Computers in Biology and Medicine 135 (2021) 104608

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

A stacked ensemble for the detection of COVID-19 with high recall and accuracy

Ebenezer Jangam [a,b], Chandra Sekhara Rao Annavarapu [b,*]

[a] *Department of Information Technology, VRSiddhartha Engineering College, Vijayawada, Andhra Pradesh, India*
[b] *Department of Computer Science and Engineering, Indian Institute of Technology (ISM), Dhanbad, Jharkhand, India*

ARTICLE INFO

Keywords:
COVID-19
False negatives
Transfer learning
Recall

ABSTRACT

The main challenges for the automatic detection of the coronavirus disease (COVID-19) from computed tomography (CT) scans of an individual are: a lack of large datasets, ambiguity in the characteristics of COVID-19 and the detection techniques having low sensitivity (or recall). Hence, developing diagnostic techniques with high recall and automatic feature extraction using the available data are crucial for controlling the spread of COVID-19. This paper proposes a novel stacked ensemble capable of detecting COVID-19 from a patient's chest CT scans with high recall and accuracy. A systematic approach for designing a stacked ensemble from pre-trained computer vision models using transfer learning (TL) is presented. A novel diversity measure that results in the stacked ensemble with high recall and accuracy is proposed. The stacked ensemble proposed in this paper considers four pre-trained computer vision models: the visual geometry group (VGG)-19, residual network (ResNet)-101, densely connected convolutional network (DenseNet)-169 and wide residual network (WideResNet)-50-2. The proposed model was trained and evaluated with three different chest CT scans. As recall is more important than precision, the trade-offs between recall and precision were explored in relevance to COVID-19. The optimal recommended threshold values were found for each dataset.

1. Introduction

According to the World Health Organization (WHO), as of 25 March 2021, globally, there have been 124,535,520 confirmed cases of COVID-19, resulting in 2,738,876 deaths [1]. The most frequently reported COVID-19 manifestations are cough, fever and difficulties with sense of smell and breathing. This virus is detrimental to the public, as many people are unaware that they are infected because they are asymptomatic. Therefore, rapid detection is of the utmost importance for those infected early. Most of the cases are diagnosed using the real-time reverse transcription-polymerase chain reaction (RRT-PCR) method. However, lower recall of the RRT-PCR may result in more false negatives and result in the pandemic's widespread [2–4]. Studies [5–8] have demonstrated that CT scans of symptomatic individuals have significantly high recall when compared to RT-PCR. Hence, using chest CT scans as a preliminary test for COVID-19 detection could be a better way to detect and control the pandemic among symptomatic patients.

The characteristics of COVID-19 vary based on the progress of the disease [9,10]. Furthermore, the diversity of characteristics of COVID-19 have been reported in various studies [4,11,12]; however, the common characteristics reported in the studies were peripheral distribution and ground glass opacities. Ground glass opaque shadows appear during early days [11,13]. In some cases, crazy paving patterns [9] may be visible. As the disease progresses, halo and reverse halo signs start appearing [14] followed by the appearance of lung lesions resembling white lungs [11]. In the next stage, the density of the lesions decreases [15]. Samples of COVID-19-positive CT scans and COVID-19-negative CT scans are shown in Figs. 5 and 6 respectively.

Due to the ambiguity in characteristics of COVID-19, extraction of the relevant features is a challenging task [16,17]. Moreover, most of the characteristics of COVID-19 are similar to other kinds of pneumonia [18]. Hence, the hand-crafted features are not suitable for the diagnosis of the pandemic [19]; however, convolution neural networks (CNNs) have proved their potential in automatic feature extraction for the complex tasks in the past [20,21]. Therefore, deep learning (DL) using CNN is a promising option for the automatic feature extraction in the case of a pandemic like COVID-19.

On the other hand, during the initial phases of COVID-19, the available datasets of chest CT and chest X-ray scans are limited. Due to the limited availability of datasets, controlling the spread of the

[*] Corresponding author.
E-mail address: acsrao@iitism.ac.in (C.S.R. Annavarapu).

https://doi.org/10.1016/j.compbiomed.2021.104608
Received 16 March 2021; Received in revised form 4 June 2021; Accepted 22 June 2021
Available online 30 June 2021

pandemic during the initial stage using DL-based analysis of CT scans, is a challenge. However, using transfer learning (TL) techniques, 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. Hence, the availability of pre-trained CNN models, such as the visual geomentry group (VGG)-19 [22], the residual network (ResNet) [23] and the densely connected convolutional network (DenseNet) [24], are promising options for the detection of COVID-19 from chest CT scans.

The contributions of the paper are as follows:

- A novel stacked ensemble was designed using a systematic approach to detect COVID-19 from chest CT scans with high recall and accuracy.
- A novel diversity measure was designed to select base classifiers such that the stacked ensemble gives high recall and accuracy.
- The performance of the proposed stacked ensemble was evaluated on three different CT-scan datasets.
- The trade-offs between recall and precision were explored at different thresholds to select the optimum threshold for high recall. It is crucial to minimise the number of false negatives in the context of COVID-19 to control the spread of the virus.
- The performance of the proposed stacked ensemble was compared with the baseline models and the existing models.

The paper is organised in the following manner. The basic terms and concepts used in the paper are defined in Section 2. In Section 3, the previous work related to the paper is presented. In Section 4, the proposed stacked ensemble, the diversity measure and the systematic approach to generate the stacked ensemble are presented. Section 5 gives details of the datasets used for experiments and the methodology followed during the experiments. Section 6 presents the results obtained and the corresponding analysis. Section 7 compares the performance of the proposed stacked ensemble with the existing models and Section 8 concludes the paper.

2. Preliminaries

This section explicates the basic concepts used in the paper.

2.1. Deep learning (DL)

DL [25] is a sub-field of machine learning that deals with artificial neural networks hundreds of layers deep and with about a million parameters. The performance of deep neural networks can be improved with more input data. For image related tasks, CNNs [26] are most prevalent. With fewer parameters compared to traditional neural networks, CNNs can achieve better performance in most of the image processing tasks.

2.2. Transfer learning (TL)

TL [27] facilitates a DL model trained on one task to perform another related task. Initialisation weights related to the second task can be carried out using the weights obtained during the training on the first task. The main advantage of TL is the data required for training can be minimised. In other words, if the available data are less, TL becomes a promising option.

Suppose that there are two tasks, A and B and both the tasks take the input in the form of images. If the data available for task B are less than task A, the DL model used to train task A can be fine-tuned for task B, even though it has less training data.

2.3. Stacking

Stacking is the technique used to combine heterogeneous base models to improve performance. There are three phases in the generation of a stacked ensemble. In the first phase, a pool of base classifiers are generated. In the second phase, a set of base classifiers are selected from the pool of base classifiers based on a diversity measure. During the third phase, the aggregation of the predictions made by the base classifiers is performed using a meta classifier.

Given a data set D, which is split into $D_1, D_2, \ldots D_N$. One of the subsets D_i is kept aside for future use. The remaining subsets generate K base classifiers using K learning algorithms. After generating base classifiers, the D_i set generates the meta-classifier. The meta-classifier's training set consists of predictions from K base classifiers over the instances in D_i. The meta-classifier data have K-attributes whose values are the predictions from K base classifiers for each instance in D_i. The process is repeated for N folds $i = 1, 2, \ldots, N$. At the end of the cross-validation process, each example of the training data for the meta-classifier has K-attributes and a target label. Once the data are available for a meta-classifier from all instances of D, any learning algorithm can generate the meta-classifier model. For the classification of a new example, the base classifier produces a vector of predictions used by the meta-classifier to predict the class [28].

2.4. Visual geometry group (VGG)-19

VGG-19 is a trained DL model for classification. It consists of 19 wt layers (16 convolutional layers, three fully connected layers, five Max-Pool layers and one SoftMax layer) and has almost 144 million parameters. The parameters obtained from training on ImageNet were used to solve problems in a variety of areas like computer graphics [29], classification of flowers [30], classification of retinal figures [31], fault diagnosis [32] and histology image classification [32].

2.5. Residual network(ResNet)-101

ResNet 101 [23] is a DL model used for image classification. ResNet-101 consists of 101 layers and approximately 45 million parameters. The uniqueness in ResNet-101 is the presence of skip connections. ResNet has shown promising results when applied in diverse areas using TL. These areas include 3D medical image analysis [33], papaya fruit classification based on maturity status [34], crop pest classification [35], brain image classification [36], seizure type classification [37], sugar beet and volunteer potato classification [38] and COVID-19 detection [39].

2.6. Densely connected convolutional network (DenseNet)-169

DenseNet-169 [24] is a DL model consisting of 169 layers and is trained for image classification. The building blocks of the DenseNet model are dense blocks. The unique property of DenseNet 169 is the connection between each layer in a Dense block and all the subsequent layers in that block. Researchers have used DenseNet to solve the classification tasks related to waste classification [40], multiple sclerosis classification [41], monocular depth estimation [42] and lung nodule classification [43].

2.7. Wide residual network (Wide ResNet)-50-2

Wide ResNet-50-2 [44] is a DL model trained for image classification. The ResNet model was modified at a depth of 50 and a width of 2 to obtain Wide ResNet-50-2 with approximately 69 million parameters. The applications of Wide ResNet using TL were in classifying malicious software [45], classifying plankton [46] and detecting COVID-19 from chest X-ray images [47].

3. Literature review

A variety of DL models were trained and tested to detect COVID-19 from CT scans and chest x-ray images [48–67].

In the early days of the outbreak of COVID-19, the main hurdle was the lack of public datasets to build and evaluate DL models [68,69]. Xu et al. [70] illustrated how chest CT scans and chest X-rays can be used to diagnose COVID-19 using a private dataset. Initially, Yang et al. [71] made their dataset public. The accuracy and F1 score reported by their DL model were 0.89 and 0.90 respectively. Another public dataset was the COVIDx dataset collected by Wang et al. [72]. Their DL model achieved an accuracy of 0.93. The accuracy of the COVIDx dataset was further improved by Farooq and Hafeez [73] as their model reported an accuracy of 0.96. Another dataset was made publicly available by He et al. [48]. They integrated self-supervised learning with TL to reduce the risk of over-fitting and achieved an F1 score of 0.85 and an Area Under Curve (AUC) of 0.94.

With the availability of the public chest X-rays and CT scans, researchers focussed on DL models with high accuracy and low average classification time [74,75]. One such attempt was made by Polsinelli et al. [49], who proposed a light CNN design based on the SqueezeNet model and their model reported an accuracy of 0.83, a recall of 0.85, a specificity of 0.81, a precision of 0.8173, and an F1 score of 0.8333. Loey et al. [50] increased the size of their dataset using classic data augmentation techniques and CGAN. Lokwani et al. [51] identified the site of infection using a 2D segmentation model based on U-Net architecture and achieved a recall of 0.96428 and a specificity of 0.8839.

Feature extraction is another challenge in the detection of COVID-19 from chest CT scan images [76]. Shaban et al. [52] proposed new hybrid feature selection methodology by combining both filter and wrapper feature selection methods. The authors in Ref. [54] constructed their model using two similar levels with different kernel sizes to capture the input chest X-ray images' local and global features. Wang et al. [58] further conducted a separate feature normalisation in latent space. Their model was able to outperform the COVID-Net model. Most of the models extracted the features automatically using DL [43,77–82].

TL was used by the researchers to achieve high accuracy and low computation time [83]. Among AlexNet, VGGNet16, VGGNet19, GoogleNet, and ResNet50, ResNet-50 provided better accuracy in the detection of COVID-19 from CT scan images. Azemin et al. [53] used a DL model based on the ResNet-101 architecture. Their model achieved an AUC of 0.82, a recall of 0.773, a specificity of 0.718, and accuracy of 0.719. Taresh et al. [55] evaluated the performance of different models on their ability to predict COVID-19-positive cases from chest X-ray images correctly. They found that the VGG-16 model had the best performance in overall scores and based-class scores. Yadav et al. [56] evaluated the two pre-trained CNN models, namely, VGG16 and InceptionV3, using data augmentation techniques. The InceptionV3 model achieved the highest classification accuracy of 0.9935 for binary classifications, whereas the VGG16 model achieved the highest accuracy of 0.9884 for multiclass classification. Rahimzadeh et al. [57] proposed a novel method for increasing the classification accuracy of CNNs. They used the ResNet50V2 network and a modified feature selection pyramid network. Their model achieved an accuracy of 0.9849, and their model was able to identify 234 out of 245 patients correctly.

However, the crucial aspect of the detection of COVID-19 from CT scans is the minimisation of false negatives. In this direction, Lokwani et al. [51] developed a method to convert slice level predictions to scan level predictions, which helped them reduce the number of false positives.

To the best of our knowledge, other studies do not consider the aspect of minimisation of false positives. Moreover, models proposed in Refs. [48–58,84–87] used a single dataset for performance evaluation. Attempts were made to use a modified version of the KNN algorithm [52], but the execution time is higher for larger datasets when compared to DL models.

4. Proposed model

This section explains the systematic approach followed to generate the proposed stacked ensemble, the diversity measure and the details of the proposed model.

4.1. Systematic approach for the generation of a stacked ensemble

There are three phases in the generation of stacked ensemble: generation, selection and aggregation.

1. In the generation phase, a pool of base classifiers consisting of models with different architectures were generated from pre-trained models. From each pre-trained model, a set of base classifiers were generated by appending a varying number of fully connected layers. All the generated base classifiers from the different pre-trained models formed the pool of base classifiers. Each base classifier differs from the other classifiers by at least one fully connected layer. Let $C = c_1, c_2, \ldots, c_m$ be the set of pool of base classifiers.
2. In the selection phase, the base classifiers that constituted the stacked ensemble were selected from the pool of base classifiers. Diversity and accuracy are the two metrics commonly used to select the base classifiers of the ensemble; however, in the case of COVID-19, as the focus is on the minimisation of false negatives and hence the recall, the selection metrics are accuracy, recall and diversity. The diversity measure used in the paper is explained in Section 4.2.
3. In the aggregation phase, the weighted average of the outputs of the base classifiers is given as input to the meta classifier.

4.2. Diversity measure

A novel pairwise diversity metric is proposed for the stacked ensemble to mimic the COVID-19 pandemic. Let N be the total number of examples in the validation set and N_1 and N_0 be the total number of positive and negative examples, respectively; hence $N = N_0 + N_1$. Let c_i and c_j be the pair of base classifiers for which the diversity is measured. Let N_1^{ab} be the number of positive examples and a be the value predicted by first classifier c_i. Let b be the value predicted by the second classifier c_j. The diversity in the set of false negatives generated by the two classifiers c_i and c_j can be measured using N_p^{ab} when $a = 0$ and $b = 1$ or $a = 1$ and $b = 0$. If the term $N_p^{01} + N_p^{01}$ is high, the diversity in the pair of base classifiers is high. On the other hand, accuracy and recall of the individual base classifiers should be high. The metric to select the base classifiers that constitute the stacked ensemble is designed with the following requirements.

- Individual base classifiers should have high accuracy. The accuracy of the first classifier c_i is

$$a_i = \frac{(N_0^{00} + N_0^{01} + N_1^{10} + N_1^{11}).}{N}$$

- The individual base classifiers should give high recall. Recall of the first classifier c_i is

$$s_i = \frac{(N_1^{10} + N_1^{11})}{(N_1^{10} + N_1^{11} + N_1^{10} + N_1^{00})}$$

The pair of classifiers should generate a diverse set of false negatives.

$$d_{ij} = \frac{(N_1^{10} + N_1^{01}).}{(N_1^{01} + N_1^{00} + N_101 + N_1^{00})}$$

The metric to select base classifiers with high recall, high accuracy and high diversity is the product

$$a_i s_i \sum_{j=1}^{m} d_{ij}$$

The pool of base classifiers are evaluated with the above-mentioned metric and the three classifiers that give the best values are selected to form the stacked ensemble.

4.3. Proposed stacked ensemble

The proposed stacked ensemble consists of three base classifiers that are diverse, highly sensitive and highly accurate. The first base classifier is comprised of a pre-trained VGG-19 model and one fully connected layer, as shown in Fig. 3. The VGG-19 model maps the input volume of size (3 × 224 × 224) to a column vector consisting of 1000 rows. The fully connected layer converts this column vector into a column vector with as many rows as the number of classes (which is two). The fully connected layer uses a softmax activation function. A dropout layer with a dropout probability of 0.5 is applied between the fully connected layer to prevent the model from over-fitting the training data.

The second base classifier is the pre-trained DenseNet-169 model and two fully connected layers, as shown in Fig. 4. The DenseNet-169 model maps the input volume of size 3 × 224 × 224 to a column vector consisting of 1000 rows, just like the VGG model. The first fully connected layer maps this column vector to a column vector of size 500. The second fully connected layer maps this column vector to a column vector with two rows a column vector with two rows (equal to the number of classes). The first fully connected layer uses a ReLU activation function, while the second fully connected layer uses a softmax activation function. This part also uses a dropout layer with a probability of 0.5.

The third base classifier is the pre-trained ResNet-101 model and one fully connected layer, as shown in Fig. 5.

Finally, the outputs of the three base classifiers are given as an input to the single neuron to get the predicted class, as shown in Fig. 6. This single neuron uses a softmax activation function. This single neuron forms the stacking model, which assigns weights to the outputs of each of the three parts, and based on these weights and the outputs of the three parts, it predicts the output class, i.e. COVID-19-positive or COVID-19-negative. The description of each layer of the model i.e. the input size, output size and number of parameters are shown in Table 1.

The proposed model uses TL so that the model can train faster. The weights of the pre-trained models are fine-tuned to the task at hand, which is to detect COVID-19. The three models are combined using stacking to predict the output class. In this model, the meta-model is a single neuron, which correctly predicts the output class based on the outputs of the three models discussed above.

The proposed stacked ensemble achieves better accuracy and sensitivity than the best performing individual base classifier when the individual base classifiers are heterogeneous. The base classifiers with high accuracy and recall were selected from the pool of classifiers to form the stacked ensemble. Additionally, the base classifiers that generate a diverse set of false negatives were selected to ensure heterogeneity. Let N be the total number of examples and N_1 be the number of positive examples, and N_0 be the number of negative examples. Accuracy and sensitivity of the base classifiers are denoted by $A_1, A_2, \ldots, A_n$ and R_1, R_2, …, R_n respectively. Initially, three best-performing base classifiers p, q, *and* r were selected and arranged in sorted order of individual recall values such that $R_p > R_q > R_r$. This implies that $FN_p < FN_q < FN_r$, where FN_i is the false negatives generated by the ith base classifier. Let FN_e be the number of false negatives generated by the stacked ensemble. Suppose N_1^{abc} denote the number of positive examples for which the prediction of p classifier is a, q classifier is b and r classifier is c. Similarly, N_0^{abc} denotes the number of negative examples for which the prediction of p classifier is a, q classifier is b and r classifier is c. The following assumptions were made based on the selection criteria of the base classifiers. The first assumption is that recall and accuracy of an individual base classifier is greater than 50%, i.e., $R_p > R_q > R_r > 0.5$, which implies that true positives are greater than false negatives for a base classifier. The second assumption is that the base classifiers p,q, and r generate diverse false negatives.

The number of false negatives of the stacked ensemble can be given by the expression $FN_e = N_1^{000} + N_1^{001} + N_1^{010} + N_1^{100}$ and the number of false negatives of the best performing base classifier can be obtained using $FN_p = N_1^{000} + N_1^{001} + N_1^{010} + N_1^{011}$. Considering the difference in false negatives, $FN_p - FN_s = N_1^{011} - N_1^{100} > 0$ since $TP > FP$ for a given base classifier and the base classifiers generate a diverse set of false negatives. Moreover, the term N_1^{100} is nearly zero as the base classifiers generate diverse set of false negatives. Hence, the FN of the stacked ensemble is less than the best performing individual base classifier. Therefore, the recall of the stacked ensemble is greater than the recall of the best-performing base classifier.

The accuracy of the stacked ensemble is proportional to the sum of true positives and true negatives, which is given by the expression $A_e = 1 - (N_1^{000} + N_1^{001} + N_1^{010} + N_1^{100} + N_0^{111} + N_0^{101} + N_0^{110} + N_0^{011})/N$. The accuracy of the individual base classifier p can be obtained using the expression $A_p = 1 - (N_1^{000} + N_1^{001} + N_1^{010} + N_1^{011} + N_0^{111} + N_0^{101} + N_0^{110} + N_0^{100})/N$. The difference between the accuracies of the stacked

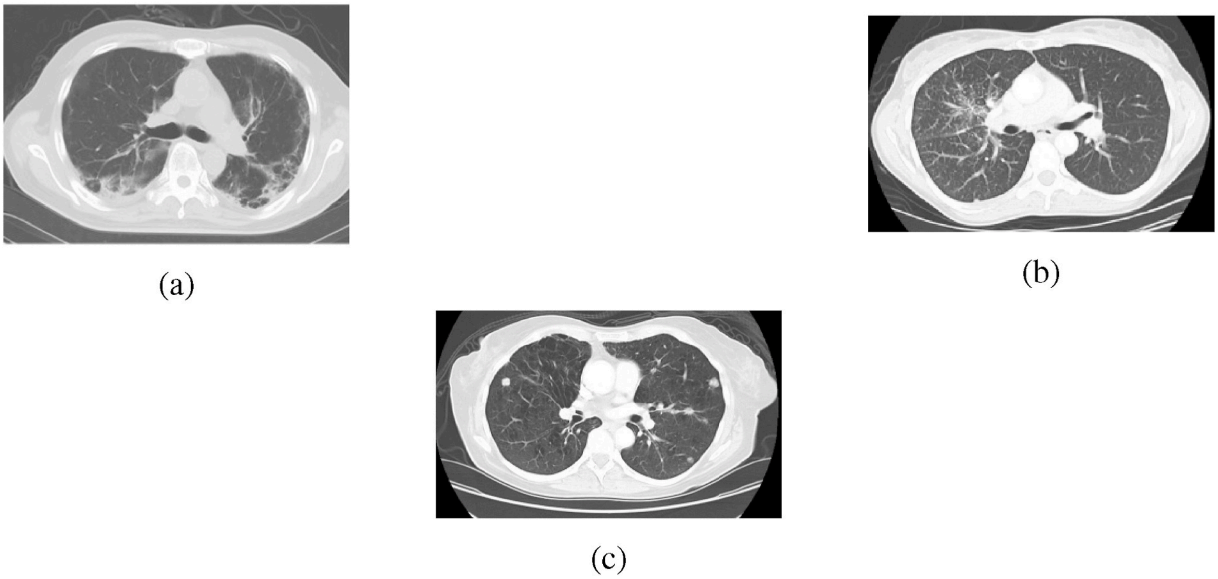
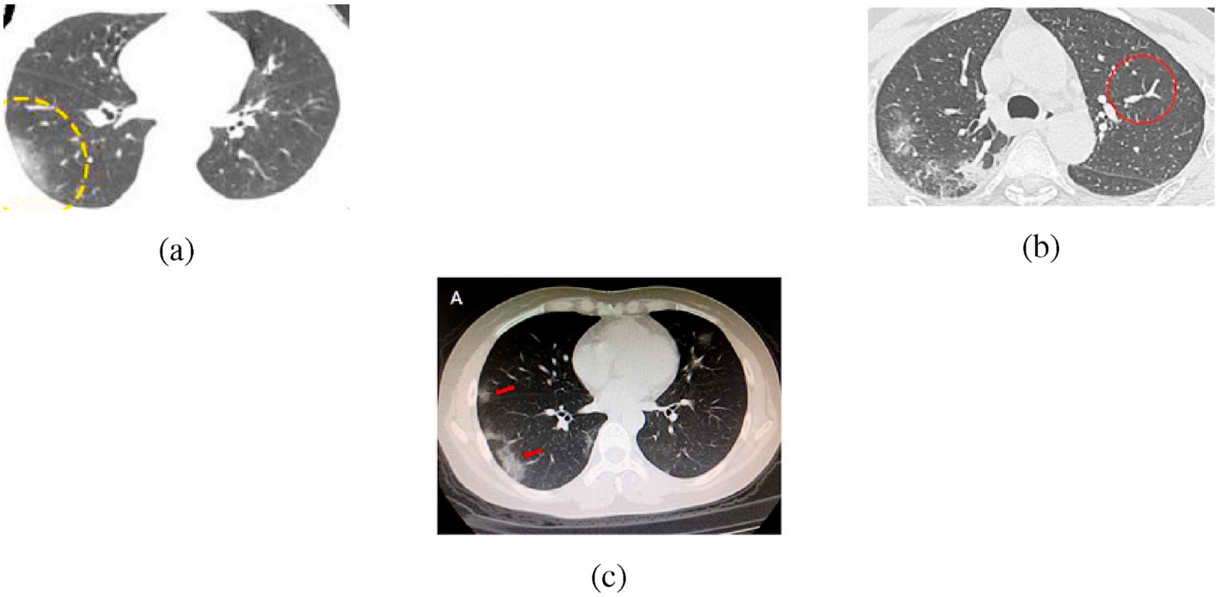


Fig. 1. COVID-19 Negative CT scan images.



(a)

(b)

(c)

Fig. 2. COVID-19 Positive CT scan images.

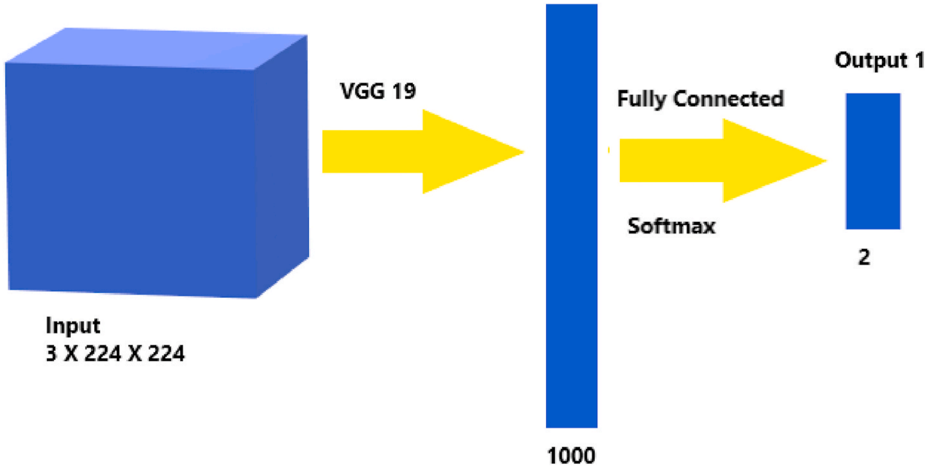


Fig. 3. Part 1 Model architecture.

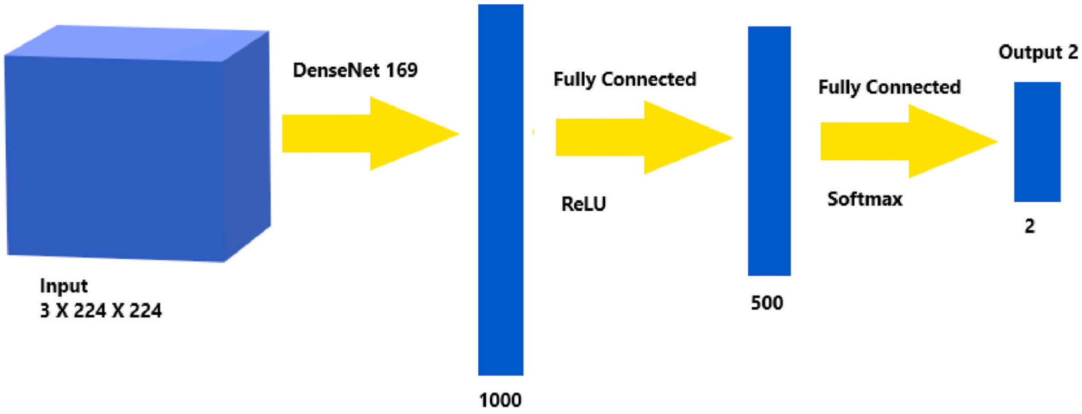


Fig. 4. Part 2 Model architecture.

ensemble and the base classifier is $A_e - A_p = (N_1^{011} + N_0^{100}) - (N_1^{100} + N_0^{011}) > 0$ as the false negatives are less than the true positives and the false positives are less than the true negatives for a given base classifier.

Moreover, the term N_1^{100} is nearly zero as the base classifiers generate diverse set of false negatives. Consequently, the accuracy of the stacked ensemble is greater than the accuracy of the best performing individual

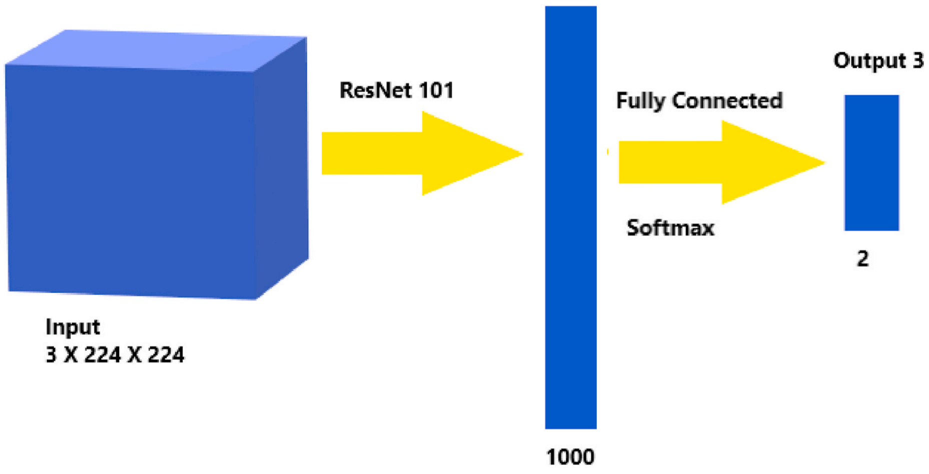


Fig. 5. Part 3 Model architecture.

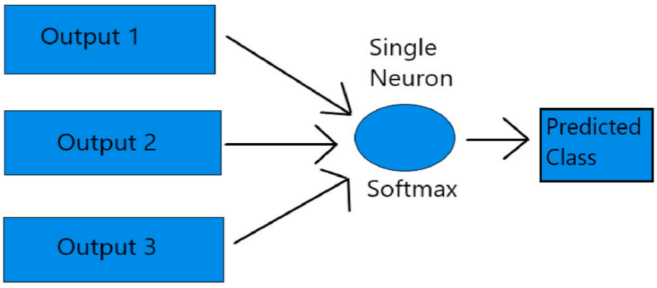


Fig. 6. Combined Model architecture.

Table 1
Description of each layer of the proposed model.

Part No.	Layer Name	Input Size	Output Size	Number of Parameters
1	VGG19	3 X 224 X 224	1000	143667240
1	Fully Connected Layer 1	1000	2	2002
2	DenseNet169	3 X 224 X 224	1000	14149480
2	Fully Connected Layer 1	1000	500	500500
2	Fully Connected Layer 2	500	2	1002
3	ResNet101	3 X 224 X 224	1000	44549160
3	Fully Connected Layer 1	1000	2	2002
–	Single Neuron	3 X 2	1 X 2	3

base classifier.

5. Experimental data and methodology

This section discusses the datasets, experimental methodology and evaluation metrics.

5.1. Datasets

The proposed stacked ensemble was evaluated on three different chest CT scan datasets obtained from different countries.

The following principles were followed in splitting each datset into test, training and validation sets. The number of images in the test set were limited to the range of 200–400 images to validate the model's generality. The size of the validation set is dependant on the size of the test set as the bigger test set gives more prominence to the validation set. All the remaining images were included in the training set. Moreover, to obtain reliable results, test and validation sets were ensured to contain same proportion of positive and negative images.

1. **COVID-CT Dataset** [88]: There are 349 COVID-19 CT images and 397 non-COVID-19 CT images in COVID-CT Dataset. Sample COVID-19-negative images and COVID-19-positive images are shown in Fig. 1 and Fig. 2 respectively. The dataset is available at https://github.com/UCSD-AI4H/COVID-CT.
   - Dataset size: 746 images
   - Number of COVID-19-positive images: 349
   - Number of COVID-19-negative images: 397
   - Train set size: 425 images
   - Validation set size: 118 images
   - Test set size: 203 images

   Data augmentation techniques were used to increase the size of the training set to 1275 images to prevent the model from over-fitting to the training data.

2. **COVID-CTset** [89]: In this original dataset, there are 15,589 and 48, 260 CT scan images belonging to 95 COVID-19 and 282 normal persons respectively. This dataset is from the Negin Medical Center, Sari, Iran and is available at/github.com/mr7495/COVID-CTset.
   - Dataset size (considered): 12,058 images
   - Number of COVID-19-positive images: 2282
   - Number of COVID-19-negative images: 9776
   - Train set size: 11,400 images
   - Validation set size: 258 images
   - Test set size: 400 images
3. **SARS-CoV-2 CT-scan dataset** [90]: This dataset contains 1252 COVID-19 positive CT-scans and 1230 COVID-19 negative CT-scans. The dataset is available at https://www.kaggle.com/plameneduardo/sarscov2-ctscan-dataset and the data are collected from hospitals in Sao Paulo, Brazil.
   - Dataset size: 2482 images
   - Number of COVID-19-positive images: 1252
   - Number of COVID-19-negative images: 1230
   - Train set size: 1800 images
   - Validation set size: 282 images
   - Test set size: 400 images

5.2. Data augmentation techniques

DL models require large datasets for efficient training. If the available datasets are small, the size of the training set can be increased using data augmentation techniques. Among the three datasets used in the study, the COVID-CT Dataset [88] is the smallest. Hence, the following data augmentation techniques were used to increase the size of the dataset.

1. **Random Rotation:** The given image was rotated by an angle chosen randomly.
2. **Random Horizontal Flip:** The given image was randomly flipped horizontally with a given probability.
3. **Colour Jittering:** The brightness, contrast, and saturation of the given image was changed randomly.

5.3. Hyper-parameter tuning

The grid search method was used to tune the values of the hyper-parameters, namely the Random Resized Crop size, the Random Resized Crop scale, the Random Rotation angle range and the Random Horizontal Flip probability. The following are the values considered for each of the hyper-parameters:

- Random Resized Crop size: 128, 200, and 224
- Random Resized Crop Scale: (0.5, 1.0), (1.0, 0.5) and (0.5, 0.5)
- Random Rotation angle: [-3°, 3°], [-5°, 5°], and [-10°, 10°] ranges
- Random Horizontal Flip probability: 0.3, 0.5, and 0.7

The batch size was initialized with a value of four and was doubled until an out of memory error was encountered. The higher the batch size, the more is the memory requirement. Finally, the batch size was chosen as the maximum possible size without getting an out of memory error. The number of epochs were initialized to 50. The number of epochs was incremented by ten until the training and validation sets' accuracy varied. When the number of epochs was 100, the training and validation sets' accuracy and F1 score remain almost constant.

The PyTorch DL framework was used for implementation and the Adam optimiser was used for optimisation. Cross entropy loss was used as a loss function.

- Number of epochs = 100
- Learning rate = 1e-3
- Batch size = 16
- Random Resized Crop size = 224
- Random Resized Crop scale = (0.5, 1.0)
- Random Rotation angle range = [-5°, 5°]
- Random Horizontal Flip probability = 0.5

5.4. Evaluation metrics

The following metrics were used to evaluate the performance of the proposed stacked ensemble.

- **Precision**: Precision is the fraction of positive predictions that actually belong to the positive class.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

- **Recall**: Recall is the fraction of positive examples in the dataset that are predicted as positive.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

- **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.

$$Accuracy = \frac{True\,Positives + True\,Negatives}{True\,Positives + False\,Positives + False\,Negatives + True\,Negatives}$$

6. Experimental results

This section explores the performance evaluation of the proposed stacked ensemble on three datasets of CT scans. In the first subsection, the results obtained for the proposed model were analysed. In the second subsection, the performance of the proposed model was obtained by varying the threshold value by a constant step size.

6.1. Performance analysis of the proposed model

The proposed model was evaluated on three different datasets of chest CT scans: the COVID-CT-Dataset [88], the COVID-CTset [89] and the SARS-CoV-2 CT-scan dataset [90].

The notation used in this study is as follows: i) $Model_0$ denotes the model with a softmax layer ii) $Model_1$ denotes the model with a single fully connected layer and a softmax layer, and iii) $Model_2$ denotes the model with two fully connected layers and a softmax layer. The model can be any of the following DL models namely: *VGG-19*, *ResNet-101*, *DenseNet-169* and *WideResNet-50-2*.

The proposed model consists of three parts, and each part consists of a pre-trained model followed by fully connected layers. Two parts contain a pre-trained model followed by one fully connected layer, whereas one part contains a pre-trained model with two fully connected layers. After eliminating the duplicates from the combinations formed by interchanging the pre-trained models, only two distinct combinations remained (apart from the proposed model). Combination 1 was obtained by interchanging the pre-trained models of part 1 and part 2, and Combination 2 was obtained by interchanging the pre-trained models of part 2 and part 3. Therefore, the two distinct ensembles resulted by interchanging the pre-trained models of the proposed stacked ensemble are: i) Ensemble 1, named as Combination 1, was designed using $DenseNet169_1$, $VGG19_2$, $ResNet101_1$, ii) Ensemble 2, named as Combination 2, was designed using $DenseNet169_1$, $ResNet101_2$, $VGG19_1$.

The COVID-CT Dataset [88] comprised of 746 images, was split into a training set consisting of 425 images, a validation set consisting of 118 images, and a test set consisting of 203 images. The test set contains 98 COVID-19-positive images and 105 COVID-19-negative images. The model could correctly classify 93 images of the 98 COVID-19-positive images resulting in seven false negatives. The proposed model could correctly classify 79 images Of the 105 COVID-19-negative images. Therefore, the accuracy and F1 score of the proposed model are 0.8473 and 0.8571, respectively. The experiment results for the proposed model, combinations, and various DL models were presented in Table 2.

The following are the observations of the experiment conducted on the COVID-CT Dataset [88]. It is evident from Table 2 that the recall of the proposed model is significantly higher than all base classifiers and ensembles. The combination 1 and combination 2 have achieved a recall slightly better than base classifiers; however, the proposed model has outperformed the base classifiers and combinations by a significant margin. The accuracy and F1 score of the proposed model were significantly higher than base classifiers and ensembles. The precision of the proposed model is better than the ensembles or combinations. The only exception is that the precision of $DenseNet169_2$ was slightly better than the proposed model.

Table 2
Comparison among the proposed model and other baseline models on COVID-CT Dataset [88].

Model	Precision	Recall	Accuracy	F1 Score
*VGG*19$_0$	0.7957	0.7551	0.7882	0.7749
*VGG*19$_1$	0.7431	0.8265	0.7783	0.7826
*VGG*19$_2$	0.7714	0.8265	0.7980	0.798
*DenseNet*169$_0$	0.7889	0.7245	0.7734	0.7553
*DenseNet*169$_1$	**0.8182**	0.8265	0.8276	0.8223
*DenseNet*169$_2$	0.7767	0.8163	0.798	0.796
*ResNet*101$_0$	0.7523	0.8367	0.7882	0.7923
*ResNet*101$_1$	0.8043	0.7551	0.7931	0.7789
*ResNet*101$_2$	0.7117	0.8061	0.7488	0.756
*WideResNet*50 2$_0$	0.7545	0.8469	0.7931	0.7981
*WideResNet*50 2$_1$	0.7196	0.7857	0.7488	0.7512
*WideResNet*50 2$_2$	0.7664	0.8367	0.798	0.8
Combination 1	0.7593	0.8367	0.7931	0.7961
Combination 2	0.7217	0.8469	0.7684	0.7793
Proposed Model	0.7815	**0.949**	**0.8473**	**0.8571**

As a second dataset, COVID-CTset [89] with 12,058 images was considered for this study. The dataset was split into a training set, validation set, and a test set consisting of 11,258 images, 400 images, and 400 images. The proposed model correctly classified 198 images out of the 200 COVID-19-positive images, i.e., the number of false negatives is two. Therefore, the proposed model achieved an accuracy of 0.99 and an F1 score of 0.99. Table 3 summarises the results obtained using the COVID-CTset [89] for different models. It is evident from Table 3 that the recall, accuracy and F1 score of the proposed model are better than base classifiers and ensembles.

The SARS-CoV-2 CT-scan dataset [90] consists of 2482 CT scans, which were split into a training set of 2082 images, a validation set of 200 images, and a test set of 200 images. The test set comprises Of 200 COVID-19-positive images and 200 COVID-19-negative images. The proposed model correctly classified 198 images and misclassified two images, i.e., the number of false negatives is two. Of the 200 COVID-19-negative images, the proposed model correctly classified 189 images, i.e., the number of false positives is 11. Therefore, the accuracy and F1 score of the proposed model are 0.935 and 0.9378, respectively. Table 4 summarises the comparison of the performance of the proposed model against other models on the SARS-CoV-2 CT-scan dataset [90].

It is evident from Table 4 that the proposed model has yielded better recall, accuracy, and F1 score compared to the ensembles. *ResNet*101$_0$ has given almost the same recall as that of the proposed model; however, the accuracy, precision, and F1 score of the proposed model are slightly better than *ResNet*101$_0$.

From Tables 2–4, it can be observed that the proposed model performed better than the combination 1, combination 2 and base classifiers in terms of accuracy and F1 score on all of the datasets.

Table 3
Comparison among the proposed model and other baseline models on COVID-CTset [89].

Model	Precision	Recall	Accuracy	F1 Score
*VGG*19$_0$	0.9899	0.985	0.9875	0.9875
*VGG*19$_1$	0.9747	0.965	0.97	0.9698
*VGG*19$_2$	0.9896	0.95	0.97	0.9694
*DenseNet*169$_0$	0.9845	0.95	0.9675	0.9669
*DenseNet*169$_1$	0.9701	0.975	0.9725	0.9726
*DenseNet*169$_2$	**1**	0.955	0.9775	0.977
*ResNet*101$_0$	0.9848	0.975	0.98	0.9799
*ResNet*101$_1$	0.9745	0.955	0.965	0.9646
*ResNet*101$_2$	0.9845	0.95	0.9675	0.9669
*WideResNet*50 2$_0$	0.9742	0.945	0.96	0.9594
*WideResNet*50 2$_1$	0.9895	0.94	0.965	0.9641
*WideResNet*50 2$_2$	0.9948	0.95	0.9725	0.9719
Combination 1	0.9898	0.975	0.9825	0.9824
Combination 2	0.9898	0.97	0.98	0.9798
Proposed Model	0.99	**0.99**	**0.99**	**0.99**

Table 4
Comparison among the proposed model and other baseline models on SARS-CoV-2 CT scan dataset [90].

Model	Precision	Recall	Accuracy	F1 Score
*VGG*19$_0$	0.8899	0.97	0.925	0.9282
*VGG*19$_1$	0.8812	0.89	0.885	0.8856
*VGG*19$_2$	0.8033	0.98	0.87	0.8829
*DenseNet*169$_0$	0.9216	0.94	0.93	0.9307
*DenseNet*169$_1$	0.8257	0.9	0.855	0.8612
*DenseNet*169$_2$	0.8763	0.85	0.865	0.8629
*ResNet*101$_0$	0.8684	**0.99**	0.92	0.9252
*ResNet*101$_1$	0.8899	0.97	0.925	0.9282
*ResNet*101$_2$	**0.9263**	0.88	0.905	0.9026
*WideResNet*50 2$_0$	0.9038	0.94	0.92	0.9216
*WideResNet*50 2$_1$	0.8807	0.96	0.915	0.9187
*WideResNet*50 2$_2$	0.8687	0.86	0.865	0.8643
Combination 1	0.9135	0.95	0.93	0.9314
Combination 2	0.9126	0.94	0.925	0.9261
Proposed Model	0.8991	0.98	**0.935**	**0.9378**

Table 5
Comparison of the training and testing time for the proposed model.

Dataset	Training time (per batch)	Testing time (per batch)
COVID-CT Dataset [88]	0.6162*s*	0.2035*s*
COVID-CTset [57]	0.5943*s*	0.2029*s*
SARS-CoV-2 CT scan dataset [90]	0.5781*s*	0.2024*s*

Table 5 present the training and testing times of the proposed model per batch for each dataset. The average time for training the model is 0.5962 s/batch and the average time for testing the model is 0.2029 s/batch.

6.2. Evaluation of the proposed model under varied thresholds

The proposed model was evaluated at different thresholds ranging from an initial value of 0.1 to the value of 0.9. Tables 6–8 summarises the performance of the proposed model on different datasets.

The observations for the COVID-CT Dataset [88] are as follows. With the increase in threshold, the number of false negatives increased, and false positives decreased. Consequently, with the increase in threshold, recall decreased, and precision increased. It can be observed that the F1 score and accuracy are maximal when the threshold is 0.5. The values obtained from the experiment at different thresholds are listed in Table 6. The evaluation metrics (precision, recall, accuracy, and F1 score) at different thresholds for the COVID-CT Dataset [88] are shown in Fig. 7.

The observations for the COVID-CTset [89] are as follows. With the increase in threshold, the number of false negatives increased, and false positives decreased. Consequently, with the increase in threshold, recall decreased, and precision increased. It can be observed that the F1 score and accuracy are maximal when the threshold is 0.5 and 0.6. The values

Table 6
Performance of the proposed model on COVID-CT Dataset [88] under varied thresholds.

Threshold	Precision	Recall	Accuracy	F1 Score
0.1	0.6299	**0.9898**	0.7143	0.7698
0.2	0.6761	0.9796	0.7635	0.8
0.3	0.7142	0.9694	0.798	0.8225
0.4	0.744	0.949	0.8177	0.8341
0.5	0.7815	0.949	**0.8473**	**0.8571**
0.6	0.7946	0.9082	0.8424	0.8476
0.7	0.8137	0.8469	0.8325	0.83
0.8	0.8933	0.6837	0.8079	0.7746
0.	**0.9138**	0.5408	0.7537	0.6795

Table 7
Performance of the proposed model on COVID-CTset [89] under varied thresholds.

Threshold	Precision	Recall	Accuracy	F1 Score
0.1	0.939	**1**	0.9675	0.9685
0.2	0.9479	**1**	0.9725	0.9732
0.3	0.966	0.995	0.98	0.9803
0.4	0.9755	0.995	0.985	0.9851
0.5	0.99	0.99	**0.99**	**0.99**
0.6	0.99	0.99	**0.99**	**0.99**
0.7	0.9899	0.98	0.985	0.9849
0.8	**1**	0.965	0.9825	0.9822
0.9	**1**	0.955	0.9775	0.977

Table 8
Performance of the proposed model on SARS-CoV-2 CT scan dataset [90] under varied thresholds.

Threshold	Precision	Recall	Accuracy	F1 Score
0.1	0.6993	**1**	0.785	0.823
0.2	0.7752	**1**	0.885	0.8734
0.3	0.8197	**1**	0.89	0.9009
0.4	0.8696	**1**	0.925	0.9302
0.5	0.8991	0.98	**0.935**	**0.9378**
0.6	0.9135	0.95	0.93	0.9314
0.7	0.9082	0.89	0.9	0.899
0.8	0.9222	0.83	0.88	0.8737
0.9	**0.95**	0.76	0.86	0.8444

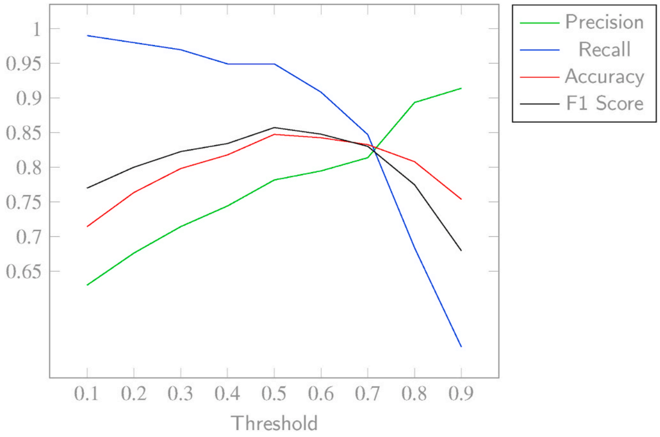


Fig. 7. Variation of Precision, Recall, Accuracy and F1 score with threshold on COVID-CT Dataset [88].

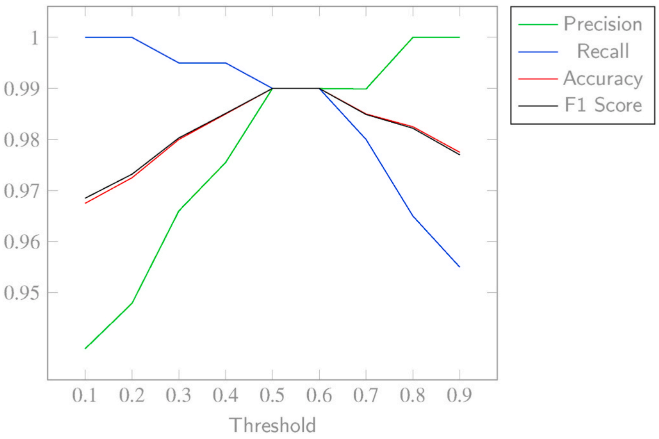


Fig. 8. Variation of Precision, Recall, Accuracy and F1 score with threshold on COVID-CTset [89].

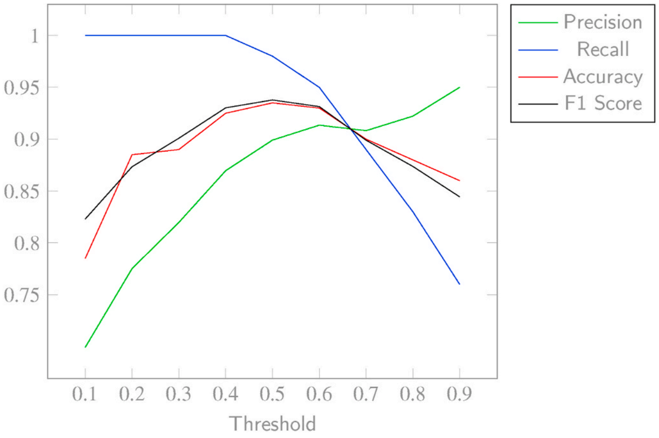


Fig. 9. Variation of Precision, Recall, Accuracy and F1 score with threshold on SARS-CoV-2 CT scan dataset [90].

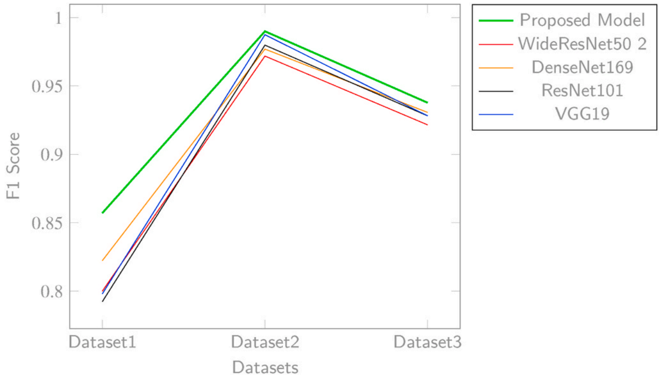


Fig. 10. F1 Score of Different models on different datasets.

obtained from the experiment at different thresholds are listed in Table 7. The evaluation metrics (precision, recall, accuracy, and F1 score) at different thresholds for the COVID-CTset [89] are presented in Fig. 8.

The observations for the SARS-CoV-2 CT-scan dataset [90] are as follows. With the increase in threshold, the number of false negatives increased, and false positives decreased. Consequently, with the increase in threshold, recall decreased, and precision increased. It can be observed that the F1 score and accuracy are maximal when the threshold is 0.5. The values obtained from the experiment at different thresholds are listed in Table 8. The evaluation metrics (precision, recall, accuracy, and F1 score) at different thresholds for the SARS-CoV-2 CT-scan dataset [90] are presented in Fig. 9.

The accuracy and F1 score of the proposed model are compared with individual pre-trained models on the three datsets in Fig. 10 and Fig. 11 respectively. The accuracies of the proposed model and pre-trained models are plotted against varying thresholds in Fig. 12.

The recommended threshold differs for each dataset and depends on the preferred metric according to the application. If high precision is required, the threshold should be higher, whereas if a high recall is preferred, the threshold should be lower. In the detection of COVID-19, minimisation of false negatives is a crucial aspect as false negatives result in the spread of the pandemic. Hence, recall along with accuracy was given preference. It can be observed from Tables 6–8 and Figs. 7–9 that recall increased with a decrease in the threshold, and hence the selection of lower thresholds yields a high recall.

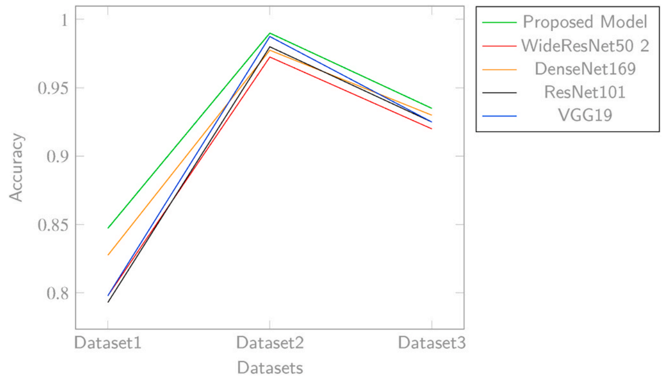


Fig. 11. Accuracy of Different models on different datasets.

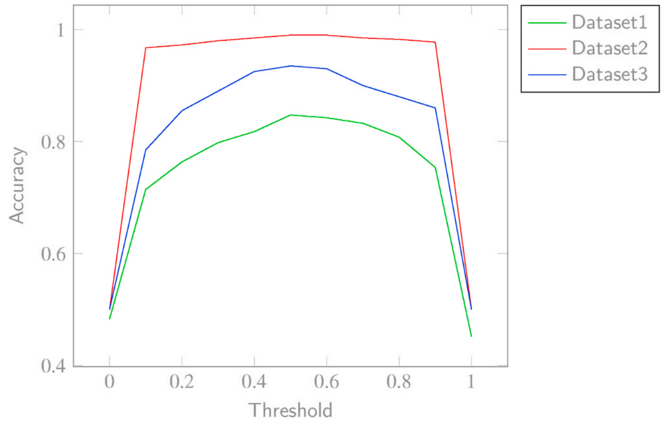


Fig. 12. Variation of Accuracy with threshold for each dataset.

7. Comparison with existing models

The performance of the proposed model was compared with the existing models [48–50,52,57,58] on the respective datasets used for evaluation of the existing models. The models [48–50,52,57,58] were chosen for comparison due to similarity in experiments conducted and the dataset composition. The evaluation metrics used for the comparison were precision, recall, accuracy, and F1 score.

The comparison of the performance of the proposed model and existing models on the COVID-CT dataset [88] are presented in Table 9. The objective of the proposed model is to minimise the false negatives without compromising the accuracy. Table 9 shows that the recall of the proposed model is significantly higher than the existing models due to the minimisation of false negatives. Moreover, the F1 score of the proposed model was better than the existing models. The accuracy of the proposed model is on par with most of the existing models.

Tables 10 and 11 present the comparison of the proposed model and existing models on the datasets COVID-CTset [89] and SARS-CoV-2 CT-scan dataset [90], respectively. It can be observed from Table 10 that the proposed model has performed well on the four metrics compared to the existing model. The observation from Table 11 is that the proposed model's recall is significantly higher than the existing model.

The following are the strengths of the proposed model. The proposed model used a stacked model of different pre-trained models and hence the proposed model could outperform existing models. Moreover, the proposed model could learn more features of COVID-19 as it was trained on three different datasets and has seen more examples than the other models. The proposed model is a stacked ensemble of three different pre-trained models and a varying number of fully connected layers. These additional fully connected layers helped the model learn the features specific to COVID-19, resulting in better performance.

Table 9
Comparison between the proposed model and models proposed in previous research papers on COVID-CT Dataset [88].

Model	Precision	Recall	F1 Score	Accuracy
DL with classical data augmentation and CGAN [50]	**0.85**	0.78	0.81	0.83
Self-Trans approach [48]	–	–	0.85	0.86
SqueezeNet based light CNN [49]	0.82	0.85	0.83	0.83
Enhanced KNN classifier [52]	0.75	0.74	0.75	**0.96**
Redesigned Net for COVID-19 CT Classification [58]	0.78	0.80	0.79	0.79
Proposed Model	0.78	**0.95**	**0.86**	0.85

Table 10
Comparison between the proposed model and models proposed in previous research papers on COVID-CTset [89].

Model	Precision	Recall	F1 Score	Accuracy
A Fully Automated Deep Learning-based Network [57]	0.81	0.95	0.87	0.98
Proposed Model	**0.99**	**0.99**	**0.99**	**0.99**

Table 11
Comparison between the proposed model and models proposed in previous research papers on SARS-CoV-2 CT scan dataset [90].

Model	Precision	Recall	F1 Score	Accuracy
Redesigned Net for COVID-19 CT Classification [58]	**0.96**	0.86	0.91	0.91
Proposed Model	0.90	**0.98**	**0.94**	**0.94**

8. Conclusion

Minimisation of false negatives is vital in controlling the spread of COVID-19. Hence, we proposed a stacked ensemble model of pre-trained models and fully connected layers to detect COVID-19 with high recall and accuracy. The stacked ensemble consisting of VGG-19, DenseNet-169, and ResNet-101 models was generated using a systematic approach and a similarity measure. The proposed stacked ensemble model performed better than the baseline and existing models. Moreover, the proposed model achieved high accuracy and recall on three chest CT-scan datasets. The trade-off between recall and precision was explored to select the recommended threshold for each dataset. For all the three CT scan datasets, the recall of the model was high when the threshold is 0.5, and it increased further by decreasing the threshold. Accuracy and F1 score were maximal when the threshold is 0.5. Hence, the recommended threshold for the three datasets is 0.5.

Declaration of competing interest

There are no known conflicts of interest.

References

[1] Coronavirus update (live): 28,988,031 cases and 925,320 deaths from COVID-19 virus pandemic - worldometer.

[2] Ai Tao, Zhenlu Yang, Hongyan Hou, Chenao Zhan, Chong Chen, Wenzhi Lv, Tao Qian, Ziyong Sun, Liming Xia, Correlation of Chest Ct and Rt-Pcr Testing in Coronavirus Disease 2019 (Covid-19) in china: a Report of 1014 Cases, Radiology (2020) 200642.

[3] Yicheng Fang, Huangqi Zhang, Jicheng Xie, Minjie Lin, Lingjun Ying, Peipei Pang, Wenbin Ji, Sensitivity of Chest Ct for Covid-19: Comparison to Rt-Pcr, Radiology (2020) 200432.

[4] Jeffrey P. Kanne, Brent P. Little, Jonathan H. Chung, Brett M. Elicker, Loren H. Ketai, Essentials for Radiologists on Covid-19: an Update—Radiology Scientific Expert Panel, 2020.

[5] Anita Kovács, Péter Palásti, Dániel Veréb, Bence Bozsik, András Palkó, Zsigmond Tamás Kincses, The sensitivity and specificity of chest ct in the diagnosis of covid-19, Eur. Radiol. (2020) 1–6.

[6] Harrison X. Bai, Ben Hsieh, Xiong Zeng, Kasey Halsey, Whae Choi Ji, Thi My Linh Tran, Ian Pan, Lin-Bo Shi, Dong-Cui Wang, Ji Mei, et al., Performance of Radiologists in Differentiating Covid-19 from Viral Pneumonia on Chest Ct, Radiology (2020) 200823.

[7] Chunqin Long, Huaxiang Xu, Qinglin Shen, Xianghai Zhang, Bing Fan, Chuanhong Wang, Bingliang Zeng, Zicong Li, Xiaofen Li, Honglu Li, Diagnosis of the coronavirus disease (covid-19): rrt-pcr or ct? Eur. J. Radiol. (2020) 108961.

[8] Xingzhi Xie, Zhong Zheng, Wei Zhao, Chao Zheng, Fei Wang, Jun Liu, Chest Ct for Typical 2019-ncov Pneumonia: Relationship to Negative Rt-Pcr Testing, Radiology (2020) 200343.

[9] Ye Zheng, Yun Zhang, Yi Wang, Zixiang Huang, Bin Song, Chest ct manifestations of new coronavirus disease 2019 (covid-19): a pictorial review, Eur. Radiol. (2020) 1–9.

[10] K. Wang, S. Kang, R. Tian, X. Zhang, Y. Wang, Imaging manifestations and diagnostic value of chest ct of coronavirus disease 2019 (covid-19) in the xiaogan area, Clin. Radiol. 75 (5) (2020) 341–347.

[11] Michael Chung, Bernheim Adam, Xueyan Mei, Ning Zhang, Mingqian Huang, Xianjun Zeng, Jiufa Cui, Wenjian Xu, Yang Yang, Zahi A. Fayad, et al., Ct imaging features of 2019 novel coronavirus (2019-ncov), Radiology 295 (1) (2020) 202–207.

[12] Li Xiao, Xu Fang, Yun Bian, Jianping Lu, Comparison of chest ct findings between covid-19 pneumonia and other types of viral pneumonia: a two-center retrospective study, Eur. Radiol. (2020) 1–9.

[13] Li Fan, Li Dong, Huadan Xue, Longjiang Zhang, Zaiyi Liu, Bing Zhang, Lina Zhang, Wenjie Yang, Baojun Xie, Xiaoyi Duan, et al., Progress and prospect on imaging diagnosis of covid-19, Chinese Journal of Academic Radiology (2020) 1–10.

[14] Yan Li, Liming Xia, Coronavirus disease 2019 (covid-19): role of chest ct in diagnosis and management, Am. J. Roentgenol. 214 (6) (2020) 1280–1286.

[15] Feng Pan, Tianhe Ye, Peng Sun, Shan Gui, Bo Liang, Lingli Li, Dandan Zheng, Jiazheng Wang, Richard L. Hesketh, Lian Yang, et al., Time Course of Lung Changes on Chest Ct during Recovery from 2019 Novel Coronavirus (Covid-19) Pneumonia, Radiology (2020).

[16] Huanhuan Liu, Fang Liu, Jinning Li, Tingting Zhang, Dengbin Wang, Weishun Lan, Clinical and ct imaging features of the covid-19 pneumonia: focus on pregnant women and children, J. Infect. 80 (5) (2020) e7–e13.

[17] Junqiang Lei, Junfeng Li, Xun Li, Xiaolong Qi, Ct imaging of the 2019 novel coronavirus (2019-ncov) pneumonia, Radiology 295 (1) (2020), 18–18.

[18] Y. Himoto, A. Sakata, M. Kirita, T. Hiroi, K. Kobayashi, K. Kubo, H. Kim, A. Nishimoto, C. Maeda, A. Kawamura, N. Komiya, Diagnostic performance of chest ct to differentiate covid-19 pneumonia in non-high-epidemic area in Japan, Jpn. J. Radiol. 38 (5) (2020) 400–406.

[19] Mumtaz Ali, Mohsin Khan, Nguyen Thanh Tung, et al., Segmentation of dental x-ray images in medical imaging using neutrosophic orthogonal matrices, Expert Syst. Appl. 91 (2018) 434–441.

[20] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, Stefan Carlsson, Cnn Features Off-The-Shelf: an Astounding Baseline for Recognition, 2014.

[21] Muhammad EH. Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, et al., Can ai help in screening viral and covid-19 pneumonia? IEEE Access 8 (2020) 132665–132676.

[22] Karen Simonyan, Andrew Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, 2014 arXiv preprint arXiv:1409.1556.

[23] Zifeng Wu, Chunhua Shen, Anton Van Den Hengel, Wider or deeper: revisiting the resnet model for visual recognition, Pattern Recogn. 90 (2019) 119–133.

[24] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, Kilian Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[25] Yann LeCun, Yoshua Bengio, Geoffrey Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[26] Saad Albawi, Tareq Abed Mohammed, Saad Al-Zawi, Understanding of a convolutional neural network, in: 2017 International Conference on Engineering and Technology (ICET), Ieee, 2017, pp. 1–6.

[27] Sinno Jialin Pan, Qiang Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2009) 1345–1359.

[28] M Paz Sesmero, I Ledezma Agapito, Araceli Sanchis, Generating ensembles of heterogeneous classifiers using stacked generalization, Wiley interdisciplinary reviews: Data Min. Knowl. Discov. 5 (1) (2015) 21–34.

[29] Tiago Carvalho, Edmar RS. De Rezende, Matheus TP. Alves, Fernanda KC. Balieiro, Ricardo B. Sovat, Exposing computer generated images by eye's region classification via transfer learning of vgg19 cnn, in: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2017, pp. 866–870.

[30] Yong Wu, Qin Xiao, Yonghua Pan, Changan Yuan, Convolution neural network based transfer learning for classification of flowers, in: 2018 IEEE 3rd International Conference on Signal and Image Processing (ICSIP),, IEEE, 2018, pp. 562–566.

[31] Joon Yul Choi, Tae Keun Yoo, Jeong Gi Seo, Jiyong Kwak, Terry Taewoong Um, Tyler Hyungtaek Rim, Multi-categorical deep learning neural network to classify retinal images: a pilot study employing small database, PloS One 12 (11) (2017) e0187336.

[32] Long Wen, X. Li, Xinyu Li, Liang Gao, A new transfer learning based on vgg-19 network for fault diagnosis, in: 2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD), IEEE, 2019, pp. 205–209.

[33] Sihong Chen, Kai Ma, Yefeng Zheng, Med3d: Transfer Learning for 3d Medical Image Analysis, 2019 arXiv preprint arXiv:1904.00625.

[34] Santi Kumari Behera, Amiya Kumar Rath, Prabira Kumar Sethy, Maturity status classification of papaya fruits based on machine learning and transfer learning approach, Inf. Process. Agric. (2020), https://doi.org/10.1016/j.inpa.2020.05.003. In press.

[35] K. Thenmozhi, U. Srinivasulu Reddy, Crop pest classification based on deep convolutional neural network and transfer learning, Comput. Electron. Agric. 164 (2019) 104906.

[36] Taranjit Kaur, Tapan Kumar Gandhi, Deep convolutional neural networks with transfer learning for automated brain image classification, Mach. Vis. Appl. 31 (3) (2020) 1–16.

[37] Shivarudhrappa Raghu, Sriraam Natarajan, Yasin Temel, Shyam Vasudeva Rao, L Kubben Pieter, Eeg based multi-class seizure type classification using convolutional neural network and transfer learning, Neural Network. 124 (2020) 202–212.

[38] Hyun K Suh, Joris Ijsselmuiden, Jan Willem Hofstee, Eldert J van Henten, Transfer learning for the classification of sugar beet and volunteer potato under field conditions, Biosyst. Eng. 174 (2018) 50–65.

[39] Sakshi Ahuja, Bijaya Ketan Panigrahi, Nilanjan Dey, Venkatesan Rajinikanth, Tapan Kumar Gandhi, Deep transfer learning-based automated detection of covid-19 from lung ct scan slices, Appl. Intell. 51 (1) (2021) 571–585.

[40] Guang-Li Huang, He Jing, Zenglin Xu, Guangyan Huang, A combination model based on transfer learning for waste classification, Concurrency Comput. Pract. Ex. 32 (19) (2020), e5751.

[41] Shui-Hua Wang, Yu-Dong Zhang, Densenet-201-based deep neural network with composite learning factor and precomputation for multiple sclerosis classification, ACM Trans. Multimed Comput. Commun. Appl 16 (2s) (2020) 1–19.

[42] Ibraheem Alhashim, Wonka Peter, High Quality Monocular Depth Estimation via Transfer Learning, 2018 arXiv preprint arXiv:1812.11941.

[43] Raul Victor Medeiros da Nóbrega, Solon Alves Peixoto, Suane Pires P da Silva, Pedro Pedrosa Rebouças Filho, Lung nodule classification via deep transfer learning in ct lung images, in: 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS), IEEE, 2018, pp. 244–249.

[44] Sergey Zagoruyko, Nikos Komodakis, Wide residual networks, 2016 arXiv preprint arXiv:1605.07146.

[45] Edmar Rezende, Guilherme Ruppert, Tiago Carvalho, Fabio Ramos, Paulo De Geus, Malicious software classification using transfer learning of resnet-50 deep neural network, in: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2017, pp. 1011–1014.

[46] Alessandra Lumini, Loris Nanni, Deep learning and transfer learning features for plankton classification, Ecol. Inf. 51 (2019) 33–43.

[47] Shervin Minaee, Rahele Kafieh, Milan Sonka, Shakib Yazdani, Ghazaleh Jamalipour Soufi, Deep-covid: predicting covid-19 from chest x-ray images using deep transfer learning, Med. Image Anal. 65 (2020) 101794.

[48] Xuehai He, Xingyi Yang, Shanghang Zhang, Jinyu Zhao, Yichen Zhang, Eric Xing, Pengtao Xie, Sample-Efficient Deep Learning for COVID-19 Diagnosis Based on CT Scans, medRxiv, April 2020, p. 2020, 04.13.20063941.

[49] Matteo Polsinelli, Luigi Cinque, Giuseppe Placidi, A Light CNN for Detecting COVID-19 from CT Scans of the Chest, arXiv:2004.12837 [cs, eess], April 2020. arXiv: 2004.12837.

[50] Loey Mohamed, Gunasekaran Manogaran, Nour Eldeen M. Khalifa, A Deep Transfer Learning Model with Classical Data Augmentation and CGAN to Detect COVID-19 from Chest CT Radiography Digital Images, May 2020.

[51] Rohit Lokwani, Ashrika Gaikwad, Viraj Kulkarni, Aniruddha Pant, and Amit Kharat. Automated Detection of COVID-19 from CT Scans Using Convolutional Neural Networks. arXiv:2006.13212 [cs, eess], June 2020. arXiv: 2006.13212.

[52] I.Saleh Ahmed, Warda M. Shaban, Asmaa H. Rabie, M.A. Abo-Elsoud, A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier, Knowl. Base Syst. 205 (October 2020) 106270.

[53] Mohd Zulfaezal Che Azemin, Radhiana Hassan, Mohd Izzuddin Mohd Tamrin, Mohd Adli Md Ali, COVID-19 Deep Learning Prediction Model Using Publicly Available Radiologist-Adjudicated Chest X-Ray Images as Training Data: Preliminary Findings, August 2020.

[54] CVDNet, A novel deep learning architecture for detection of coronavirus (Covid-19) from chest x-ray images, Chaos, Solit. Fractals 140 (November 2020) 110245.

[55] Mundher Taresh, Ningbo Zhu, Talal Ahmed Ali Ali, Transfer Learning to Detect COVID-19 Automatically from X-Ray Images, Using Convolutional Neural Networks, medRxiv, August 2020, p. 2020, 08.25.20182170.

[56] Samir S. Yadav, Mininath R. Bendre, Pratap S. Vikhe, Shivajirao M. Jadhav, Analysis of Deep Machine Learning Algorithms in COVID-19 Disease Diagnosis, August 2020 arXiv:2008.11639 [cs, eess], arXiv: 2008.11639.

[57] Mohammad Rahimzadeh, Abolfazl Attar, Seyed Mohammad Sakhaei, A Fully Automated Deep Learning-Based Network for Detecting COVID-19 from a New and Large Lung CT Scan Dataset, medRxiv, September 2020, p. 2020, 06.08.20121541.

[58] Zhao Wang, Quande Liu, Dou Qi, Contrastive cross-site learning with redesigned net for COVID-19 CT classification, IEEE Journal of Biomedical and Health Informatics 24 (10) (2020) 2806–2813.

[59] Stefanos Karakanis, Georgios Leontidis, Lightweight deep learning models for detecting covid-19 from chest x-ray images, Comput. Biol. Med. 130 (2021) 104181.

[60] Tulin Ozturk, Muhammed Talo, Eylul Azra Yildirim, Ulas Baran Baloglu, Ozal Yildirim, U Rajendra Acharya, Automated detection of covid-19 cases using deep neural networks with x-ray images, Comput. Biol. Med. 121 (2020) 103792.

[61] Luca Brunese, Francesco Mercaldo, Alfonso Reginelli, Antonella Santone, Explainable deep learning for pulmonary disease and coronavirus covid-19 detection from x-rays, Comput. Methods Progr. Biomed. 196 (2020) 105608.

[62] Tahmina Zebin, Shahadate Rezvy, Covid-19 detection and disease progression visualization: deep learning on chest x-rays for classification and coarse localization, Appl. Intell. 51 (2) (2021) 1010–1021.

[63] Zheng Wang, Ying Xiao, Yong Li, Jie Zhang, Fanggen Lu, Muzhou Hou, Xiaowei Liu, Automatically discriminating and localizing covid-19 from community-acquired pneumonia on chest x-rays, Pattern Recogn. 110 (2021) 107613.

[64] Turker Tuncer, Fatih Ozyurt, Sengul Dogan, Abdulhamit Subasi, A novel covid-19 and pneumonia classification method based on f-transform, Chemometr. Intell. Lab. Syst. 210 (2021) 104256.

[65] Tanvir Mahmud, Md Awsafur Rahman, Shaikh Anowarul Fattah, 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 (2020) 103869.

[66] Anunay Gupta, Shreyansh Gupta, Rahul Katarya, et al., Instacovnet-19: a deep learning classification model for the detection of covid-19 patients using chest x-ray, Appl. Soft Comput. 99 (2021) 106859.

[67] Pedro Ras Bassi, Romis Attux, A deep convolutional neural network for covid-19 detection using chest x-rays, Research on Biomedical Engineering (2021) 1–10.

[68] Mohamed Esmail Karar, Ezz El-Din Hemdan, Marwa A. Shouman, Cascaded deep learning classifiers for computer-aided diagnosis of covid-19 and pneumonia diseases in x-ray scans, Complex & Intelligent Systems 7 (1) (2021) 235–247.

[69] Mizuho Nishio, Shunjiro Noguchi, Hidetoshi Matsuo, Takamichi Murakami, 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) (2020) 1–6.

[70] Xiaowei Xu, Xiangao Jiang, Chunlian Ma, Du Peng, Xukun Li, Shuangzhi Lv, Yu Liang, Yanfei Chen, Junwei Su, Guanjing Lang, Yongtao Li, Hong Zhao, Kaijin Xu, Lingxiang Ruan, Wei Wu, Deep Learning System to Screen Coronavirus Disease 2019 Pneumonia, arXiv:2002.09334 [physics], February 2020, 09334. arXiv: 2002.

[71] Xingyi Yang, Xuehai He, Jinyu Zhao, Yichen Zhang, Shanghang Zhang, and Pengtao Xie. COVID-CT-Dataset: A CT Scan Dataset about COVID-19. arXiv: 2003.13865 [cs, eess, stat], June 2020. arXiv: 2003.13865.

[72] Linda Wang, Alexander Wong, 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], May 2020. arXiv: 2003.09871.

[73] Muhammad Farooq and Abdul Hafeez. COVID-ResNet: A Deep Learning Framework for Screening of COVID19 from Radiographs. arXiv:2003.14395 [cs, eess], March 2020. arXiv: 2003.14395.

[74] K.C. Kamal, Zhendong Yin, Mingyang Wu, Zhilu Wu, Evaluation of deep learning-based approaches for covid-19 classification based on chest x-ray images, Signal, Image and Video Processing (2021) 1–8.

[75] Abdullahi Umar Ibrahim, Mehmet Ozsoz, Sertan Serte, Fadi Al-Turjman, Polycarp Shizawaliyi Yakoi, Pneumonia classification using deep learning from chest x-ray images during covid-19, Cognitive Computation (2021) 1–13.

[76] Sheetal Rajpal, Navin Lakhyani, Ayush Kumar Singh, Rishav Kohli, Naveen Kumar, Using handpicked features in conjunction with resnet-50 for improved detection of covid-19 from chest x-ray images, Chaos, Solit. Fractals 145 (2021) 110749.

[77] Asmaa Abbas, Mohammed M. Abdelsamea, Mohamed Medhat Gaber, Classification of covid-19 in chest x-ray images using detrac deep convolutional neural network, Appl. Intell. 51 (2) (2021) 854–864.

[78] M Ibrahim Dina, Nada M. Elshennawy, Amany M. Sarhan, Deep-chest: multi-classification deep learning model for diagnosing covid-19, pneumonia, and lung cancer chest diseases, Comput. Biol. Med. (2021) 104348.

[79] Sohaib Asif, Wenhui Yi, Hou Jin, Yi Tao, Si Jinhai, Classification of Covid-19 from Chest X-Ray Images Using Deep Convolutional Neural Networks, MedRxiv, 2020.

[80] Morteza Heidari, Seyedehnafiseh Mirniaharikandehei, Abolfazl Zargari Khuzani, Gopichandh Danala, Yuchen Qiu, Bin Zheng, Improving the performance of cnn to predict the likelihood of covid-19 using chest x-ray images with preprocessing algorithms, Int. J. Med. Inf. 144 (2020) 104284.

[81] Mohammad Rahimzadeh, Abolfazl Attar, A modified deep convolutional neural network for detecting covid-19 and pneumonia from chest x-ray images based on the concatenation of xception and resnet50v2, Informatics in Medicine Unlocked 19 (2020) 100360.

[82] Ankita Shelke, Madhura Inamdar, Vruddhi Shah, Amanshu Tiwari, Aafiya Hussain, Talha Chafekar, Ninad Mehendale, Chest X-Ray Classification Using Deep Learning for Automated Covid-19 Screening, medRxiv, 2020.

[83] Rajesh Mehra, et al., Breast cancer histology images classification: training from scratch or transfer learning? ICT Express 4 (4) (2018) 247–254.

[84] Weiqiu Jin, Shuqin Dong, Changzi Dong, Xiaodan Ye, Hybrid ensemble model for differential diagnosis between covid-19 and common viral pneumonia by chest x-ray radiograph, Comput. Biol. Med. 131 (2021) 104252.

[85] M Ismael Aras, Abdulkadir Şengür, Deep learning approaches for covid-19 detection based on chest x-ray images, Expert Syst. Appl. 164 (2021) 114054.

[86] Asif Iqbal Khan, Junaid Latief Shah, Mohammad Mudasir Bhat, Coronet: A deep neural network for detection and diagnosis of covid-19 from chest x-ray images, Comput. Methods Progr. Biomed. 196 (2020) 105581.

[87] Emtiaz Hussain, Mahmudul Hasan, Md Anisur Rahman, Ickjai Lee, Tasmi Tamanna, Mohammad Zavid Parvez, Corodet: a deep learning based classification for covid-19 detection using chest x-ray images, Chaos, Solit. Fractals 142 (2021) 110495.

[88] Jinyu Zhao, Yichen Zhang, Xuehai He, Pengtao Xie, Covid-ct-dataset: a Ct Scan Dataset about Covid-19, 2020 arXiv preprint arXiv:2003.13865.

[89] Mohammad Rahimzadeh, Abolfazl Attar, Seyed Mohammad Sakhaei, A Fully Automated Deep Learning-Based Network for Detecting Covid-19 from a New and Large Lung Ct Scan Dataset, medRxiv, 2020.

[90] Eduardo Soares, Plamen Angelov, Sarah Biaso, Michele Higa Froes, Daniel Kanda Abe, Sars-cov-2 Ct-Scan Dataset: A Large Dataset of Real Patients Ct Scans for Sars-Cov-2 Identification, medRxiv, 2020.