in the experiment are explained in this section.

5.1 Datasets

Two different chest X-ray datasets were used for the evaluation of the multi-class classification framework. The first dataset is a large chest X-ray image dataset, and the second dataset contains limited chest X-ray images.

1. **COVID-19 Radiography Database** [15]: This is a large dataset that contains three classes of chest X-ray images.
   - Source: https://www.kaggle.com/tawsifurrahman/covid19-radiography-database
   - Number of images in the dataset: 15,153
   - COVID-19 Positive labelled images: 3,616
   - Viral pneumonia labelled images: 1,345
   - Normal or healthy images: 10,192
   - Number of images used for training: 13,953
   - Number of images used for validation: 600
– Number of images used for testing: 600

2. **Chest X-ray images pneumonia and COVID-19** [11]: This is relatively smaller dataset with limited number of COVID-19 pneumonia chest X-ray images.
   - Source: https://www.kaggle.com/masumrefat/chest-xray-images-pneumonia-and-covid19
   - Number of images in the dataset: 6,118
   - COVID-19 Positive labelled images: 262
   - Viral pneumonia labelled images: 1,583
   - Normal or healthy images: 4,273
   - Number of images used for training: 5,600
   - Number of images used for validation: 218
   - Number of images used for testing: 300

The two datasets have been summarized in Table 2.

### 5.2 Hyper-parameter tuning and data augmentation

During the validation phase, we used a grid search technique to initialize the values of the hyper-parameters. The following values were investigated for each hyper-parameter:

- The values considered for random resized crop size were 128, 200, and 224.
- The values considered for the random resized crop scale were (0.5, 1.0), (1.0, 0.5) and (0.5, 0.5).
- The range of values considered for the random rotation angle were [-3°, 3°], [-5°, 5°], and [-10°, 10°].
- The values considered for the random horizontal flip probability were 0.3, 0.5, and 0.7.
- The values considered for Batch Size were 4, 8, 16, 32.
- The values considered for number of epochs were 50, 60, 70, 80, 90, 100.
- The values considered for learning rate were 0.1, 0.03, 0.01, 0.003, 0.001, 0.0003

During the training phase, the initial value of the batch size was set at four. The batch size was doubled until an error associated with memory limitations was encountered. The final value of the batch size was chosen such that it was the maximum value that did not

| Name                  | Dataset1 [15] | Dataset2 [11] |
|-----------------------|---------------|---------------|
| Total Images          | 15,153        | 6,118         |
| COVID-19 Positive     | 3,616         | 262           |
| Viral Pneumonia       | 1,345         | 1,583         |
| Normal Images         | 10,192        | 4,273         |
| Images used for training | 13,953   | 5,600         |
|                       | (9792 N, 945 P, 3216 C) | (4100 N, 1400 P, 100 C) |
| Images used for Validation | 600       | 218           |
|                       | (200 of each class) | (73 N, 83 P, 62 C) |
| Images used for Testing | 600 (200 of each class) | 300 (100 of each class) |

N : Normal, P : Pneumonia, C : COVID-19
generate an error related to memory. The number of epochs was initialized with a value of 50. After that, epochs were incremented by a step size often until the training and validation set accuracies remained constant. We used a PyTorch framework for the experiments using Adam as an optimizer and a cross-entropy loss function.

The following hyper-parameter values were used throughout the experiment:

- The total number of epochs was 100.
- The learning rate was set to 1e-3.
- The maximum value for the batch size was 16.
- The final value for the random resized crop size was 224.
- The final value for the random resized crop scale was (0.5, 1.0).
- The selected range of values for the random rotation angle was [-5°, 5°].
- The final value for the random horizontal flip probability was 0.5.

Tenfold cross validation was used to avoid overfitting. A dropout layer is added at the first fully connected layer of the base classifiers to reduce the chances of overfitting. Moreover, as the proposed models are stacked ensembles, the output of the model was received from multiple models that reduce the chances of overfitting.

5.3 Evaluation metrics

Four evaluation metrics were used to measure the performance of the proposed model as follows. The metrics for binary classification for a single class are:

- **Precision**: Precision is the fraction of positive predictions that belong to the positive class.
  
  \[ Precision = \frac{TP}{TP + FP} \]

- **Recall**: Recall is the fraction of positive examples in the dataset that are predicted positive.
  
  \[ Recall = \frac{TP}{TP + FN} \]

- **Specificity**: Specificity is the fraction of negative examples in the dataset that are predicted negative.
  
  \[ Specificity = \frac{TN}{TN + FP} \]

- **F1 Score**: F1 Score is the harmonic mean of precision and recall.
  
  \[ F1\ Score = \frac{2 \times precision \times recall}{precision + recall} \]

- **Accuracy**: Accuracy is the fraction of the total predictions that are correct.
  
  \[ Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \]

where \( TP \) denote true positives, \( TN \) denote true negatives, \( FP \) denote false positives and \( FN \) denote false negatives.

The metrics for multi-class classification were obtained by using macro averaging. Initially, Precision, Recall, F1 score, and accuracy were calculated for each class. The metrics for multi-class classification were obtained using macro averaging, i.e., by taking an average of the values obtained for three classes.
6 Experimental results

The results obtained in the experiments are presented in this section. The experimental results section is divided into two subsections. The first subsection consists of analyzing the results obtained from the proposed stacked ensemble using three CT scan datasets. The second subsection provides details of the proposed model at different threshold values.

6.1 Performance analysis of the proposed multi-class classification framework

The proposed framework was evaluated using two datasets of chest X-rays i.e. COVID-19 Radiography Database [15] and Chest X-ray images pneumonia and COVID-19 [11].

The following notation is used to denote the models:

1. $M_0$ denotes the base classifier formed using a pre-trained model $M$ appended with a softmax layer.
2. $M_1$ denotes the base classifier formed using a pre-trained model $M$ appended with additional fully connected and softmax layers.
3. $M_2$ denotes the base classifier generated using a pre-trained model $M$ appended with two additional fully connected layers and a softmax layer.

The candidates considered for the pre-trained model $M$ were ResNet-101, DenseNet-169, VGG-19, and WideResNet-50-2. The selection was based on their performances, as reported by the existing studies.

The precision is calculated as the weighted average of the precisions of the three classes. Where the precision of the COVID-19 class is given a higher weight as compared to the other two classes. The same has been done for recall also.

When the first dataset COVID-19 Radiography Database [15] was used for evaluation, the proportion of the train-validation-test was as follows. The test set was ensured to contain 600 images in order to test the generality of the model. The conditions required for the testing were facilitated and tested using a validation set containing 600 images. The remaining 13,953 images were included in the training set.

The precision values of the base classifiers and the proposed models were compared in Table 3. The base classifiers provided high values of precision for each class. The disease diagnosis model yielded fewer false positives than base classifiers and provided high precision values for Pneumonia and COVID-19 classes. However, the disease screening model minimized the false positives of the normal class; therefore, the precision value for the normal class is high. Similarly, recall values of three classes are depicted in Table 4. The disease screening model has provided high recall values for pneumonia and COVID-19 classes. However, the disease diagnosis model yielded a high recall value for the normal or healthy class. When average values of precision, recall, accuracy, and F1 score are considered in Table 6, disease screening and diagnosis models yielded better values than base classifiers. The average precision obtained for the disease diagnosis model was higher than other models, and the average recall resulting from the disease screening model was better than other models. However, the margin of difference is small as the dataset under consideration contained enough images for training (Tables 5 and 6).

When the second dataset Chest Xray Images PNEUMONIA and Covid-19 dataset [11] was used for evaluation, the proportion of the train-validation-test was as follows. The test set was ensured to contain 600 images in order to test the generality of the model. The conditions required for the testing were facilitated and tested using a validation set containing 600 images. The remaining 13,953 images were included in the training set (Tables 7 and 8).
### Table 3
Comparison of multi-class classification framework’s precision with base classifiers on COVID-19 Radiography Database [15]

| Model                  | Precision-Normal | Precision-Pneumonia | Precision-COVID-19 |
|------------------------|------------------|---------------------|--------------------|
| *VGG* 190              | 0.9091           |                     | 0.99               |
| *VGG* 191              | 0.9346           | 0.9947              | 0.9848             |
| *VGG* 192              | 0.9132           |                     | 0.9898             |
| *DenseNet* 1690        | 0.9171           |                     | 0.9747             |
| *DenseNet* 1691        | 0.8924           | 0.9945              | 0.9897             |
| *DenseNet* 1692        | 0.9337           | 0.9648              | 0.9366             |
| *ResNet* 1010          | 0.9458           | 0.9945              | 0.9213             |
| *ResNet* 1011          | 0.8959           |                     | 0.9522             |
| *ResNet* 1012          | 0.9471           | 0.9896              | 0.96               |
| *WideResNet* 50 20     | 0.9327           |                     | 0.9655             |
| *WideResNet* 50 21     | 0.9561           | 0.9701              | 0.9794             |
| *WideResNet* 50 22     | 0.9843           | 0.9707              | 0.9608             |
| Disease diagnosis model| 0.9346           |                     | 1.0                |
| Disease screening model| 0.9852           | 0.9947              | 0.9476             |

Bold entries denote highest values for a specific evaluation metric (Column)

The precision values of the base classifiers and the proposed models were compared in Table 9. The base classifiers provided high values of precision for each class. The disease diagnosis model yielded fewer false positives than base classifiers and provided high precision values for Pneumonia and COVID-19 classes. However, the disease screening model

### Table 4
Comparison of multi-class classification framework’s recall with base classifiers on COVID-19 Radiography Database [15]

| Model                  | Recall-Normal | Recall-Pneumonia | Recall-COVID-19 |
|------------------------|--------------|-----------------|-----------------|
| *VGG* 190              | 1            | 0.9             | 0.99            |
| *VGG* 191              | 1            | 0.94            | 0.97            |
| *VGG* 192              | 1            | 0.92            | 0.975           |
| *DenseNet* 1690        | 0.995        | 0.925           | 0.965           |
| *DenseNet* 1691        | 0.995        | 0.905           | 0.965           |
| *DenseNet* 1692        | 0.915        | 0.96            | 0.96            |
| *ResNet* 1010          | 0.96         | 0.9             | 0.995           |
| *ResNet* 1011          | 0.99         | 0.85            | 0.995           |
| *ResNet* 1012          | 0.985        | 0.95            | 0.96            |
| *WideResNet* 50 20     | 0.97         | 0.945           | 0.98            |
| *WideResNet* 50 21     | 0.98         | 0.975           | 0.95            |
| *WideResNet* 50 22     | 0.94         | 0.995           | 0.98            |
| Disease diagnosis model| 1            | 0.98            | 0.95            |
| Disease screening model| 1            | 0.93            | 0.995           |

Bold entries denote highest values for a specific evaluation metric (Column)
Table 5 Comparison of multi-class classification framework’s specificity with base classifiers on COVID-19 Radiography Database [15]

| Model               | Specificity-Normal | Specificity-Pneumonia | Specificity-COVID-19 |
|---------------------|--------------------|-----------------------|----------------------|
| VGG19               | 0.9497             | 1                     | 0.9948               |
| VGG19               | 0.9646             | 0.9975                | 0.9923               |
| VGG19               | 0.9523             | 1                     | 0.9948               |
| DenseNet169         | 0.9545             | 1                     | 0.9871               |
| DenseNet169         | 0.9397             | 0.9975                | 0.9948               |
| DenseNet169         | 0.9673             | 0.9817                | 0.9665               |
| ResNet101           | 0.9718             | 0.9974                | 0.9563               |
| ResNet101           | 0.9413             | 1                     | 0.9735               |
| ResNet101           | 0.9720             | 0.9949                | 0.9797               |
| WideResNet50 20     | 0.9649             | 1                     | 0.9821               |
| WideResNet50 21     | 0.9772             | 0.9847                | 0.9899               |
| WideResNet50 22     | 0.9923             | 0.9846                | 0.9797               |
| Disease diagnosis model | 0.965              | 1                     | 1                     |
| Disease screening model | 0.9923           | 0.9975                | 0.9723               |

Bold entries denote highest values for a specific evaluation metric (Column)

minimizes false positives of a normal class, and therefore the precision value for the normal class is high. Similarly, recall values of three classes are depicted in Table 10. The disease screening model has provided high recall values for pneumonia and COVID-19 classes. However, the disease diagnosis model yielded a high recall value for the normal or healthy class. When average values of precision, recall, accuracy, and F1 score are considered in

Table 6 Comparison of proposed multi-class classification framework with base classifiers on COVID-19 Radiography Database [15]

| Model               | Precision | Recall  | Accuracy | F1 Score |
|---------------------|-----------|---------|----------|----------|
| VGG19               | 0.9722    | 0.97    | 0.9633   | 0.9711   |
| VGG19               | 0.9747    | 0.97    | 0.97     | 0.9723   |
| VGG19               | 0.9732    | 0.9675  | 0.965    | 0.9704   |
| DenseNet169         | 0.9666    | 0.9625  | 0.9617   | 0.9646   |
| DenseNet169         | 0.9666    | 0.9575  | 0.955    | 0.962    |
| DenseNet169         | 0.9429    | 0.9488  | 0.945    | 0.9458   |
| ResNet101           | 0.9457    | 0.9625  | 0.9517   | 0.954    |
| ResNet101           | 0.9501    | 0.9575  | 0.945    | 0.9538   |
| ResNet101           | 0.9642    | 0.9638  | 0.965    | 0.964    |
| WideResNet50 20     | 0.9659    | 0.9688  | 0.965    | 0.9673   |
| WideResNet50 21     | 0.9713    | 0.9638  | 0.9683   | 0.9675   |
| WideResNet50 22     | 0.9691    | 0.9738  | 0.9717   | 0.9714   |
| Disease diagnosis model | **0.9836** | 0.97 | **0.9767** | **0.9768** |
| Disease screening model | 0.9688   | **0.98** | 0.975    | 0.9744   |

Bold entries denote highest values for a specific evaluation metric (Column)
### Table 7  Confusion Matrix of the Disease Diagnosis Model on the COVID-19 Radiography Database [15]

| Predicted Labels | Normal | Pneumonia | COVID 19 | Total |
|------------------|--------|-----------|---------|-------|
| True Labels      |        |           |         |       |
| Normal           | 200    | 0         | 0       | 200   |
| Pneumonia        | 4      | 196       | 0       | 200   |
| COVID 19         | 10     | 0         | 190     | 200   |
| Total            | 214    | 196       | 190     | 600   |

### Table 8  Confusion Matrix of the Disease Screening Model on the COVID-19 Radiography Database [15]

| Predicted Labels | Normal | Pneumonia | COVID 19 | Total |
|------------------|--------|-----------|---------|-------|
| True Labels      |        |           |         |       |
| Normal           | 200    | 0         | 0       | 200   |
| Pneumonia        | 3      | 186       | 11      | 200   |
| COVID 19         | 0      | 1         | 199     | 200   |
| Total            | 203    | 187       | 210     | 600   |

### Table 9  Comparison of multi-class classification framework’s precision with base classifiers on Chest Xray Images PNEUMONIA and Covid-19 dataset [11]

| Model              | Precision-Normal | Precision-Pneumonia | Precision-COVID-19 |
|--------------------|------------------|----------------------|--------------------|
| VGG19<sub>0</sub>  | 0.9151           | 0.9515               | 1                  |
| VGG19<sub>1</sub>  | 0.949            | 0.8205               | 0.9882             |
| VGG19<sub>2</sub>  | 0.9417           | 0.9612               | 1                  |
| DenseNet169<sub>0</sub> | 0.9375       | 0.8919               | 1                  |
| DenseNet169<sub>1</sub> | 0.9381       | 0.9074               | 1                  |
| DenseNet169<sub>2</sub> | 0.9684       | 0.9252               | 1                  |
| ResNet101<sub>0</sub> | 0.9327       | 0.97                 | 0.9896             |
| ResNet101<sub>1</sub> | 0.96         | 0.9252               | 1                  |
| ResNet101<sub>2</sub> | 0.9886       | 0.8772               | 1                  |
| WideResNet50 2<sub>0</sub> | 0.96         | 0.9245               | 1                  |
| WideResNet50 2<sub>1</sub> | 0.9588       | 0.9159               | 1                  |
| WideResNet50 2<sub>2</sub> | 0.9681       | 0.9083               | 1                  |
| Disease detection model | 0.9794 | 0.934               | 1                  |
| Disease screening model | **0.9897** | 1                    | 0.9429             |

Bold entries denote highest values for a specific evaluation metric (Column)
Table 10  Comparison proposed multi-class classification framework’s recall with base classifiers on Chest Xray Images PNEUMONIA and Covid-19 dataset [11]

| Model               | Recall-Normal | Recall-Pneumonia | Recall-COVID-19 |
|---------------------|---------------|------------------|-----------------|
| VGG190              | 0.97          | 0.98             | 0.91            |
| VGG191              | 0.93          | 0.96             | 0.84            |
| VGG192              | 0.97          | 0.99             | 0.94            |
| DenseNet1690        | 0.9           | 0.99             | 0.93            |
| DenseNet1691        | 0.91          | 0.98             | 0.95            |
| DenseNet1692        | 0.92          | 0.99             | 0.98            |
| ResNet1010          | 0.97          | 0.97             | 0.95            |
| ResNet1011          | 0.96          | 0.99             | 0.93            |
| ResNet1012          | 0.87          | 1                | 0.98            |
| WideResNet50 20     | **0.98**      | 0.96             | 0.94            |
| WideResNet50 21     | **0.98**      | 0.9567           | 0.96            |
| WideResNet50 22     | 0.91          | 0.99             | 0.97            |
| Disease diagnosis Model | 0.95      | 0.99             | 0.97            |
| Disease screening Model | 0.96      | 0.98             | **0.99**        |

Bold entries denote highest values for a specific evaluation metric (Column)

Table 12, disease screening and diagnosis models yielded better values than base classifiers. The average precision obtained for the disease diagnosis model was higher than other models, and the average recall resulting from the disease screening model was better than other models. However, the margin of difference is small as the dataset under consideration contained enough images for training. The margin of difference between stacked ensembles and base classifiers for the second dataset is high compared to the first dataset, implying that the stacked ensemble is more efficient with limited data (Tables 11, 12, 13 and 14).

Table 11  Comparison of multi-class classification framework’s specificity with base classifiers on Chest Xray Images PNEUMONIA and Covid-19 dataset [11]

| Model               | Specificity-Normal | Specificity-Pneumonia | Specificity-COVID-19 |
|---------------------|---------------------|-----------------------|----------------------|
| VGG190              | 0.9545              | 0.9741                | 1                    |
| VGG191              | 0.973               | 0.8939                | 0.9947               |
| VGG192              | 0.9698              | 0.9795                | 1                    |
| DenseNet1690        | 0.9697              | 0.9385                | 1                    |
| DenseNet1691        | 0.9698              | 0.949                 | 1                    |
| DenseNet1692        | 0.985               | 0.9596                | 1                    |
| ResNet1010          | 0.9648              | 0.9846                | 0.9949               |
| ResNet1011          | 0.9796              | 0.9594                | 1                    |
| ResNet1012          | **0.995**           | 0.9296                | 1                    |
| WideResNet50 20     | 0.9796              | 0.9596                | 1                    |
| WideResNet50 21     | 0.9798              | 0.9545                | 1                    |
| WideResNet50 22     | 0.9849              | 0.9495                | 1                    |
| Disease diagnosis model | 0.9899     | 0.9648                | 1                    |
| Disease screening model | **0.995** | 1                    | 0.97                |

Bold entries denote highest values for a specific evaluation metric (Column)
### Table 12: Comparison of proposed multi-class classification framework with base classifiers using chest X-ray images pneumonia and COVID-19 [11]

| Model                | Precision | Recall | Accuracy | F1 Score |
|----------------------|-----------|--------|----------|----------|
| VGG19                | 0.9666    | 0.9425 | 0.9533   | 0.9544   |
| VGG19                | 0.9365    | 0.8925 | 0.91     | 0.914    |
| VGG19                | 0.9757    | 0.96   | 0.9667   | 0.9678   |
| DenseNet169          | 0.9573    | 0.9375 | 0.94     | 0.9473   |
| DenseNet169          | 0.9614    | 0.9475 | 0.9467   | 0.9544   |
| DenseNet169          | 0.9734    | 0.9675 | 0.9633   | 0.9704   |
| ResNet101            | 0.9705    | 0.96   | 0.9633   | 0.9652   |
| ResNet101            | 0.9713    | 0.9525 | 0.96     | 0.9618   |
| ResNet101            | 0.9665    | 0.9575 | 0.95     | 0.962    |
| WideResNet50         | 0.9711    | 0.955  | 0.96     | 0.963    |
| WideResNet50         | 0.9687    | 0.9575 | 0.9567   | 0.963    |
| WideResNet50         | 0.9691    | 0.96   | 0.9567   | 0.9645   |
| Disease diagnosis Model | **0.9783** | 0.97  | 0.97     | 0.9742   |
| Disease screening model | 0.9689  | **0.98** | **0.9767** | **0.9744** |

Bold entries denote highest values for a specific evaluation metric (Column)

### Table 13: Confusion Matrix of the Disease Diagnosis Model on the Chest Xray Images PNEUMONIA and Covid-19 dataset [11]

| Predicted Labels | Normal | Pneumonia | COVID 19 | Total |
|------------------|--------|-----------|----------|-------|
| True Labels      |        |           |          |       |
| Normal           | 95     | 5         | 0        | 100   |
| Pneumonia        | 1      | 99        | 0        | 100   |
| COVID 19         | 1      | 2         | 97       | 100   |
| Total            | 97     | 106       | 97       | 300   |

### Table 14: Confusion Matrix of the Disease Screening Model on the Chest Xray Images PNEUMONIA and Covid-19 dataset [11]

| Predicted Labels | Normal | Pneumonia | COVID 19 | Total |
|------------------|--------|-----------|----------|-------|
| True Labels      |        |           |          |       |
| Normal           | 96     | 0         | 4        | 100   |
| Pneumonia        | 0      | 98        | 2        | 100   |
| COVID 19         | 1      | 0         | 99       | 100   |
| Total            | 97     | 98        | 105      | 300   |
Table 15  Precision of the proposed model on COVID-19 Radiography Database [15] under varied threshold

| Threshold | Precision-Normal | Precision-Pneumonia | Precision-COVID-19 |
|-----------|------------------|---------------------|-------------------|
| 0.1       | 0.9041           | 0.9944              | 0.9754            |
| 0.2       | 0.9005           | 0.9944              | 0.985             |
| 0.3       | **0.9434**       | **0.9948**          | **0.9949**        |
| 0.4       | 0.9302           | 0.9948              | 0.9948            |
| 0.5       | 0.9259           | 0.9948              | 0.9948            |
| 0.6       | 0.9091           | 0.9948              | 0.9947            |
| 0.7       | 0.8969           | 0.9948              | 0.9946            |
| 0.8       | 0.8734           | 0.9948              | 0.9944            |
| 0.9       | 0.8621           | 0.9845              | 0.9943            |

Bold entries denote highest values for a specific evaluation metric (Column)

6.2 Evaluation of the multi-class classification framework under varied thresholds

For the COVID-19 Radiography Database [15], with the rise in the threshold, the precision of all the classes initially increases and then decreases, and the recall of COVID-19 decreases, while for the other two classes, the recall increases. This is because as the threshold increased, the number of false negatives for COVID-19 increases, while for the other two classes, the number of false negatives decreases. The F1 score and accuracy are maximum when the threshold is 0.3. Tables 15, 16 and 17 show the experimental results and the results on the first dataset [15] are shown in Fig. 4.

For the Chest X-ray images pneumonia and COVID-19 [11] dataset, the precision of the COVID-19 class remains the same, while that of the other two classes decreases with an increase in the threshold. The recall of the COVID-19 class decreases, while with an increase in threshold, that of the other two classes remains the same. The accuracy and F1 score are maximum at a threshold of 0.1. Tables 18, 19 and 20 show evaluation metrics obtained by varying threshold values for the chest X-ray images pneumonia and COVID-19 [11] are shown in Fig. 5.

Table 16  Recall of the proposed model on COVID-19 Radiography Database [15] under varied threshold

| Threshold | Recall-Normal | Recall-Pneumonia | Recall-COVID-19 |
|-----------|---------------|------------------|----------------|
| 0.1       | 0.99          | 0.885            | **0.99**       |
| 0.2       | 0.995         | 0.89             | 0.985          |
| 0.3       | 1             | **0.95**         | 0.98           |
| 0.4       | 1             | **0.95**         | 0.965          |
| 0.5       | 1             | **0.95**         | 0.96           |
| 0.6       | 1             | **0.95**         | 0.94           |
| 0.7       | 1             | **0.95**         | 0.925          |
| 0.8       | 1             | **0.95**         | 0.895          |
| 0.9       | 1             | **0.95**         | 0.87           |

Bold entries denote highest values for a specific evaluation metric (Column)
Table 17 Performance of the proposed model on COVID-19 Radiography Database [15] under varied thresholds

| Threshold | Precision | Recall | Accuracy | F1 Score |
|-----------|-----------|--------|----------|----------|
| 0.1       | 0.958     | 0.955  | 0.955    | 0.9565   |
| 0.2       | 0.96      | 0.9567 | 0.9567   | 0.9583   |
| 0.3       | **0.9777**| **0.9767**| **0.9767**| **0.9772**|
| 0.4       | 0.9733    | 0.9717 | 0.9717   | 0.9725   |
| 0.5       | 0.9718    | 0.97   | 0.97     | 0.9709   |
| 0.6       | 0.9662    | 0.9633 | 0.9633   | 0.9648   |
| 0.7       | 0.9621    | 0.9583 | 0.9583   | 0.9602   |
| 0.8       | 0.9542    | 0.9483 | 0.9483   | 0.9513   |
| 0.9       | 0.9469    | 0.94   | 0.94     | 0.9435   |

Bold entries denote highest values for a specific evaluation metric (Column)

6.3 Evaluation of the proposed model under varied noise levels

The performance of the Disease Diagnosis Model and the Disease Screening Model on the two datasets was explored under varied noise levels. Results obtained were listed in Tables 21, 22, 23 and 24. As seen from tables from the above mentioned tables the performance decreases with increase in the percent of noise, which is expected. Both models performed well when the noise percentage is less than 2.5%.

7 Explainable AI using LIME

In this study, we used Local Interpretable Model-agnostic Explanations (LIME) [54] to interpret the results produced by our proposed model. LIME provides interpretations for the
| Threshold | Precision-Normal | Precision-Pneumonia | Precision-COVID-19 |
|-----------|------------------|---------------------|-------------------|
| 0.1       | 0.9694           | 0.9519              | 1                 |
| 0.2       | 0.9596           | 0.9519              | 1                 |
| 0.3       | 0.9596           | 0.9519              | 1                 |
| 0.4       | 0.9596           | 0.9519              | 1                 |
| 0.5       | 0.9596           | 0.9519              | 1                 |
| 0.6       | 0.9596           | 0.9519              | 1                 |
| 0.7       | 0.95             | 0.9429              | 1                 |
| 0.8       | 0.95             | 0.934               | 1                 |
| 0.9       | 0.9406           | 0.9252              | 1                 |

Bold entries denote highest values for a specific evaluation metric (Column)

| Threshold | Recall-Normal | Recall-Pneumonia | Recall-COVID-19 |
|-----------|---------------|------------------|----------------|
| 0.1       | 0.95          | 0.99             | 0.98           |
| 0.2       | 0.95          | 0.99             | 0.97           |
| 0.3       | 0.95          | 0.99             | 0.97           |
| 0.4       | 0.95          | 0.99             | 0.97           |
| 0.5       | 0.95          | 0.99             | 0.97           |
| 0.6       | 0.95          | 0.99             | 0.97           |
| 0.7       | 0.95          | 0.99             | 0.97           |
| 0.8       | 0.95          | 0.99             | 0.94           |
| 0.9       | 0.95          | 0.99             | 0.92           |

Bold entries denote highest values for a specific evaluation metric (Column)

| Threshold | Precision | Recall  | Accuracy | F1 Score |
|-----------|-----------|---------|----------|----------|
| 0.1       | 0.9737    | 0.9733  | 0.9733   | 0.9736   |
| 0.2       | 0.9705    | 0.97    | 0.97     | 0.9703   |
| 0.3       | 0.9705    | 0.97    | 0.97     | 0.9703   |
| 0.4       | 0.9705    | 0.97    | 0.97     | 0.9703   |
| 0.5       | 0.9705    | 0.97    | 0.97     | 0.9703   |
| 0.6       | 0.9705    | 0.97    | 0.97     | 0.9703   |
| 0.7       | 0.9643    | 0.9633  | 0.9633   | 0.9638   |
| 0.8       | 0.9613    | 0.96    | 0.96     | 0.9607   |
| 0.9       | 0.9533    | 0.9533  | 0.9533   | 0.9543   |

Bold entries denote highest values for a specific evaluation metric (Column)
results generated by a machine learning model. LIME does not depend on the model and can provide interpretations for the results of any model. LIME helps us understand the regions that our model is looking at while making predictions. LIME initially generates samples of a particular test image by perturbing the pixels of the original image. It will then generate the prediction for each sample of the test image. Next, it computes the weight for the sample of the test image using the cosine distance. Finally, it tries to fit a linear classifier using the generated samples and their prediction. Using the weights of this classifier, LIME then selects the most important features.

The LIME interpretations of a few images belonging to the class pneumonia and COVID-19 are shown in Figs. 6 and 7. In Fig. 6, the regions marked in red are the areas that may contain COVID-19-pneumonia and hence the red regions decrease the probability of the Other Pneumonia class, and the regions marked in green are the areas that increase the probability of Other Pneumonia class.

| Noise Percentage | Precision | Recall | Accuracy | F1 Score |
|------------------|-----------|--------|----------|----------|
| 0                | 0.9836    | 0.97   | 0.9767   | 0.9768   |
| 2.5              | 0.9687    | 0.9513 | 0.96     | 0.9599   |
| 5                | 0.913     | 0.9013 | 0.91     | 0.9071   |
| 10               | 0.8184    | 0.8013 | 0.81     | 0.8097   |
| 30               | 0.6185    | 0.605  | 0.605    | 0.6117   |
Table 22  Performance of the Disease Screening Model on COVID-19 Radiography Database [15] under varied noise levels

| Noise Percentage | Precision | Recall | Accuracy | F1 Score |
|------------------|-----------|--------|----------|----------|
| 0                | 0.9688    | 0.98   | 0.975    | 0.9744   |
| 2.5              | 0.9389    | 0.9425 | 0.9417   | 0.9407   |
| 5                | 0.8834    | 0.8988 | 0.8917   | 0.891    |
| 10               | 0.7973    | 0.8122 | 0.8048   | 0.8047   |
| 30               | 0.5833    | 0.5913 | 0.5833   | 0.5873   |

In Fig. 7, the regions marked in green are the areas that increase the probability of COVID-19-pneumonia, and the regions marked in red are the areas that decrease the probability of the COVID-19-pneumonia. From the images mentioned above, it can be seen that the regions marked in green are part of the lung. Therefore our model was looking at the right areas to make a prediction.

8 Comparison with existing models

Multi-class classification framework’s performance was compared with the existing models [14, 35, 49, 63, 65] and the observation was multi-class classification framework performance was better than the existing models that were evaluated using the same datasets under similar environment conditions as shown in Table 25. The better performance of the proposed model can be ascribed to the following reasons. Unlike the other models or ensembles, our ensemble was generated using a systematic approach using a selection metric. The other two base classifiers compensated for the misclassification made by one base classifier.

Another reason for the better performance was the inclusion of fully connected layers.

The performance of the multi-class classification framework was compared with existing models, and the results are presented in Table 25. The accuracy and F1 score of the proposed model are compared with pre-trained models as depicted in Figs. 8 and 9 respectively.

The experimental results reported in the study should be considered in the light of some limitations. The first limitation is that the margin of difference of the proposed stacked ensemble and the base classifiers is not significantly high as Base classifiers considered in

Table 23  Performance of the Disease Diagnosis Model on the chest X-ray images pneumonia and COVID-19 [11] under varied noise levels

| Noise Percentage | Precision | Recall | Accuracy | F1 Score |
|------------------|-----------|--------|----------|----------|
| 0                | 0.9783    | 0.97   | 0.97     | 0.9742   |
| 2.5              | 0.9353    | 0.9275 | 0.9233   | 0.9314   |
| 5                | 0.8804    | 0.8675 | 0.8667   | 0.8739   |
| 10               | 0.8037    | 0.77   | 0.7733   | 0.7865   |
| 30               | 0.6385    | 0.6025 | 0.6      | 0.6      |
Table 24  Performance of the Disease Screening Model on the chest X-ray images pneumonia and COVID-19 [11] under varied noise levels

| Noise Percentage | Precision | Recall | Accuracy | F1 Score |
|------------------|-----------|--------|----------|----------|
| 0                | 0.9689    | 0.98   | 0.9767   | 0.9744   |
| 2.5              | 0.923     | 0.9375 | 0.93     | 0.9302   |
| 5                | 0.8684    | 0.8825 | 0.8733   | 0.8754   |
| 10               | 0.7886    | 0.7975 | 0.7967   | 0.793    |
| 30               | 0.628     | 0.64   | 0.6367   | 0.634    |

the current study have performed well for the two chest X-ray datasets. Hence, the difference in the performance of base classifiers and ensembles is not significantly different.

The second limitation correspond to the evaluation of the proposed model in terms of parameters like memory requirement, parameters and time of execution.

Fig. 6  Pneumonia Images : Output from LIME
Fig. 7 COVID-19 Images: Output from LIME

9 Conclusion and future work

A multi-class classification framework for disease detection was designed using stacked ensembles of pre-trained models. A stacked ensemble was generated for disease screening mode, and another stacked ensemble was designed for diagnosis mode. Different metrics for

| Model               | Accuracy | F1 Score |
|---------------------|----------|----------|
| CNN Model [63]      | 0.8422   | 0.8421   |
| Automated Model [49]| 0.8702   | 0.8737   |
| COVID NET Model [65]| 0.924    | 0.9      |
| CORONET Model [35]  | 0.95     | 0.956    |
| CVDNET Model [14]   | 0.9669   | 0.9668   |
| Proposed framework  | 0.9767   | 0.9774   |
the selection of base classifiers were designed for the two modes. The proposed multi-class classification disease screening and diagnosis framework were trained and tested using two chest X-ray image datasets that contain three classes of images. The proposed framework has outperformed the base classifiers and existing models on two chest X-ray datasets. The average values of recall and precision for the proposed framework was 98%.

The main objective of the study is to develop deep learning based stacked ensemble that are useful to develop efficient computer aided diagnosis system. Unlike the existing studies, the current study focused on minimising false negatives or false positives based on the context using stacked ensemble of pre-trained models and fully connected layers. We designed selection metrics and systematic method to generate efficient stacked ensemble that minimises false positives or false negatives.

Lightweight deep learning architectures can be used as base models to reduce the computational time and resources. Federated learning can be used to preserve the privacy of different data sets and to reduce the local computations. The proposed framework can be
further extended for the classification of multiple lung diseases that can be detected using chest X-rays and CT scans. Based on the characteristics of the disease, suitable metrics can be selected to generate corresponding ensemble.

**Funding** This study was not funded by any of the grants.

**Declarations**

**Ethics approval** This article does not contain any studies with human participants or animals performed by any of the authors.

**Conflict of Interests** The authors declare that they have no conflict of interest.

**References**

1. Abbas A, Abdelsamea MM, Gaber MM (2021) Classification of Covid-19 in chest x-ray images using detrac deep convolutional neural network. Appl Intell 51(2):854–864
2. Abdul Salam M, Taha S, Ramadan M (2021) Covid-19 detection using federated machine learning. PLOS ONE 16(6):1–25, 06
3. Ai T, Yang Z, Hou H, Zhan C, Chen C, Lv W, Tao Q, Sun Z, Xia L (2020) Correlation of chest ct and rt-pcr testing in coronavirus disease 2019 (Covid-19) in china: a report of 1014 cases. Radiology, p 200642
4. Ali M, Khan M, Tung NT et al (2018) Segmentation of dental x-Ray images in medical imaging using neutrosophic orthogonal matrices. Expert Syst Appl 91:434–441
5. Alliou H, Mohammed MA, Benneau N, Al-Khatteeb B, Abdulkareem KH, Garcia-Zapirain B, Damasevicius R, Maskeliunas R (2022) A multi-agent deep reinforcement learning approach for enhancement of covid-19 CT image segmentation. J Pers Med 12(2):309
6. Alyasseri ZAA, Al-Betar MA, Doush IA, Awadallah MA, Abasi AK, Makhadmeh MN, Alomari OA, Abdulkareem KH, Adam A, Damasevicius R et al (2021) Review on Covid-19 diagnosis models based on machine learning and deep learning approaches. Expert systems, pe12759
7. Asif S, Yi W, Jin H, Yi T, Si J (2020) Classification of Covid-19 from chest x-Ray images using deep convolutional neural networks. MedRxiv
8. Bai HX, Hsieh B, Xiong Z, Halsey K, Choi JW, Tran TML, Pan I, Shi LB, Wang D-C, Mei J et al (2020) Performance of radiologists in differentiating Covid-19 from viral pneumonia on chest ct. Radiology, p 200823
9. Bassi PRAS, Attux R (2021) A deep convolutional neural network for Covid-19 detection using chest X-rays. Research on Biomedical Engineering, pp 1–10
10. Brunese L, Mercaldo F, Reginelli A, Santone A (2020) Explainable deep learning for pulmonary disease and coronavirus Covid-19 detection from x-Rays. Comput Methods Prog Biomed 196:105608
11. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, McKeown A, Yang G, Wu X, Yan F, Dong J, Prasadaka MK, Pei J, Ting MYL, Zhiu J, Li C, Hewett S, Dong J, Ziyar I, Shi A, Zhang R, Zheng L, Hou R, Shi W, Fu X, Duan Y, Huu VAN, Wen C, Zhang ED, Zhang CL, Li O, Wang X, Singer MA, Sun X, Xu J, Tafreshi A, Lewis MA, Xia H, Zhang K (2018) Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell 172(5):1122–1131.e9. https://doi.org/10.1016/j.cell.2018.02.010. ISSN 0092–8674
12. Che Azemin MZ, Hassan R, Mohd Tamrin MI, Md Ali MA (August 2020) COVID-19 deep learning prediction model using publicly available radiologist-Adjudicated chest x-Ray images as training data: preliminary findings
13. Chowdhury MEH, Rahman T, Khandakar A, Mazhar R, Kadir MA, Mahbub ZB, Islam KR, Khan MS, Iqbal A, Al Emadi N et al (2020) Can ai help in screening viral and Covid-19 pneumonia? IEEE Access 8:132665–132676
14. CVDNet (November 2020) A novel deep learning architecture for detection of coronavirus (Covid-19) from chest x-ray images. Chaos Soliton Fract 140:110245
15. Chowdhury MEH et al (2020) Can ai help in screening viral and covid-19 pneumonia? IEEE Access 8:132665–132676
16. Fang Y, Zhang H, Xie J, Lin M, Ying L, Pang P, Ji W (2020) Sensitivity of chest ct for Covid-19: comparison to rt-pcr. Radiology, p 200432
17. Gupta A, Gupta S, Katarya R et al (2021) Instacovnet-19: a deep learning classification model for the detection of Covid-19 patients using chest x-ray. Appl Soft Comput 99:106859
18. Gupta RK et al (2021) Deep learning based mathematical model for feature extraction to detect corona virus disease using chest x-ray images. International Journal of Uncertainty Fuzziness and Knowledge-Based Systems, pp 921–947
19. He X, Yang X, Zhang S, Zhao J, Zhang Y, Xing E, Xie P (April 2020) Sample-efficient, deep learning for COVID-19 diagnosis based on CT Scans. medRxiv, p 2020.04.13.20063941
20. Himoto Y, Sakata A, Kiriti M, Hiroi T, Kobayashi K-I, Kubo K, Kim H, Nishimoto A, Maeda C, Kawamura A et al (2020) Diagnostic performance of chest ct to differentiate Covid-19 pneumonia in non-high-epidemic area in Japan. Japanese Journal of Radiology, p 1
21. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
22. Hussain E, Hasan M, Rahman MA, Lee I, Tamanna T, Parvez MZ (2021) Corodet: a deep learning based classification for Covid-19 detection using chest x-Ray images. Chaos Soliton Fract 142:110495
23. Ibrahim DM, Elshennawy NM, Sarhan AM (2021) Deep-chest: Multi-classification deep learning model for diagnosing Covid-19, pneumonia, and lung cancer chest diseases. Computers in Biology and Medicine, p 104348
24. Ibrahim AU, Ozsoz M, Serte S, Al-Turjman F, Yakoi PS (2021) Pneumonia classification using deep learning from chest x-Ray images during Covid-19. Cognitive Computation, pp 1–13
25. Islam MD, Karray F, Alhajj R, Zeng J (2021) Milon a review on deep learning techniques for the diagnosis of novel coronavirus (covid-19). IEEE Access 9:30551–30572
26. Islam N et al (2021) Thoracic imaging tests for the diagnosis of Covid-19. Cochrane Database of Systematic Reviews (3)
27. Ismael AM, Şengür A (2021) Deep learning approaches for Covid-19 detection based on chest x-Ray images. Expert Syst Appl 164:114054
28. Jangam E, Annavarapu CSR (2021) A stacked ensemble for the detection of covid-19 with high recall and accuracy. Comput Biol Med 135:104608
29. Jangam E, Barreto AAD, Annavarapu CSR (2021) Automatic detection of covid-19 from chest ct scan and chest x-rays images using deep learning, transfer learning and stacking. Applied Intelligence, pp 1–17
30. Kamal KC, Yin Z, Wu M, Wu Z (2021) Evaluation of deep learning-based approaches for Covid-19 classification based on chest x-Ray images. Signal, Image and Video Processing, pp 1–8
31. Kanne JP, Little BP, Chung JH, Elicker BM, Ketai LH (2020) Essentials for radiologists on Covid-19: an update—radiology scientific expert panel
32. Karakanis S, Leontidis G (2021) Lightweight deep learning models for detecting Covid-19 from chest x-Ray images. Comput Biol Med 130:104181
33. Karar ME, Hemdan EE-D, Shouman MA (2021) Cascaded deep learning classifiers for computer-aided diagnosis of Covid-19 and pneumonia diseases in x-Ray scans. Complex Intell Syst 7(1):235–247
34. Khan MA, Alhaisoni M, Tariq U, Hussain N, Majid A, Damaševičius R, Maskeliūnas R (2021) Covid-19 case recognition from chest ct images by deep learning, entropy-controlled firefly optimization, and parallel feature fusion. Sensors 21(21):7286
35. Khan AI, Shah JL, Bhat MM (2020) Coronet: a deep neural network for detection and diagnosis of Covid-19 from chest x-ray images. Comput Methods Prog Biomed 196:105581
36. Kongtao C, Franko K, Sang R (2021) Structured model pruning of convolutional networks on tensor processing units. arXiv:2107.04191
37. Kovács A, Palásti P, Verebb D, Boszik B, Palkó A, Kincses ZT (2020) The sensitivity and specificity of chest ct in the diagnosis of Covid-19. European Radiology, pp 1–6
38. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
39. Lei J, Li J, Li X, Qi X (2020) Ct imaging of the 2019 novel coronavirus (2019-ncov) pneumonia. Radiology 295(1):18–18
40. Liu H, Liu F, Li J, Zhang T, Wang D, Lan W (2020) Clinical and ct imaging features of the Covid-19 pneumonia: Focus on pregnant women and children. Journal of infection
41. Loey M, Manogaran G, Khalifa NEM (May 2020) A deep transfer learning model with classical data augmentation and CGAN to detect COVID-19 from chest CT radiography digital images
42. Lokwani R, Gaikwad A, Kulkarni V, Pant A, Kharat A (June 2020) Automated detection of COVID-19 from CT scans using convolutional neural networks. arXiv:2006.13212 [cs, eess]
43. Long C, Xu H, Shen Q, Zhang X, Fan B, Wang C, Zeng B, Li Z, Li X, Li H (2020) Diagnosis of the coronavirus disease (Covid-19): rrt-pcr or ct?. European journal of radiology, p 108961
44. Mahmud T, Rahman MA, Fattah SA (2020) Covxnet: a multi-dilation convolutional neural network for automatic Covid-19 and other pneumonia detection from chest x-ray images with transferable multi-receptive feature optimization. Comput Biol Med 122:103869
45. Muhammad G, Hossain MS (2021) Covid-19 and non-covid-19 classification using multi-layers fusion from lung ultrasound images. Inform Fusion 72:80–88
46. Mukherjee H, Ghosh S, Dhar A, Obaidullah SM, Santosh KC, Roy K (2020) Deep neural network to detect Covid-19: one architecture for both ct scans and chest x-Rays. Appl Intell 51(5):2777–2789
47. Mukherjee H, Ghosh S, Dhar A, Obaidullah SM, Santosh KC, Roy K (2021) Shallow convolutional neural network for Covid-19 outbreak screening using chest x-Rays. Cognitive Computation
48. Nishio M, Noguchi S, Matsuo H, Murakami T (2020) Automatic classification between Covid-19 pneumonia, non-Covid-19 pneumonia, and the healthy on chest x-Ray image: combination of data augmentation methods. Sci Rep 10(1):1–6
49. Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Acharya UR (2020) Automated detection of Covid-19 cases using deep neural networks with x-Ray images. Comput Biol Med 121:103792
50. Pan SJ, Yang Q (2009) A survey on transfer learning. IEEE Trans Knowledge Data Eng 22(10):1345–1359
51. Polsinelli M, Cinque L, Placidi G (April 2020) A Light CNN for detecting COVID-19 from CT scans of the chest. arXiv:2004.12837 [cs, eess]
52. Rahimzadeh M, Attar A, Sakhaei SM (September 2020) A fully automated deep learning-based network for detecting COVID-19 from a new and large lung ct scan Dataset. medRxiv, page 2020.06.08.20121541
53. Rahimzadeh M, Attar A (2020) A modified deep convolutional neural network for detecting Covid-19 and pneumonia from chest x-Ray images based on the concatenation of sception and resnet50v2. Inform Med Unlocked 19:100360
54. Ribeiro MT, Singh S, Guestrin C (August 2016) Why should i trust you?: Explaining the predictions of any classifier. arXiv:1602.04938 [cs, stat]
55. Sahlol AT, Yousri D, Ewees AA, Al-Qaness MAA, Damasevicius R, Elaziz MA (2020) Covid-19 image classification using deep features and fractional-order marine predators algorithm. Sci Rep 10(1):1–15
56. Sarvamangala DR, Kulkarni RV (2021) Convolutional neural networks in medical image understanding: A survey. Evolutionary intelligence, pp 1–22
57. Sesmero MP, Ledezma AI, Sanchis A (2015) Generating ensembles of heterogeneous classifiers using stacked generalization. Wiley Interdiscip Rev: Data Min Knowl Discov 5(1):21–34
58. Sharif Razavian A, Azizpour H, Sullivan J, Carlsson S (2014) Cnn features off-the-shelf: an astounding baseline for recognition
59. Shaban WM, Rabie AH, Saleh AI, Abo-Elsoud MA (October 2020) A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier. Knowl-Based Syst 205:106270
60. Shelke A, Inamdar M, Shah V, Tiwari A, Hussain A, Chafekar T, Mehendale N (2020) Chest x-ray classification using deep learning for automated Covid-19 screening. medRxiv
61. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
62. Taresh MM, Zhu N, Ali TAA, Hameed AS, Mutar ML (2021) Transfer learning to detect COVID-19 automatically from X-ray images, using convolutional neural networks. International Journal of Biomedical Imaging, 2021
63. Toraman S, Alakus TB, Turkoglu I (2020) Convolutional capsnet: a novel artificial neural network approach to detect covid-19 disease from x-ray images using capsule networks. Chaos Soliton Fract 140:110122
64. Trivedi M, Gupta A (2022) A lightweight deep learning architecture for the automatic detection of pneumonia using chest x-Ray images. Multimed Tools Appl 81(4):5515–5536
65. Wang L, Wong A (May 2020) COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest x-ray images. arXiv:2003.09871 [cs, eess]
66. Wang S-H, Govindaraj VV, Górriz JM, Zhang X, Zhang Y-D (2021) Covid-19 classification by fcsnet with deep feature fusion from graph convolutional network and convolutional neural network. Inform Fusion 67:208–229
67. Wang Z, Liu Q, Qi D (2020) Contrastive cross-site learning with redesigned net for COVID-19 CT classification. IEEE Journal of Biomedical and Health Informatics, pp 1–1
68. Wang S-H, Nayak DR, Guttery DS, Zhang X, Zhang Y-D (2021) Covid-19 classification by ccsnet with deep fusion using transfer learning and discriminant correlation analysis. Inform Fusion 68:131–148
69. Wang H, Ros AC, Basavanahally AN, Gilmore HL, Shih N, Feldman M, Tomszewski J, Gonzalez F, Madabhushi A (2014) Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. J Med Imaging 1(3):034003
70. Wu Z, Shen C, Van Den Hengel A (2019) Wider or deeper: Revisiting the resnet model for visual recognition. Pattern Recogn 90:119–133
71. Xie X, Zhong Z, Zhao W, Zheng C, Wang F, Liu J (2020) Chest CT for typical 2019-nCoV pneumonia: relationship to negative RT-PCR testing. Radiology
72. Yadav SS, Bendre MR, Vikhe PS, JadHAV SM (August 2020) Analysis of deep machine learning algorithms in COVID-19 disease diagnosis. arXiv:2008.11639 [cs, eess]
73. Zagoruyko S, Komodakis N (2016) Wide residual networks. arXiv:1605.07146
74. Zebin T, Rezvy S (2021) Covid-19 detection and disease progression visualization: Deep learning on chest x-Rays for classification and coarse localization. Appl Intell 51(2):1010–1021

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.