Neural Computing and Applications (2021) 33:8871–8892
https://doi.org/10.1007/s00521-020-05636-6

ORIGINAL ARTICLE

COVIDScreen: explainable deep learning framework for differential diagnosis of COVID-19 using chest X-rays

Rajeev Kumar Singh[1] · **Rohan Pandey**[1] · **Rishie Nandhan Babu**[1]

Received: 11 August 2020 / Accepted: 15 December 2020 / Published online: 8 January 2021
© The Author(s), under exclusive licence to Springer-Verlag London Ltd. part of Springer Nature 2021

Abstract

COVID-19 has emerged as a global crisis with unprecedented socio-economic challenges, jeopardizing our lives and livelihoods for years to come. The unavailability of vaccines for COVID-19 has rendered rapid testing of the population instrumental in order to contain the exponential rise in cases of infection. Shortage of RT-PCR test kits and delays in obtaining test results calls for alternative methods of rapid and reliable diagnosis. In this article, we propose a novel deep learning-based solution using chest X-rays which can help in rapid triaging of COVID-19 patients. The proposed solution uses image enhancement, image segmentation, and employs a modified stacked ensemble model consisting of four CNN base-learners along with Naive Bayes as meta-learner to classify chest X-rays into three classes viz. COVID-19, pneumonia, and normal. An effective pruning strategy as introduced in the proposed framework results in increased model performance, generalizability, and decreased model complexity. We incorporate explainability in our article by using Grad-CAM visualization in order to establish trust in the medical AI system. Furthermore, we evaluate multiple state-of-the-art GAN architectures and their ability to generate realistic synthetic samples of COVID-19 chest X-rays to deal with limited numbers of training samples. The proposed solution significantly outperforms existing methods, with 98.67% accuracy, 0.98 Kappa score, and F-1 scores of 100, 98, and 98 for COVID-19, normal, and pneumonia classes, respectively, on standard datasets. The proposed solution can be used as one element of patient evaluation along with gold-standard clinical and laboratory testing.

Keywords COVID-19 · Chest X-rays · Deep learning · Ensemble learning · ExplainableAI · GANs

1 Introduction

COVID-19 which began from Wuhan, China on Dec 1, 2019, quickly engulfed the entire globe and became one of the first global pandemics in around 100 years, killing 12,31,017 humans and infecting close to 48.5 million people as of 6th November 2020 [23]. With almost 216 countries getting affected and an estimated financial loss of 28 trillion USD over the next five years, it is pertinent that the world must look for a fast and effective solution for large-scale population testing to find the presence of SARS-CoV-2 that causes COVID-19 [65]. Given the enormous human and financial cost, COVID-19 has thrown a big challenge in front of the research community.

The basic reproduction number R_0 represents the average number of people who could be infected by an infected person and this value indicates the speed of disease progression, i.e. the transmissibility of the disease. The severity of COVID-19 can be understood by the fact that the 1918 influenza pandemic which resulted in 50 million deaths worldwide had an average R_0 of 2.7 whereas COVID-19 has an average R_0 of 3.28 [48, 86]. Figure 1 gives an overview of how the COVID-19 cases have increased over the last few months.

Transmission of SARS-CoV-2 can happen through direct, indirect, and close contact with infected persons through infected secretions, respiratory droplets, and other respiratory secretions. Airborne transmission of this virus can occur during the medical procedures that result in the generation of aerosols. Fomite transmission may also

✉ Rajeev Kumar Singh
rajeev.kumar@snu.edu.in

[1] Shiv Nadar University, NCR, Gautam Budh Nagar, India

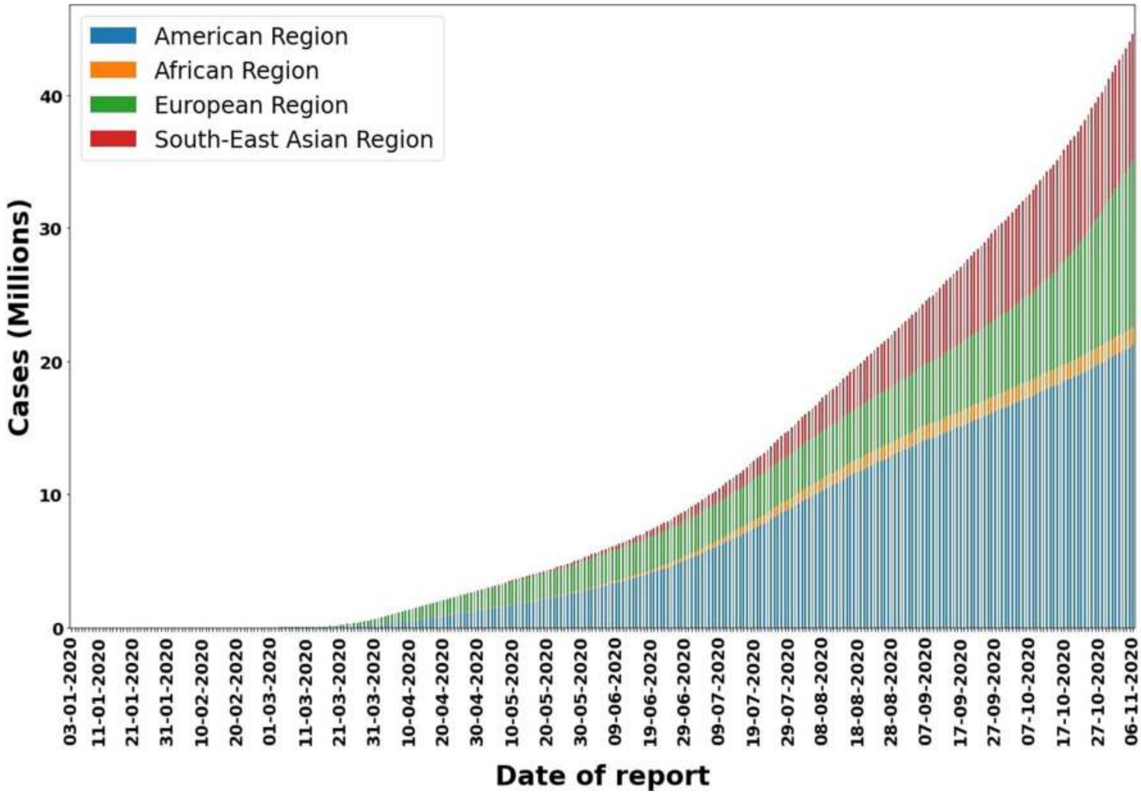


Fig. 1 Number of COVID-19 cases reported worldwide, as per WHO (6th November 2020). These numbers are divided according to WHO regions [23]

happen due to touching of surfaces or objects contaminated with this virus [88]. The ease of transmission of the virus along with the high density of the population in many countries forced nations to shut down industries and restrict the movement of people. Countries follow WHO guidelines of rigorous tracking, contact tracing, rapid diagnosis, and immediate isolation of cases. These approaches face roadblocks due to the gradual reopening of economic activities, wherein social distancing norms are getting eased while testing kits are still in short supply. The incubation period of COVID-19 virus is 5–6 days on average and can be as long as 14–21 days in some cases. This makes the whole process of testing more complicated since one can infect many before they become ill [4, 38, 46, 51, 81, 93]. Modelling study by [32] estimates that many cases are asymptotic and up to 44% of transmission may have already happened before the symptoms appear in the infected person. Thereby, early diagnosis of COVID-19 is crucial for timely referral of the patients to quarantine, rapid incubation of serious cases in specialized hospitals, and containment of the spread of this disease. [95] have demonstrated that the proportion of nosocomial infection in patients affected with SARS-Cov-2 is 44% highlighting the need for testing facilities to be located outside of hospitals. This helps prevent overburdening of hospital resources and reduces the risk of nosocomial transmission to other patients and healthcare warriors.

The unprecedented nature of COVID-19 presents challenges on multiple fronts. Widespread accessibility to testing is critical; however, the high cost of COVID-19 diagnostic tests is a constraint, especially in countries with private health and testing centres. Currently, reverse transcription-polymerase chain reaction, RT-PCR is the gold standard for COVID-19 testing. Serological testing or antibody testing has also been used in certain settings though it is fairly unreliable. RT-PCR testing is a time-consuming process and is currently available in limited supply which is leading to a lower number of people getting tested daily [2, 8]. The test may take up to 2 days to produce results [56]. In this duration, if the resources for isolation of suspected patients are unavailable, they may spread the virus to others, resulting in the proliferation of the virus.

People are hoping for a vaccine to defeat COVID-19; however, traditional vaccine development pathways take on average over 10 years involving stages like R&D, preclinical stage, clinical trials, regulatory review, manufacturing, and quality control [68]. With the combined might of doctors, scientists and policymakers, we might reduce the time of development of COVID-19 vaccine; however, a vaccine that is affordable and accessible to all will still elude us for quite some time [20].

In light of this, many computer scientists entered into innovative partnerships and collaborated with hospitals and doctors to explore other ways of bringing faster diagnosis of COVID-19. Chest X-ray is not recommended for COVID-19 diagnosis and screening; however, WHO has recommended the use of chest X-rays in case RT-PCR is not available or results are delayed [89]. Chest X-rays are less-resource intensive and are associated with lower radiation dose which helps to repeat the test sequentially for monitoring disease progression. The fact that portable devices can be used at the point of care can help minimize the risk of infection while travelling for getting tested. Patients with a high risk of disease progression and associated comorbidities can greatly benefit from this one element of patient evaluation before the final result of RT - PCR is available to the doctors. The triage, allocation, and reallocation of medical resources can be greatly helped by an early warning system which can be achieved through X-ray imaging. It is important to reiterate that chest radiography is not an alternative to clinical testing but an element of patient evaluation that must be corroborated by further tests.

In the last few years, deep learning has grown exponentially and in the medical imaging world, the potential of automated disease discovery framework has been highlighted by many scientists [13, 25, 40, 47, 66, 76]. Considering the success and potential of AI and deep learning in the medical imaging field, many computer scientists are exploring the possibility of automatic detection of COVID-19 using chest X-rays. However, any deep learning-based solution needs sufficient training data to produce generalizable results. The research community has therefore been pooling a lot of data to enhance the knowledge bank which we use for the purpose of this study. Motivated by the recent progress made by the scientific community, we propose to explore the use of chest X-rays for the detection of COVID-19 in this article. It is understood that in any automated disease discovery framework, it is pertinent to have quality images to train the model. We thereby propose to preprocess the image by using noise attenuation and contrast enhancement along with image transformation methods. To remove unwanted annotations, image segmentation has been employed in this work. Generative adversarial networks have been used to create some realistic artificial images to deal with the need for large training data samples. One of the main challenges in the effective use of any deep learning-based solution in the medical context is the black-box nature of such models due to which medical practitioners do not completely understand the logic of a particular machine prediction. To create trust in the medical fraternity, we propose to use an explainable AI technique called Grad-CAM in this article [74].

The main contributions of this paper are:

- Employing preprocessing and segmentation techniques for Chest X-rays enhancement which results in a 6% increase in overall accuracy as compared to the original dataset.
- Evaluation of multiple hypotheses and proposal of an incremental framework to select optimal settings for training deep learning networks detecting COVID-19 cases using chest X-rays. These hypotheses include weight initialization, training class distribution, preprocessing, segmentation, and ensemble learning.
- A novel pruned meta-learning algorithm and framework is proposed addressing the issues of generalizability and model complexity using multiple CNNs as base-learners.
- Qualitatively evaluating the effectiveness of multiple state of the art GAN architectures, and their ability to generate realistic artificial samples for COVID-19 chest X-rays.
- Explainability is built into the proposed model in the form of Grad-CAM visualization to build the confidence and trust of the medical community in using such models.

The article has been organized as given. Section 2 gives a basic introduction to related work in this domain. Section 3 which is named Materials and methods has been divided into multiple subsections. Section 3.1 describes the data sets used for training validation and testing whereas Sect. 3.2 gives a detailed description of the proposed pipeline including image preprocessing, segmentation, and pruned ensemble learning method using CNNs. Section 4 is about experimentation and includes detailed experimentation and results along with appropriate visualizations and implementational details. Section 5 gives a brief overview of the results along with a short and crisp conclusion.

2 Related works

Due to the unexpected rise of coronavirus, there exists a humongous bridge between the existing and the required medical infrastructure, with shortages of essential equipments like PPE kits and lack of qualified doctors and nurses [71]. Over the years, the usage of deep learning methodologies in the medical domain has grown immensely. Evaluation of images by a human expert is tedious, expensive, time-consuming, impractical in many large settings, and introduces inter-observer variability. This has necessitated increased usage of deep learning methods to gain the statistical power for drawing conclusions across a whole patient population bereft of the aforementioned modalities. The development of appropriate algorithms has therefore become a major research focus in medical AI with the potential to deliver objective, reproducible, and scalable approaches to medical imaging tasks. A plethora of computer aided diagnostic systems have come up over the past few years, especially in the detection of multiple chest pathologies using chest X-rays. These work are centred on using convolutional neural networks around the detection of many diseases such as pneumonia, right pleural effusion, cardiomegaly, abnormal mediastinum, pulmonary edema, tuberculosis, etc [5, 6, 36, 70].

Chest radiography can potentially be the first-line imaging modality used for patients with suspected COVID-19 [91]. Chest radiography is a fast and relatively inexpensive imaging modality that is available in many resource-constrained healthcare settings. However, one of the biggest bottlenecks faced is the need for expert radiologists to interpret the radiography images, which may not be available in every setting. Research studies have proven that COVID-19 causes abnormalities that are visible in the chest X-rays and CT images, in the form of ground-glass opacities [41, 43]. The existence of X-Ray laboratories across the globe coupled with reliable computation-based methodologies can potentially ease the pressure on the front-line COVID-19 warriors. [3] evaluated the performance of state-of-the-art convolutional neural networks including MobileNet-v2, VGG-19, Inception, Xception and Inception ResNet-v2 for the detection of COVID-19 from chest X-rays. The work by [84] proposes a SqueezeNet-based architecture tuned for the COVID-19 diagnosis with Bayes optimization along with the validation phase. [87] have proposed a deep convolutional neural network design named COVID-Net using a lightweight residual projection-expansion projection-extension design pattern. The work by [63] proposes a patch-based convolutional neural network approach with a relatively small number of trainable parameters along with statistical analysis of the potential imaging biomarkers of the chest X-rays.

Authors in [9] propose a two-stage classification model where the first stage involves classifying the chest X-ray as belonging to a healthy or non-healthy person with some pulmonary disease. The second stage then finds the presence of COVID-19 caused pneumonia or generic viral pneumonia. The authors used VGG-16 architecture along with transfer learning for this task. The research article [60] aims to build a deep transfer learning-based method for the detection of COVID-19 using Xception which is derived from the standard Inception network architecture. The study from [61] evaluates the effect of five pre-trained CNNs (ResNet50, ResNet101, ResNet152, InceptionV3 and Inception-ResNetV2) for the detection of COVID-19 infected patients from chest X-rays. Authors in [94] propose a combination of anomaly detection, shared feature extractor, and confidence prediction modules called the confidence aware anomaly detection (CAAD) model involving the use of EfficientNet [79]. Through progressive resizing of input images and network fine-tuning, the study in [22] involves the use of pre-trained ResNet-50 architecture for the screening of COVID-19 from chest X-rays called COVID-ResNet. The study from [1] introduces a deep CNN method called DeTraC for the identification of COVID-19 from chest radiographs. DeTraC stands for Decompose, Transfer, and Compose and involves a class decomposition method to handle irregularities in data. Authors in [53] present a deep learning procedure to identify patients with COVID-19 using chest X-rays. Their approach involves a pre-trained variation of CheXNet [70] for disease identification. The research article from [27] involves a variation of the CheXNet [70] deep learning model to build COVID-CXNet which is then used to detect COVID-19-based pneumonia from chest radiographs. The study from [83] introduces an AI framework for the detection of COVID-19 from chest X-rays which used a standard version of the Inception-V3 network architecture pre-trained on ImageNet [21] dataset. The study also introduces attention maps to validate detected regions of interest in chest radiographs.

Many recent studies have thus highlighted the significance of deep learning for the detection of patients with COVID-19 using chest X-rays. The majority of these studies have solely focused on either proposing a new deep learning architecture which is a slight modification of the existing one or exploring the feasibility of standard state-of-the-art architectures for COVID-19 detection. The proposed framework in this paper makes several new contributions to the existing literature. We explore various hypothesis testing to justify decisions taken during CNN model training for the classification task without holding any assumptions which most studies fail to incorporate into their work. This study additionally uses U-Net-based segmentation and the proposed pruned ensemble framework to reduce computational cost and model complexity. In most of the COVID-19 research articles, the emphasis is fundamentally on improving model performance whereas our article endeavors to improve performance while holding the computational expense to the base. This research also addresses the issue of explainability by incorporating Grad-CAM visualization which instills confidence in the medical practitioners to adopt the model in a clinical setting. To the best of our knowledge, this is one of the novel studies that evaluate the usefulness of GANs for the task of image augmentation to improve the model performance for the detection of COVID-19.

3 Materials and methods

3.1 Dataset

Multiple datasets are used in this study for the purpose of classification, segmentation, and weight initialization.

Classification: The datasets used for classification are constructed by using the open data sources provided in Table 1. Several image data repositories have been leveraged in order to gather publicly available COVID-19 Chest X-ray images. Normal and pneumonia samples have been extracted from the open-source NIH chest X-ray dataset used in the RSNA pneumonia detection challenge on Kaggle. Due to the overlap of images in the publically available COVID-19 dataset collections, we provide the number of unique samples in each class of these datasets.

Unbalanced data is a common problem in the image classification task wherein some classes have fewer samples as compared to others. This issue has the potential to make deep CNNs profoundly biased against the less frequent class [39]. In this study, we evaluate the effectiveness of class distribution and thereby create the following dataset splits: as shown in Table 2. Set A dataset split has a balanced distribution of all training classes by undersampling pneumonia and normal class [92], In dataset B we upsample the COVID-19 class using random rotation of 25%, horizontal flipping and Gaussian blur [28]. Set C is imbalanced and we use class weighting to train the network wherein we assign weights of respective proportions conditioned on the initial class sizes while training to avoid model bias [44]. For all datasets, the split is patient-based and samples of patients in the test and validation set have no overlap with the training set at any stage of the study. Each test and validation sample is corresponding to a unique patient.

Segmentation: In order to train the segmentation architecture, we use the Shenzhen and Montgomery County datasets consisting of 662 and 138 chest X-ray samples, respectively. Both the datasets include the manifestation of Tuberculosis and normal cases along with their respective masks [35].

Table 1 All datasets used for the task of COVID-19 classification

Class	Sources	Samples
COVID-19	COVID-19 image data collection [19]	422
	Figure 1 COVID-19 chest X-rays [16]	35
	Actualmed COVID-19 chest X-rays [15]	58
	COVID-19 radiography database [80]	58
Pneumonia	RSNA pneumonia detection challenge [73]	6041
Normal	RSNA pneumonia detection challenge [73]	8851

Weight Initialization: In order to test the effect of weight initialization, we use the CheXpert dataset [34]. CheXpert is a large public dataset for chest radiograph interpretation, consisting of 224,316 chest radiographs of 65,240 patients. We use this dataset to train base learners, and use the trained weights as the initial values for one of the initialization methods in Hypothesis 3.

3.2 Proposed framework

In this study, we evaluate multiple research hypotheses to find optimal model parameters and training methodologies for COVID-19 classification from chest X-ray samples using deep learning models. The proposed paradigm consists of preprocessing, segmentation, and pruned ensemble learning technique as shown in Fig. 2. Detailed explanations for each part is provided in the following subsections.

3.2.1 Deep convolutional neural networks

With the advent of convolutional neural networks, deep learning has been able to effectively outperform existing methodologies for tasks such as segmentation and classification [45, 50, 72, 78]. For most medical imaging tasks, convolutional neural networks are currently state of the art, inspiring us to investigate the efficacy of these for COVID-19 detection using chest X-rays [17, 18, 30, 59].

Thanks to the excellent efficiency of CNN architectures in medical imaging activities, we use the following state of the art standard architectures as the baseline for the proposed pipeline: VGG-19, VGG-16, ResNet-50, DenseNet-161, and DenseNet-169 [31, 33, 75]. The baseline models are truncated at the last fully-connected layer and the following layers have been added as the new head to each of the baseline models: (i) average pooling with 7×7 pool size, (ii) flatten layer, (iii) dense layer with 128 hidden units and reLU activation (iv) dropout layer with 0.5 dropout ratio, and (v) dense layer with 3 hidden units and softmax activation. The input image size for all base learners is 224×224.

A common drawback of these standard architectures is their tendency to over fit the training set. In order to address this drawback, we employ a dropout of 0.5 and L-2 regularization of 1e−3. Stochastic gradient descent, SGD optimizer has been used with initial learning rate of 1e−4, and a momentum of 0.95. The categorical cross-entropy loss function is used for training the baseline models, which is widely used for multi-class classification tasks [24]:

Table 2 Dataset splits used for classification task in this study

Dataset	Class	Train	Validation	Test	Description
A	COVID-19	473	50	50	Balanced (downsampled)
	Normal	473	50	50	
	Pneumonia	473	50	50	
B	COVID-19	1419	50	50	Balanced (upsampled)
	Normal	1419	50	50	
	Pneumonia	1419	50	50	
C	COVID-19	473	50	50	Imbalanced
	Normal	1500	50	50	
	Pneumonia	1500	50	50	

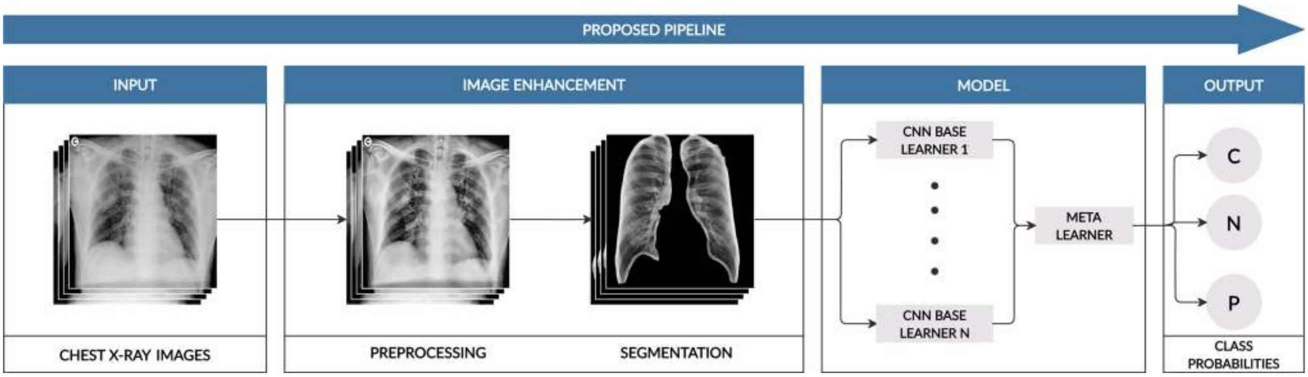


Fig. 2 The proposed pipeline works as follows: the input images (chest X-rays) are passed through the preprocessing stage followed by segmentation. These images are then fed simultaneously into multiple base learners to generate class probabilities which are then passed into the meta-learner to predict final class labels belonging to one of three classes: COVID-19 (C), Normal (N), and Pneumonia (P)

$$\text{Loss} = -\sum_{i=1}^{N} y_i \cdot \log \hat{y}_i \tag{1}$$

here $\hat{y}_i$ is the i-th scalar value in the model output, y_i is the corresponding target value, and N is the number of scalar values in the model output.

The adaptive learning rate has been used for SGD optimizer using a learning rate scheduler which reduces the learning rate to half if the validation accuracy does not improve for 10 epochs. Grid search has been used to optimize the following hyperparameters: (i) initial learning rate of optimizer, (ii) momentum of the optimizer (iii) dropout ratio, and (iv) L-2 regularization. Grid search builds a model for every combination of hyperparameters specified and evaluates each model accordingly. The search ranges were $[1e{-}15, 1e{-}1]$, $[0.85, 0.99]$, $[0.1, 0.8]$, and $[1e{-}10, 1e{-}3]$, respectively. Model checkpoints have been used to save the best weights of the models which were further used in the pipeline.

3.2.2 Image preprocessing

Visual analysis of the dataset showed that a majority of the chest X-rays are either over-exposed or under-exposed and noisy at the time of capture, which can severely impact a clear understanding of the medical problem it depicts. Consequently, there is a compelling need for preliminary image enhancement techniques such as histogram equalization for contrast correction and image filtering methods for denoising. Moreover, multiple studies prove that image preprocessing is significant in standardizing the dataset and thereby resulting in superior performance [7, 37].

The first step in the proposed preprocessing pipeline includes a variant of histogram equalization referred to as CLAHE, contrast limited adaptive histogram equalization, which is frequently used to enhance different types of medical images [67]. This technique effectively spreads out the most frequent intensity values in the images. The more common histogram equalization technique considers the global contrast of the image which can sometimes lead to a loss of information due to over-brightness [77]. CLAHE acts as an alternative which divides the image into smaller blocks called “tiles” and each of these tiles are histogram equalized to confine the spread of intensity values to that particular region using the general histogram equalization formula:

$$h(v) = \text{round}\left(\frac{\text{CDF}(v) - \text{CDF}_{\min}}{(M \times N) - \text{CDF}_{\min}} \times (L - 1)\right) \tag{2}$$

here $\mathrm{CDF}_{\min}$ is the minimum non-zero value of the cumulative distribution function of the pixel intensities, $M \times N$ gives the chosen tile's number of pixels where M denotes the width and N denotes the height. L is the number of grey levels which is set to 256 in this study. However, there is a possibility of noise being confined in a small area that could get amplified. To prevent this, contrast limiting is applied. If any histogram bin is above the specified contrast limit, those pixels are clipped and distributed uniformly to other bins. The *clipLimit* threshold for contrast limiting is tested for different values during experimentation and has been empirically set at 2.0 owing to its improved performance as observed through manual inspection of sample images. Post equalization, bilinear interpolation is applied to remove possible artifacts at the tile borders. Figure 3 suitably demonstrates the impact of CLAHE by showing the histogram distribution of intensity values using a COVID-19 X-ray sample.

To prepare images for further processing such as segmentation and classification, certain image denoising filters capable of removing a significant amount of noise are desirable. We have thus applied a more advanced and dynamic image filtering technique called NLMD, non-local means denoising. This technique can potentially result in much greater post-filtering clarity and lower loss of information in the image compared to local mean algorithms [10, 11].

Noise is largely treated as a random variable with zero mean. Thus a noisy pixel is represented as $p = p_0 + \eta$ where p_0 is the true value of pixel and η is iid zero means Gaussian noise with unknown variance, $\eta \sim \mathcal{N}(0,\sigma^2)$. Averaging of similar pixels from different images should give $p = p_0$ which is the true value of the pixel. But, there is sometimes only one noisy image and no more of its kind. Therefore, instead of seeking similar pixels from different images, we consider a small window in the image and use a fixed sliding window across the image to look for similar patches in the same picture. It is highly probable that a similar patch is found in a small neighbourhood around it. So, we take a pixel and a small window, scan the image for similar windows, average all the windows, and substitute the normal pixel with the average. Although it consumes more time than other blurring techniques, its results are very promising as verified through manual inspection of

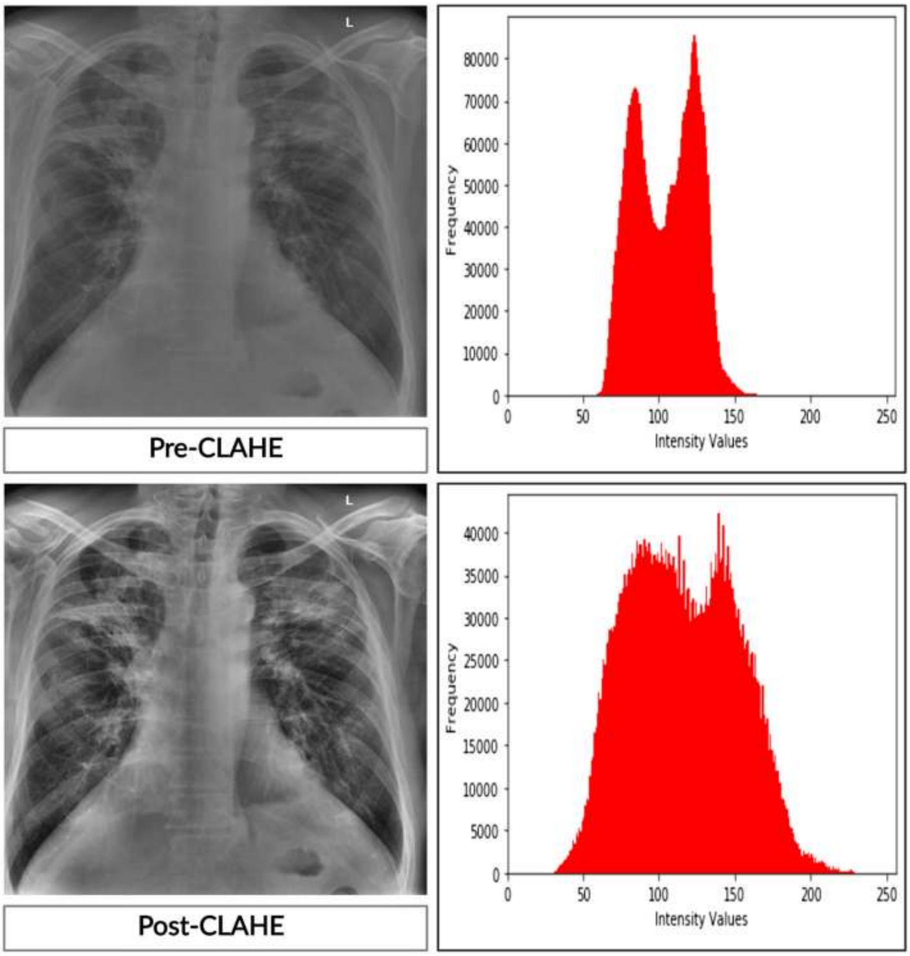


Fig. 3 Sample images of COVID-19 pre and post CLAHE along with histogram

image samples [49]. This non-local means filter is characterized by the following function [11]:

$$\mathrm{NL}_u(p) = \frac{1}{C(p)} \int f(d(B(p), B(q)))u(q)\mathrm{d}q \tag{3}$$

here, $d(B(p), B(q))$ is an Euclidean distance between image patches centered, respectively at p and q, f is a decreasing function and $C(p)$ is the normalizing factor.

The parameters involved in the *NLMD* method include *templateWindowSize*, defined as the size in pixels of the template patch that is used to compute weights, *searchWindowSize*, defined as the size in pixels of the window that is used to compute the weighted average for a given pixel and h, regulates the filter strength having a tradeoff between the removal of noise and image detail. The best-fitting parameters have been empirically found to have $templateWindowSize = 7$, $searchWindowSize = 21$ and $h = 7$ post experimentation with different settings and from manual inspection of image samples. The original image and the enhanced image for each class are shown in Fig. 4.

3.2.3 Image segmentation

Image segmentation has been used extensively in multiple medical imaging tasks and has a twofold benefit of superior model performance and reduced computational cost [12, 18]. In this study, the dataset used for training has been derived from various data sources as mentioned in Sect. 3.1.

A qualitative exploration of data elicits that the image widths and heights across the classes are not equal and these differences lead to wide and asymmetric distribution of image areas. Moreover, there are multiple instances of possible erroneous visual indicators outside the region of interest, ROI such as markings and annotations on the chest X-rays. Thus, it is essential to select relevant image area, in this case, the left and right lung areas as ROI which contain

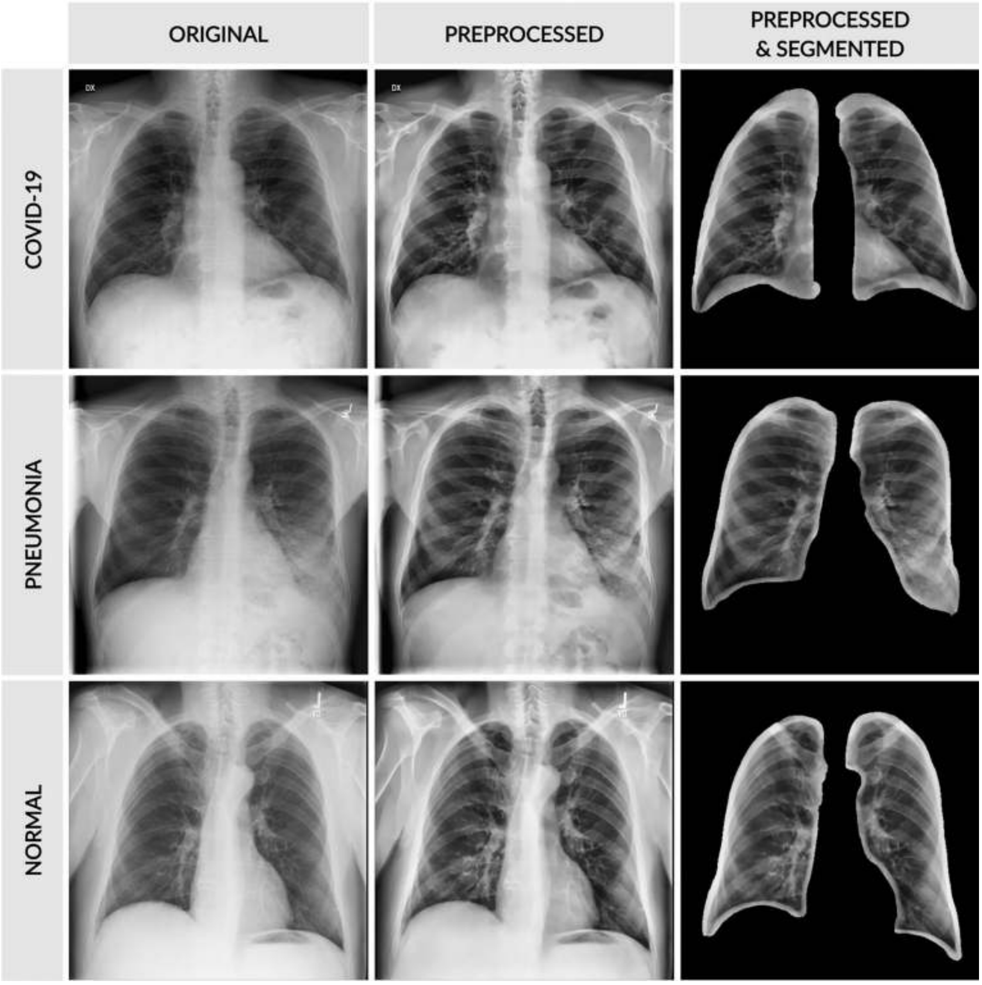


Fig. 4 Sample images of COVID-19, normal and pneumonia classes at various stages through the proposed pipeline

vital information for diagnosis. Detection of ROI reduces the required computational cost by extracting the features from a smaller part of the image. To capture the ROI by excluding insignificant regions of the image, we have deployed a U-Net architecture, which has consistently shown promising results in biomedical image segmentation tasks [72].

A brief overview of U-Net architecture is described to give insight to the reader. The U-Net architecture consists of a contracting path and an expansive path. The contracting path consists of the repeated application of two 3×3 unpadded convolutions, each followed by a rectified linear unit (ReLU) and a 2×2 max pooling operation with stride 2 for downsampling. At each step during downsampling, the number of feature channels is multiplied by 2. Each step in the expansive path involves the upsampling of the feature map followed by a 2×2 convolution that halves the number of feature channels, a concatenation with the correspondingly cropped feature map from the contracting path, and two 3×3 convolutions, each followed by a ReLU. In the final layer, a 1×1 convolution is used to map every 64-component feature vector to the desired number of classes which, in this case, is 3. The network consists of 23 convolutional layers in total.

In order to train the U-Net architecture, we use the segmentation dataset mentioned in Sect. 3.1. The lung segmentation masks were dilated to load lung boundary information within the training net and the images were resized to 512×512 pixels. The input images and their respective segmentation maps were used to train the network with binary cross-entropy loss function optimized using Adam optimizer with a learning rate of 0.001 and with Pixel Accuracy as the reporting metric that returns the percent of pixel rightly classified in the image as belonging to the binary mask:

$$\text{Pixel accuracy} = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{4}$$

where T_P represents the number of pixels correctly predicted as belonging to a given class. T_N represents the number of pixels correctly identified as not belonging to a given class. F_P and F_N represent the false positive and false negative pixel predictions, respectively.

The trained U-Net is then used to generate masks for the classification dataset containing three classes viz 'COVID-19', 'Normal', and 'Pneumonia'. On visual assessment, the masks acquired for COVID-19 class were not found to incorporate the lung region. In order to improve the quality of masks obtained for COVID-19, we manually segmented 200 samples of COVID-19 from the training set and fine tuned the above-mentioned model using these 200 samples. The ROIs extracted using the above-mentioned segmentation techniques are displayed in Fig. 4.

3.2.4 Pruned ensemble learning

In tasks involving classification of medical images and in particular, COVID -19 cases where new data sets are emerging on a daily basis, it is of paramount importance that the models not only perform robustly on new data sets but also on extreme cases of noise and outliers. The variance of a single CNN classifier during prediction is usually too high resulting in poor generalizability to real-world applications where type classification of images leads to sensitive decision making, such as selecting the course of care for the patient. To address this shortcoming, a combination of learners can be employed to help lower the variance and improve generalizability. The accuracy of predictions made by a set of base learners is often better than a single best learner [29]. In this work, the term 'base-learner' refers to a single deep CNN learner used in the ensemble.

One of the commonly used ensemble techniques is model averaging in which different base-learners contribute equally to the combined prediction by directly averaging the base learner's output score or predicted probability. The predicted probability is obtained using softmax function which noramalizes the output scores into a probability distribution:

$$p_{ij} = \text{softmax}\left(\overrightarrow{s_i}\right)[j] = \frac{\overrightarrow{s_i}[j]}{\sum_{k=1}^{K} e^{s_i[k]}} \tag{5}$$

where, vector $\overrightarrow{s_i}$ is the output from the last layer of the neural network for ith unit, $s_i[k]$ is the score corresponding to kth class/label, and p_{ij} is the predicted probability for unit i in class j.

Model averaging ensembles are constrained due to equal contribution from each base learner so one can employ an alternate method to allow unequal contribution depending on the confidence or performance of the specific base learner. We are thereby training a completely separate model, known as the meta-learner, to learn how best to incorporate each prediction made by base learners into the final combined prediction.

Algorithm 1 Proposed Framework

```
procedure MetaTrain($train_{base}, train_{meta}, test$)
    for $b_i \in B$ do
        Train $b_i$ using $train_{base}$
        Use $b_i$ to generate output score $C_i$ on $train_{meta}$
        Push $C_i$ to S
    for $m_i \in M$ do
        Train $m_i$ using S as input
        $ACC_{m_i} \leftarrow$ Evaluate $m_i$ on test
        if $ACC_{best} < ACC_{m_i}$
            $ACC_{best} \leftarrow ACC_{m_i}$
            $META_{best} \leftarrow m_i$
    return $META_{best}$
procedure MetaPrune($META_{best}$, $test$)
    Initialize empty stack A
    for $b_i \in B$ do
        Train $Meta_{best}$ using S - $C_i$ as input
        $ACC_{b_i} \leftarrow$ Evaluate $Meta_{best}$ on test
        Push $ACC_{b_i}$ to A
    $Remove_{b_i} \leftarrow \underset{b_i}{\operatorname{argmax}} A$
    if $ACC_{best} <= ACC_{Remove_{b_i}}$
        $ACC_{best} \leftarrow ACC_{Remove_{b_i}}$
        return $Best_{b_i}$
    else
        return null
procedure Main()
    B: Set of all base-learners
    M: Set of all meta-learners
    Initialize Stack S, $META_{best}$ = null,
    $ACC_{best}$ = 0, $Base_{remove}$ = 0, $C_0$=0
    $META_{best} \leftarrow METATRAIN(train_{base}, train_{meta}, test)$
    while $Base_{remove} \neq$ null do
        $Base_{remove} \leftarrow METAPRUNE(META_{best}, test)$
        Update B $\leftarrow$ B - $Base_{remove}$, S - $C_{Base_{remove}}$
```

The meta-learner hypothesis function takes predictions made by the base-models as input and learns to combine them to make a more accurate and robust output prediction. This is referred to as the stacked generalization ensemble technique and can result in improved predictive performance than any individual base-learner [82, 90]. In this study, we propose a modified stacked generalization procedure incorporating a pruning method for the selection of optimum base-learners. We have used the prediction of class probabilities from outputs of base-learners as input to the meta-learner instead of class labels where the class probabilities serve as the confidence measure for the predictions made. In this study, we evaluate the performance of the following meta-learners: (i) support vector machines, (ii) random forests, (iii) neural network, (iv) XGBoost, and (v) Naive Bayes.

The proposed algorithm requires a set of base-learners B and a set of meta-learners M as declared in Algorithm 1. The initial step of the algorithm requires training of all base-learners b_i as mentioned in Sect. 3.2.1 on $\text{train}_{\text{base}}$ which is the training set. By freezing the base-learner weight updates, these models then use the hold-out set $\text{train}_{\text{meta}}$ as input to predict class scores C_i. These class scores C_i are then pushed onto the stack S at each iteration. Post training of all base-learners and generation of stack S, each meta-learner m_i belonging to the set of all meta-learners M is trained using the stack S as input to make the final output label predictions. The meta-learner uses the set of predictions from base-learners and conditionally weighs each prediction, potentially resulting in better performance [82]. The meta-learner models are thus effectively trained on this holdout set $\text{train}_{\text{meta}}$ to avoid overfitting. We now evaluate each meta-learner on the test set, and choose the meta-learner $\text{META}_{\text{best}}$ with the best performing metric, in this case, accuracy ACC_{best}.

Another important challenge to consider is the selection of the base-learners amongst all suitable learners. After finalizing the meta-learner $\text{META}_{\text{best}}$ and starting with all the N base-learners, we use a pruning approach as shown Algorithm 1 to remove redundant base-learners resulting in increased model performance, generalizability, and decreased model complexity. We first iterate through the set of base-learners b_i, by removing the class scores C_i corresponding to b_i from the stack S. We then evaluate the performance of the meta-learner on the test data using the updated stack S and push the obtained accuracy to stack A. Argmax function is used to obtain the model Remove_{b_i}, whose removal corresponds to the best performance. If the removal of Remove_{b_i} leads to similar or better performance, we update the best accuracy ACC_{best} and subsequently remove the base-learner from the set of all base-learners B. In case of identical performance between more than one base-learner, we break the tie by removing the base-learner with a higher model complexity to ensure faster model deployment. At the end of each removal cycle, if removing any particular base-learner results in a similar or better performance, we repeat this process again on the updated set B now without the redundant base-learner. This is repeated until the outer while loop returns a 'null' value, signifying that the removal of any more base-learners will not result in improved performance. Thus, we are able to prune the set of all base-learners to the selected few for better generalizability and also lower ensemble model complexity for faster real-time model deployment.

3.2.5 Generative adversarial networks

In the medical AI domain, especially in the case of COVID-19, a lack of sufficient imaging data is a fundamental problem. Supervised deep learning is currently the state of the art in many computer vision and medical image analysis tasks, but its success is heavily dependent on the large-scale availability of labeled training data. Acquisition and labelling of medical image data are tedious, time-consuming, costly, and subject to many regulations. The scarcity of data and imbalanced classes are thus inherent. GANs can generate realistic-looking images from a latent distribution that follows the real data distribution and help balance the dataset for improved performance. In this study, we evaluate the feasibility of state of the art GAN architectures in generating realistic chest X-ray samples for COVID-19.

As shown in Fig. 5, the GAN training strategy is to define a game between two competing networks. The generator network maps a source of noise to the input space. The discriminator network receives either a generated sample or a true data sample and must distinguish between the two. The generator is trained to deceive the discriminator. Formally, the game between the generator G and the discriminator D is the minimax objective function:

$$\min_G \max_D \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}_r}[\log(D(\mathbf{x}))] + \mathop{\mathbb{E}}_{\tilde{\mathbf{x}}\sim\mathbb{P}_g}[\log(1-D(\tilde{\mathbf{x}}))] \tag{6}$$

where $\mathbb{P}_r$ is the data distribution and $\mathbb{P}_g$ is the model distribution implicitly defined by $\tilde{x} = \mathrm{G(z)}$, $z \sim p(z)$. The input z to the generator is sampled from some simple noise distribution p, such as the uniform distribution or a spherical Gaussian distribution.

3.2.6 Visualization

The lack of tools to understand the behaviour of black-box models affects the use of deep learning in medical imaging scenario where explainability and reliability are the key elements for establishing trust amongst the clinicians and patients.

The probability of the chest X-ray classification model gathering distinguishing characteristics from outside the lung area is high due to the complex nature of some deep learning models which often results in low generalizability [52]. It is also possible for deep learning models to identify biologically novel patterns by understanding underlying features possibly overlooked during diagnosis. These insights will only be available; however, if the model can be interpreted, and the examiner can understand the pattern used by the model to make its predictions. The trust on forecast is always questionable if the reasons behind the prediction is unknown or poorly understood [58].

We deploy gradient-weighted class activation mapping known as Grad-CAM that is class-discriminatory and locates relevant regions of the image. It is a gradient-based visualization method that calculates the scores in a trained model for a given image category using the feature maps of

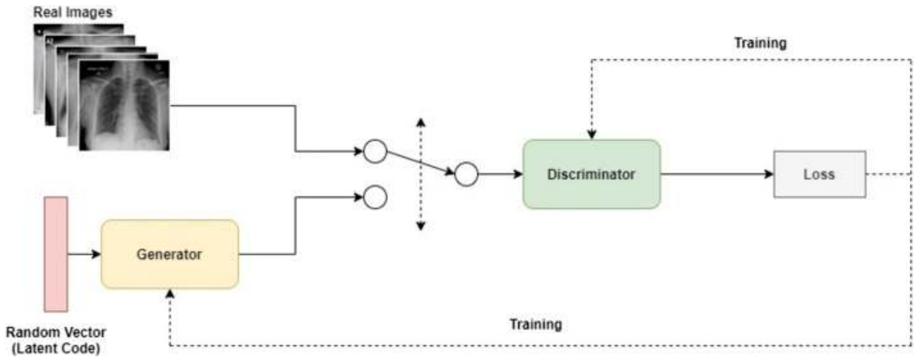


Fig. 5 General GAN architecture. The real images and synthetic images generated using the generator and fed into the discriminator. Using the WGAN-GP loss function backpropagation is done to improve both the networks

Table 3 Performance metrics of base learners where C, N and P are COVID-19, normal and pneumonia, respectively

CNN model	Accuracy	Kappa score		Precision	Recall	F-1 score
Densenet169	82.667	0.74	C	64	100	78.04878049
			N	94	71.21	81.03310938
			P	90	86.53	88.2308956
Densenet121	84	0.76	C	64	96.97	77.10852954
			N	96	72.72	82.75391181
			P	92	90.19	91.08600911
ResNet50	79.34	0.69	C	52	100	68.42105263
			N	92	68.65	78.62807345
			P	94	82.45	87.84698215
Vgg19	89.34	0.84	C	82	97.61	89.12666333
			N	94	81.03	87.03445124
			P	92	92	92
Vgg16	84	0.76	C	64	96.97	77.10852954
			N	94	74.6	83.18386714
			P	94	87.03	90.38082086

the deepest convolutional layer [74]. The gradients that are flowing backward are pooled globally to measure the importance of the weights in the decision-making process. This can be applied to off-the-shelf CNN-based architectures without any modifications in the standard network architecture.

4 Experimentation and results

4.1 Evaluation metrics

Based on classifying images into three separate groups, viz. COVID-19, Pneumonia, and Normal, the issue of multi-class classification poses a greater challenge as compared to a binary classification task due to the increased complexity of models. In addition, this study tests the classification model's ability to distinguish between COVID-19 and pneumonia which have similar imaging modalities. We deploy the following metrics to evaluate the performance of the proposed multi-class classification task.

4.1.1 Overall accuracy (OA)

Overall accuracy is a metric for evaluating classification models [57]. Informally, overall accuracy is the fraction of predictions the model gets right. Formally, it is defined as:

$$\mathrm{OA} = \frac{(T_p + T_n)}{(T_p + T_n + F_p + F_n)} \tag{7}$$

where, T_p represents the number of samples correctly predicted as belonging to a given class. T_n represents the number of samples correctly identified as not belonging to a given class. F_p and F_n represent the false positive and false negative sample predictions, respectively.

4.1.2 Precision (P)

Precision refers to the proportion of correct positive identifications to all positive identifications [85]. A low precision will correspond to high false positives and would result in an unwanted burden on the health care systems, catering to patients incorrectly classified as having a

Table 4 Performance metrics of VGG-19 using different training techniques, where C, N and P are COVID-19, normal and pneumonia, respectively.

Training method	Accuracy	Kappa score		Precision	Recall	F-1 score
Entire	89.34	0.84	C	82	97.61	89.12666333
			N	94	81.03	87.03445124
			P	92	92	92
TopHead	70	0.55	C	48	75	58.53658537
			N	92	59.74	72.44075392
			P	70	85.37	76.92476025

Here TopHead refers to training just the custom head, and entire refers to training the entire network

Table 5 Performance metrics of entirely trained VGG-19 using different weight initialization, where C, N and P are COVID-19

Weight initialization	Accuracy	Kappa score		Precision	Recall	F-1 score
Image Net	89.34	0.84	C	82	97.61	89.12666333
			N	94	81.03	87.03445124
			P	92	92	92
Chexpert	84	0.76	C	68	97.14	79.99903113
			N	92	76.67	83.63834707
			P	92	83.63	87.61555543
Random	70	0.55	C	50	69.45	58.14148179
			N	82	57.74	67.76413339
			P	78	90.69	83.86768629

Table 6 Performance metrics of entirely trained VGG-19 using ImageNet weights using different training set distributions, where C, N and P are COVID-19, normal and pneumonia, respectively

Dataset	Accuracy	Kappa score		Precision	Recall	F-1 score
A	89.34	0.84	C	82	97.61	89.12666333
			N	94	81.03	87.03445124
			P	92	92	92
B	86	0.79	C	74	97.36	84.08776844
			N	94	79.66	86.2379362
			P	90	84.9	87.37564322
C	74.67	0.62	C	32	100	48.48484848
			N	100	63.29	77.51852532
			P	92	83.63	87.61555543

particular pathology. Precision lies between [0,1] and is defined as:

$$P = \frac{T_p}{T_p + F_p} \tag{8}$$

4.1.3 Recall (R)

Recall is defined as the number of true positives (T_p) over the number of true positives plus the number of false negatives (F_n) [85]. Recall takes into account the false negative rate which is of utmost significance in medical tasks. Thus, a lower recall rate would result in incorrect diagnosis and course of treatment for the patients. The precision values obtained are between [0,1] and is defined as:

Table 7 Enhancement techniques used for hypothesis 5

Dataset	Preprocessing	Segmentation
Raw	X	X
Preprocessed	✓	X
Segmented	X	✓
Both	✓	✓

$$R = \frac{T_p}{T_p + F_n} \tag{9}$$

4.1.4 F1-score (F1)

F1-score is defined as the harmonic mean of precision and recall [96]. For classification tasks where both precision and recall are of high significance as in this study focusing on the detection of COVID-19, F1-score should be maximized. The values obtained are between [0,1] with 1 being the highest.

$$F1 = 2\left(\frac{P \times R}{P + R}\right) \tag{10}$$

4.1.5 Kappa score (κ)

Kappa score, also known as Cohen's kappa is a statistic that measures inter-annotator agreement [55]. It is defined as:

$$\kappa = \frac{(p_o - p_e)}{(1 - p_e)} \tag{11}$$

where p_o is the empirical probability of agreement on the label assigned to any sample, and p_e is the expected agreement when both annotators assign labels randomly. p_e

Table 8 Performance metrics obtained using the preprocessing and segmentation technique proposed in the study, where C, N and P are COVID-19, normal and pneumonia, respectively

Input type	Accuracy	Kappa score		Precision	Recall	F-1 score
Preprocessed	94.67	0.92	C	90	100	94.73684211
			N	98	90.74	94.23036982
			P	96	94.11	95.04560518
Segmented	94.67	0.92	C	100	98.03	99.00520123
			N	94	92.15	93.06580714
			P	90	93.75	91.83673469
Both	95.34	0.93	C	100	98.03	99.00520123
			N	94	92.15	93.06580714
			P	92	95.83	93.87595166
Raw	89.34	0.84	C	82	97.61	89.12666333
			N	94	81.03	87.03445124
			P	92	92	92

Table 9 Performance metrics of base-learners retrained with best model obtained obtained through hypotheses 1–5 where C, N and P are COVID-19, normal and pneumonia, respectively

CNN Model	Accuracy	Kappa score		Precision	Recall	F-1 score
Densenet169	93.34	0.9	C	100	100	100
			N	86	93.47	89.57953976
			P	94	87.03	90.38082086
Densenet121	96	0.94	C	100	98.03	99.00520123
			N	94	94	94
			P	94	95.91	94.94539519
ResNet50	96	0.94	C	100	98.03	99.00520123
			N	96	94.11	95.04560518
			P	92	95.83	93.87595166
Vgg19	94.67	0.92	C	100	100	100
			N	96	88.89	92.30829142
			P	88	95.65	91.66566839
Vgg16	95.34	0.93	C	100	98.03	99.00520123
			N	94	92.15	93.06580714
			P	92	95.83	93.87595166

is estimated using a per-annotator empirical prior over the class labels.

4.2 Experimentation

The proposed study is centered on testing various model configurations to justify decisions taken during model training. We explore various hypothesis testing to justify decisions taken during CNN model training for the classification task without holding any assumptions. Training and testing were performed using an 11 GB RTX 2080 Ti GPU. The proposed deep learning framework has been implemented using Keras deep learning library with TensorFlow as a backend. The experiments are as follows:

Hypothesis 1: base models In the initial round of experiments, we train the following base models: VGG–16, VGG–19, ResNet-50, DenseNet-121 and DenseNet-169 as shown in Table 3. The architectural and training specifications of all models have been specified in Sect. 3.2.1. These models have been evaluated on test-split of dataset A. Post evaluation, VGG-19 emerges as the best performing model with an overall accuracy of 89.34 and a Kappa score of 0.84. We fix the base model as VGG-19 for further hypothesis testing.

Hypothesis 2: training approach In the field of deep learning, common training approaches include training the entire network and training the top few layers of the network. We evaluate the effectiveness of both these training approaches using the VGG-19 architecture using the test-split of the dataset A. It is evident from Table 4 that training the entire network outperforms training the custom

Table 10 Performance metrics of different meta-learners obtained when using all base models, where C, N and P are COVID-19, normal and pneumonia, respectively

Meta-Learner	Accuracy	Kappa score		Precision	Recall	F-1 score
SVC	96.67	0.95	C	100	100	100
			N	94	95.91	94.94539519
			P	96	94.11	95.04560518
RF	95.34	0.93	C	100	100	100
			N	92	93.87	92.92559316
			P	94	92.15	93.06580714
Neural Net	97.34	0.96	C	100	100	100
			N	96	96	96
			P	96	96	96
XGBoost	94	0.91	C	100	100	100
			N	94	88.67	91.25723983
			P	88	93.61	90.71835251
Naive Bayes	98	0.97	C	100	100	100
			N	96	97.95	96.96519722
			P	98	96.07	97.02540321
Pruned Naive	98.67	0.98	C	100	100	100
			N	98	98	98
			P	98	98	98

head. Hence, we fix the entire network training approach for all successive stages of hypothesis testing.

Hypothesis 3: weight initialization An important hyper-parameter tuning required during deep neural network training is at the weight initialization stage. In these experiments, we train the entire VGG–19 architecture and evaluate the effect of the following three weight initialization methods: (1) ImageNet—pretrained ImageNet weights, (2) CheXpert—VGG-19 pre-trained on CheXpert dataset for 100 epochs, and (3) random initial weights. Table 5 highlights the superior results of ImageNet weight initialization, we thus, use it for all successive stages of hypothesis testing.

Hypothesis 4: class distribution Post evaluation of previous hypotheses, we have trained all layers of VGG-19 architecture initialized with ImageNet weights on three datasets with varying training class distributions as mentioned in Sect. 3.1.

The dataset split as mentioned in Table 2 uses the original images without preprocessing or segmentation to help decide which training split provides superior performance. The results obtained from the hypothesis experiments are provided in Table 6 which clearly indicates that upsampling and unbalanced training sets do not improve model performance on the hold-out test set. As a consequence, downsampled balanced class distribution, Dataset A, is used for further hypothesis testing.

Hypothesis 5: preprocessing and segmentation We have used the same VGG–19 architecture with the configuration resulting from previous hypothesis testing to now evaluate the effect of preprocessing and segmentation in the model performance. For this purpose, we have created four different versions of the train and test sets to distinguish and study the effects of preprocessing and segmentation separately. The four versions of the aforementioned dataset are referred to in Table 7. Observing the results in Table 8, it is now evident that both preprocessing and segmentation individually result in superior model performance, and the model performs best with the effect of both preprocessing and segmentation together which falls in line with the expected results. For further hypothesis testing, we have used the preprocessed and segmented dataset for further experimentation.

Hypothesis 6: proposed pruned ensemble learning The ensemble learning set-up requires choosing an optimum meta-learner hypothesis function to best combine the predictions made by the base learners as discussed in Sect. 3.2.4. To evaluate the performance of various meta-learner algorithms, we have first deployed all base learners, i.e. VGG-19, VGG-16, DenseNet-121, DenseNet-169, and ResNet-50 retrained using the best obtained training methods from earlier experiments. The performance of these models on various metrics are summarized in Table 9. We have then iteratively used various meta-learners to combine the trained base learners' output

Table 11 Comparative evaluation of the proposed framework with other studies, where C, N and P are COVID-19, normal and pneumonia, respectively

Study	Accuracy	Kappa score		Precision	Recall	F-1 score
Proposed method	98.67	0.98	C	100	100	100
			N	98	98	98
			P	98	98	98
Wang and Wong [87]	93.34	0.9	C	98.91	91	94.79026907
			N	90.47	95	92.67967865
			P	91.26	94	92.60973767
Oh et al. [63]	88.9	–	C	76.9	100	86.94177501
			N	95.7	90	92.76252019
			P	90.3	93	91.63011457
Khan et al. [42]	90.21	0.83	C	97	89	92.82795699
			N	92	85	88.36158192
			P	87	95	90.82417582
Ozturk et al. [64]	87.022	0.776	C	80.702	97.872	88.46154697
			N	89.635	86.642	88.11309099
			P	85.714	85.366	85.53964606
Apostolopoulos and Mpesiana [3]	93.55	0.895	C	93.671	93.277	93.47358482
			N	93.156	95.286	94.20896208
			P	94.07	91.27	92.64884968
Ucara and Korkmaz [84]	98.257	0.974	C	100	99.351	99.67444357
			N	98.039	97.403	97.71996518
			P	96.732	98.013	97.3682869
Chowdhury et al. [14]	97.935	0.966	C	99.291	99.057	99.17386197
			N	97.973	97.849	97.91096074
			P	97.508	97.706	97.60689959
Haghanifar et al. [27]	87.208	0.778	C	94.203	90.278	92.19874604
			N	82.515	95.166	88.39012038
			P	92.416	77.976	84.58413559
Mangal et al. [53]	90.52	0.808	C	96.774	74.359	84.09854167
			N	98.864	99.487	99.17452161
			P	86.801	100	92.934192

predictions and thereby evaluate their performance. It is clear from Table 10 that Naive Bayes outperformed all other Hypothesis functions with a high overall accuracy of 98 and a Kappa score of 0.97. A model hyperparameter is a characteristic of a model that is external to the model and whose value cannot be estimated from data. We have used grid search to perform hyperparameter tuning to finalize the optimum hyperparameters for all Hypothesis functions before evaluating them as a meta-learner in the ensemble set-up.

The various hyperparameters used in the grid search for this hypothesis are learning rate for XGBoost algorithm; kernel type, gamma for support vector machines; regularization parameter "C" for logistic regression; no. of estimators, max depth, min samples split, min samples leaf for random forests; and activation, hidden layer size, number

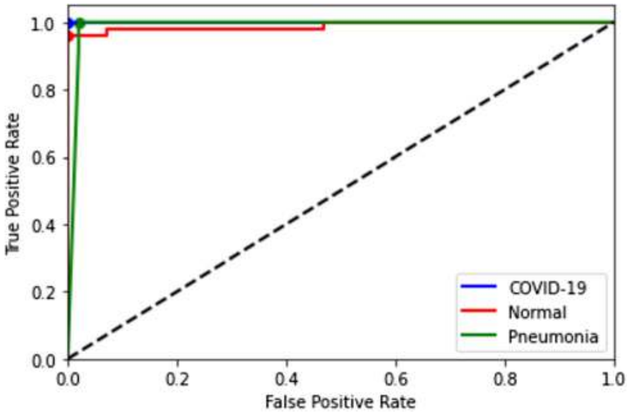
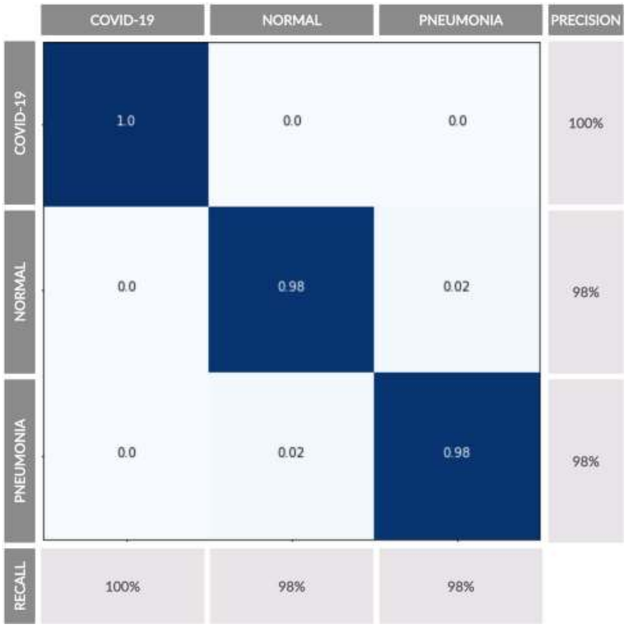


Fig. 6 Receiver operating characteristic curve for the final pruned ensemble model



	COVID-19	NORMAL	PNEUMONIA	PRECISION
COVID-19	1.0	0.0	0.0	100%
NORMAL	0.0	0.98	0.02	98%
PNEUMONIA	0.0	0.02	0.98	98%
RECALL	100%	98%	98%	

Fig. 7 The final confusion matrix obtained on test set, with *X*-axis representing the actual class labels and *Y*-axis representing the predicted class labels from the final pruned ensemble model

of layers for neural network. Naive Bayes algorithm does not have any hyper-parameters. The precision and recall for each class as summarized in Table 10 reveal that Naive Bayes hypothesis function as a meta-learner offers the highest degree of generalizability.

We then continue with the pruning process to select the optimal base learners as mentioned in Sect. 3.2.4. At the end of the pruning process, among all base-learner functions in the member set, VGG-16, ResNet-50, DenseNet-169, and DenseNet-121 together show the best performance. The removal of one base learner resulted in improved overall performance. This improved performance can be attributed to the discarded model negatively affecting the final output generated by the meta-learner. The results obtained from Table 11 show that the above-mentioned set of base-learners along with Naive Bayes as the meta-learner performs best in the classification task with an overall accuracy of 98.67%, average precision of 98.67%, average recall of 98.67%, average F1 score of 98.67% and a kappa score of 0.98. The ROC curve and confusion matrix for this best performing ensemble model is shown in Figs. 6 and 7, respectively. The Grad-CAM visualization for the final four base-learners using a COVID-19 sample has been shown in Fig. 8.

4.2.1 GANs

Given the recent promise that adversarial networks have shown, various GAN models have been explored to generate COVID-19 chest X-ray samples. Wasserstein GAN with gradiant penalty, auxiliary classifier GAN, least square GAN, and deep convolution GAN has been trained and images generated using each of these methods are shown in Fig. 9 [26, 54, 62, 69]. Clearly, WGAN with gradient penalty is able to generate images of extremely high quality. The input and output images generated are 128x128. The standard parameters as used by Gulrajani et al. have been used for training the WGAN with $\lambda = 10$, $n_{\text{critic}} = 5$, $\alpha = 0.0001$, $\beta_1 = 0$, $\beta_2 = 0.9$ [26], Adam optimizer with learning rate of 0.0002, a latent vector of 100 dimensions and a batch size of 64 has been used [26].

5 Discussion and conclusion

In a pandemic situation such as COVID-19, rapid triaging of patients is critical to contain the spread. The commonly used RT-PCR nucleic acid-based test, although extremely useful, is time-consuming, expensive, and in short supply, especially in low resource settings. In order to address these shortcomings, we have proposed an artificial intelligence-based solution to help triage COVID-19 patients faster and eliminate the scope of human error. The final model deploys VGG-16, ResNet-50, DenseNet-121, and DenseNet-169 as base learners along with a Naive Bayes meta-learner in a pruned ensemble learning framework. The proposed model demonstrates state of the art results, with 98.67% accuracy, 0.98 Kappa score, and F-1 scores of 100, 98, and 98 for COVID-19, normal, and pneumonia classes, respectively.

The experiments in this study evaluate the effectiveness of different training methods such as weight initialization, training class distribution, preprocessing, segmentation, and ensemble learning. We have used the latest publicly available datasets with regards to COVID-19 and compared the proposed model with the results of recent papers using similar datasets as summarized in Table 11. The proposed diagnosis model outperforms all existing methods and we assume that, with the increased size of the training data set, we can produce even better results. This research not only concentrated on overall accuracy but also illustrated the generalizability of the proposed model to different classes by carefully analysing precision and recall measures on each class during various hypothesis testing stages as outlined in the results of the experiment.

In order to validate the proposed framework with regard to interpretability, Grad-CAM visualization has been performed. Public availability of data for COVID-19 cases has

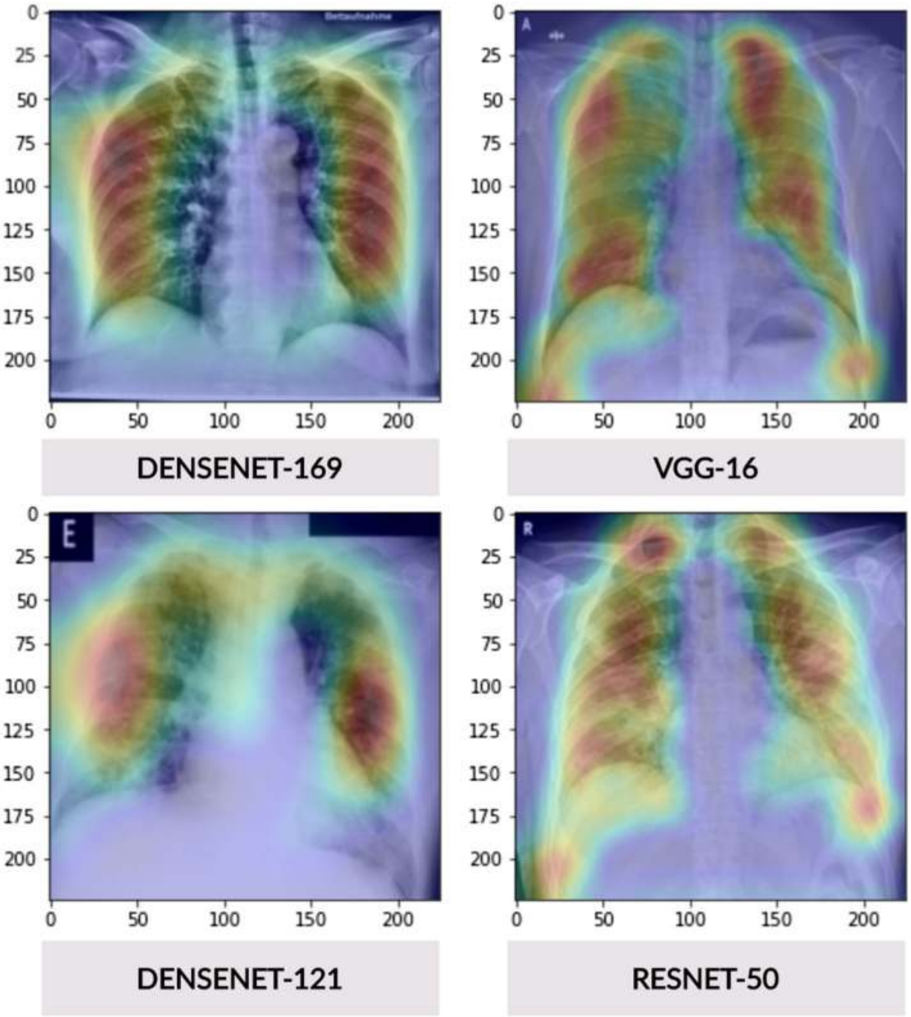


Fig. 8 COVID-19 Grad-CAM visualization for the base learners used in the final model

been minimal, which has hindered the fast progress of research studies. To address this issue, we have deployed multiple generative adversarial networks and qualitatively evaluated these samples through visual inspection. The availability of sufficient public data can pave the path for succeeding studies to exploit these observations and explore greater effectiveness of generative models in such a setting. Nonetheless, we strongly believe that this work can be successfully implemented for a low-cost, quick, and automatic COVID-19 disease diagnosis.

Most classification tasks assume equal costs of false negatives and false positives. However, in medical image classification problems such as this, a false negative error rate is far more expensive than a false positive error rate since the failure of diagnosis of a disease such as COVID-19 can not only endanger the patient's life but also promote further community spread. The obtained results clearly underwrite the ability of the proposed model in successfully removing all false negatives and false positives for detecting COVID-19. This can be attributed to the manual segmentation of some of the COVID-19 chest X-rays highlighting the importance of collaboration between AI and the medical community.

The proposed model does suffer from some limitations. One major shortcoming associated with all COVID-19 related studies including ours is the small sample size. Despite the encouraging results of using deep learning-based models to screen patients with COVID-19, further data collection is required to test the generalizability of such AI models to other patient populations which are not a part of the training set. A collaborative effort in data collection may facilitate improving the AI model. Further studies should explore combining X-ray imaging and clinical information and confirmation in hospital settings.

Chest X-rays are commonly used for the detection of multiple pathologies. The proposed pipeline can be used effectively for the detection of other diseases such as pneumonia, pneumothorax, edema, etc. which have clear

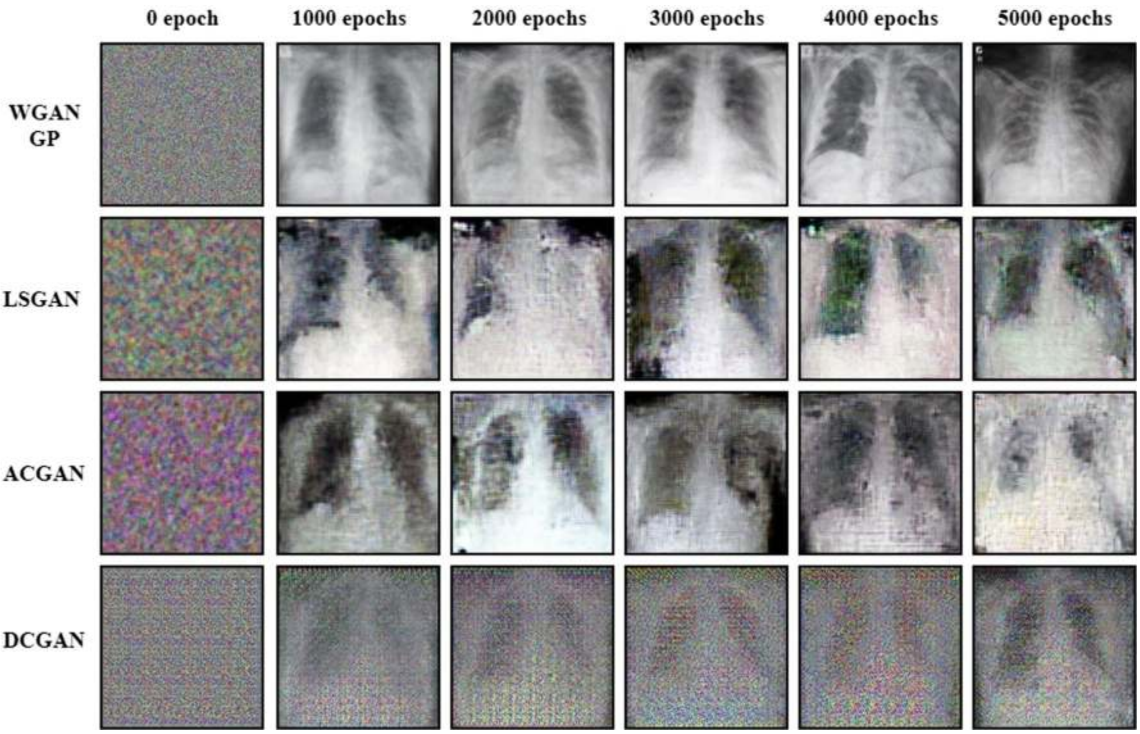


Fig. 9 Images generated by various GAN models. The models explored include Wasserstein GAN (WGAN), least squares GAN (LSGAN), auxiliary classifier GAN (ACGAN), and deep convolution GAN. The synthetic images generated using WGAN are extremely realistic

indicators in chest radiographs. This can be done either by fine-tuning or complete retraining of our proposed model. Moreover, the proposed pipeline can also be incorporated in the detection of any other disease where visual indicators of distinction are present in the corresponding medical images.

Precise diagnosis of any disease especially in radiology can be challenging even to expert radiologists owing to the minute details in chest X-ray images that can easily go unnoticed. While this model is able to efficiently pick up the necessary details for diagnosis, it reduces monotonous procedural elements such as manual examination of chest X-ray images thus, allowing doctors to focus on more demanding tasks. In conclusion, these results illustrate the potential impact of a highly accurate and explainable Artificial Intelligence-based algorithm for the rapid identification of COVID-19 patients, which could be of profound help in combating the rapid proliferation of this disease. Moreover, such studies can also be incorporated in the detection of any similar disease with basic fine tuning and retraining of the proposed framework.

Acknowledgements We would like to thank the reviewers for their insightful comments which helped improve the overall quality of our manuscript. We would also like to thank Nishanth T. Arun for his valuable suggestions during the course of this work.

Compliance with ethical standards

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

References

1. Abbas A, Abdelsamea MM, Gaber MM (2020) Classification of covid-19 in chest X-ray images using detrac deep convolutional neural network. Appl Intell. https://doi.org/10.1007/s10489-020-01829-7
2. Adams HJ, Kwee TC, Yakar D, Hope MD, Kwee RM (2020) Chest CT imaging signature of coronavirus disease 2019 infection: in pursuit of the scientific evidence. Chest. https://doi.org/10.1016/j.chest.2020.06.025
3. Apostolopoulos ID, Mpesiana TA (2020) Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. Phys Eng Sci Med 43(2):635–640. https://doi.org/10.1007/s13246-020-00865-4
4. Arons MM, Hatfield KM, Reddy SC, Kimball A, James A, Jacobs JR, Taylor J, Spicer K, Bardossy AC, Oakley LP, Tanwar S, Dyal JW, Harney J, Chisty Z, Bell JM, Methner M, Paul P, Carlson CM, McLaughlin HP, Thornburg N, Tong S, Tamin A, Tao Y, Uehara A, Harcourt J, Clark S, Brostrom-Smith C, Page LC, Kay M, Lewis J, Montgomery P, Stone ND, Clark TA, Honein MA, Duchin JS, Jernigan JA (2020) Presymptomatic sars-cov-2 infections and transmission in a skilled nursing facility. New Engl J Med 382(22):2081–2090. https://doi.org/10.1056/NEJMoa2008457

5. Bar Y, Diamant I, Wolf L, Greenspan H (2015) Deep learning with non-medical training used for chest pathology identification. In: Medical imaging 2015: computer-aided diagnosis, international society for optics and photonics, SPIE, vol 9414, pp 215–221. https://doi.org/10.1117/12.2083124
6. Bar Y, Diamant I, Wolf L, Lieberman S, Konen E, Greenspan H (2015) Chest pathology detection using deep learning with non-medical training. In: 2015 IEEE 12th international symposium on biomedical imaging (ISBI), pp 294–297. https://doi.org/10.1109/ISBI.2015.7163871
7. Bieniecki W, Grabowski S, Rozenberg W (2007) Image preprocessing for improving OCR accuracy. In: 2007 international conference on perspective technologies and methods in MEMS design, pp 75–80. https://doi.org/10.1109/MEMSTECH.2007.4283429
8. Binnicker MJ (2020) Emergence of a novel coronavirus disease (COVID-19) and the importance of diagnostic testing: why partnership between clinical laboratories, public health agencies, and industry is essential to control the outbreak. Clin Chem 66(5):664–666. https://doi.org/10.1093/clinchem/hvaa071
9. 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 Programs Biomed 196:105608. https://doi.org/10.1016/j.cmpb.2020.105608
10. Buades A, Coll B, Morel J (2005) A non-local algorithm for image denoising. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol 2, pp 60–65. https://doi.org/10.1109/CVPR.2005.38
11. Buades A, Coll B, Morel JM (2011) Non-local means denoising. Image Process Line 1:208–212. https://doi.org/10.5201/ipol.2011.bcm_nlm
12. Chen H, Dou Q, Yu L, Qin J, Heng PA (2018) Voxresnet: deep voxelwise residual networks for brain segmentation from 3D MR images. NeuroImage 170:446–455. https://doi.org/10.1016/j.neuroimage.2017.04.041
13. Cheng JZ, Ni D, Chou YH, Qin J, Tiu CM, Chang YC, Huang CS, Shen D, Chen CM (2016) Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. Sci Rep 6(1):24454. https://doi.org/10.1038/srep24454
14. Chowdhury MEH, Rahman T, Khandakar A, Mazhar R, Kadir MA, Mahbub ZB, Islam KR, Khan MS, Iqbal A, Emadi NA, Reaz MBI, Islam MT (2020) Can AI help in screening viral and covid-19 pneumonia? IEEE Access 8:132665–132676. https://doi.org/10.1109/ACCESS.2020.3010287
15. Chung A (2020a) Actualmed covid-19 chest X-ray data initiative. https://github.com/agchung/Actualmed-COVID-chestxray-dataset
16. Chung A (2020b) Figure 1 covid-19 chest X-ray data initiative. https://github.com/agchung/Figure1-COVID-chestxray-dataset
17. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O (2016) 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin S, Joskowicz L, Sabuncu MR, Unal G, Wells W (eds) Medical image computing and computer-assisted intervention—MICCAI 2016. Springer, Cham, pp 424–432
18. Ciresan DC, Giusti A, Gambardella LM, Schmidhuber J (2012) Deep neural networks segment neuronal membranes in electron microscopy images. In: NIPS, pp 2852–2860
19. Cohen JP, Morrison P, Dao L (2020) Covid-19 image data collection. arXiv 200311597
20. Deming ME, Michael NL, Robb M, Cohen MS, Neuzil KM (2020) Accelerating development of sars-cov-2 vaccines: the role for controlled human infection models. New Engl J Med 383(10):e63. https://doi.org/10.1056/NEJMp2020076
21. Deng J, Dong W, Socher R, Li L, Kai Li, Li Fei-Fei (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, pp 248–255. https://doi.org/10.1109/CVPR.2009.5206848
22. Farooq M, Hafeez A (2020) Covid-resnet: a deep learning framework for screening of covid19 from radiographs. arxiv2003.14395
23. Geneva: World Health Organization (2020) WHO coronavirus disease (COVID-19) dashboard. https://covid19.who.int/. Accessed 7 Nov 2020
24. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. The MIT Press, Cambridge, MA, USA
25. Gorantla R, Singh RK, Pandey R, Jain M (2019) Cervical cancer diagnosis using cervixnet: a deep learning approach. In: 2019 IEEE 19th international conference on bioinformatics and bioengineering (BIBE), pp 397–404. https://doi.org/10.1109/BIBE.2019.00078
26. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC (2017) Improved training of Wasserstein Gans. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems 30. Curran Associates, Inc, Red Hook, pp 5767–5777
27. Haghanifar A, Majdabadi MM, Choi Y, Deivalakshmi S, Ko S (2020) COVID-CXNet: detecting COVID-19 in frontal chest X-ray images using deep learning. arxiv2006.13807
28. Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G (2017) Learning from class-imbalanced data: review of methods and applications. Expert Syst Appl 73:220–239. https://doi.org/10.1016/j.eswa.2016.12.035
29. Hansen LK, Salamon P (1990) Neural network ensembles. IEEE Trans Pattern Anal Mach Intell 12(10):993–1001. https://doi.org/10.1109/34.58871
30. Havaei M, Davy A, Warde-Farley D, Biard A, Courville A, Bengio Y, Pal C, Jodoin PM, Larochelle H (2017) Brain tumor segmentation with deep neural networks. Med Image Anal 35:18–31. https://doi.org/10.1016/j.media.2016.05.004
31. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), pp 770–778. https://doi.org/10.1109/CVPR.2016.90
32. He X, Lau EHY, Wu P, Deng X, Wang J, Hao X, Lau YC, Wong JY, Guan Y, Tan X, Mo X, Chen Y, Liao B, Chen W, Hu F, Zhang Q, Zhong M, Wu Y, Zhao L, Zhang F, Cowling BJ, Li F, Leung GM (2020) Temporal dynamics in viral shedding and transmissibility of covid-19. Nat Med 26(5):672–675. https://doi.org/10.1038/s41591-020-0869-5
33. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 2261–2269. https://doi.org/10.1109/CVPR.2017.243
34. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H, Haghgoo B, Ball R, Shpanskaya K et al (2019) Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. Proc AAAI Conf Artif Intell 33:590–597
35. Jaeger S, Candemir S, Antani S, Wáng YXJ, Lu PX, Thoma G (2014) Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quant Imaging Med Surg 4(6):475–477. https://doi.org/10.3978/j.issn.2223-4292.2014.11.20
36. Jaiswal AK, Tiwari P, Kumar S, Gupta D, Khanna A, Rodrigues JJ (2019) Identifying pneumonia in chest X-rays: a deep learning approach. Measurement 145:511–518. https://doi.org/10.1016/j.measurement.2019.05.076
37. Jamal I, Akram MU, Tariq A (2012) Retinal image preprocessing: background and noise segmentation. Telkomnika 10(3):537–544

38. Jang S, Han SH, Rhee JY (2020) Cluster of coronavirus disease associated with fitness dance classes, South Korea. Emerg Infect Dis 26(8):1917–1920. https://doi.org/10.3201/eid2608.200633
39. Johnson JM, Khoshgoftaar TM (2019) Survey on deep learning with class imbalance. J Big Data 6(1):27. https://doi.org/10.1186/s40537-019-0192-5
40. Kallenberg M, Petersen K, Nielsen M, Ng AY, Diao P, Igel C, Vachon CM, Holland K, Winkel RR, Karssemeijer N, Lillholm M (2016) Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. IEEE Trans Med Imaging 35(5):1322–1331. https://doi.org/10.1109/TMI.2016.2532122
41. Kanne JP, Little BP, Chung JH, Elicker BM, Ketai LH (2020) Essentials for radiologists on covid-19: an update-radiology scientific expert panel. Radiology 296(2):E113–E114. https://doi.org/10.1148/radiol.2020200527
42. 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 Programs Biomed 196:105581. https://doi.org/10.1016/j.cmpb.2020.105581
43. Kooraki S, Hosseiny M, Myers L, Gholamrezanezhad A (2020) Coronavirus (covid-19) outbreak: what the department of radiology should know. J Am Coll Radiol 17(4):447–451. https://doi.org/10.1016/j.jacr.2020.02.008
44. Krawczyk B (2016) Learning from imbalanced data: open challenges and future directions. Progress Artif Intell 5(4):221–232. https://doi.org/10.1007/s13748-016-0094-0
45. Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. Commun ACM 60(6):84–90. https://doi.org/10.1145/3065386
46. Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, Azman AS, Reich NG, Lessler J (2020) The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: estimation and application. Ann Intern Med 172(9):577–582. https://doi.org/10.7326/M20-0504
47. Liu S, Liu S, Cai W, Pujol S, Kikinis R, Feng D (2014) Early diagnosis of Alzheimer's disease with deep learning. In: 2014 IEEE 11th international symposium on biomedical imaging (ISBI), pp 1015–1018. https://doi.org/10.1109/ISBI.2014.6868045
48. Liu Y, Gayle AA, Wilder-Smith A, Rocklöv J (2020) The reproductive number of COVID-19 is higher compared to SARS coronavirus. J Travel Med. https://doi.org/10.1093/jtm/taaa021
49. Baozhong LIU, Jianbin LIU (2018) Overview of image noise reduction based on non-local mean algorithm. MATEC Web Conf 232:03029. https://doi.org/10.1051/matecconf/201823203029
50. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR), pp 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965
51. Luo L, Liu D, Liao Xl, Wu Xb, Jing Ql, Zheng Jz, Liu Fh, Yang Sg, Bi B, Li Zh, Liu Jp, Song Wq, Zhu W, Wang Zh, Zhang Xr, Chen Pl, Liu Hm, Cheng X, Cai Mc, Huang Qm, Yang P, Yang Xf, Huang Zg, Tang Jl, Ma Y, Mao C (2020) Modes of contact and risk of transmission in covid-19 among close contacts. medRxiv https://doi.org/10.1101/2020.03.24.20042606
52. Maguolo G, Nanni L (2020) A critic evaluation of methods for covid-19 automatic detection from X-ray images. arXiv preprint arXiv:200412823
53. Mangal A, Kalia S, Rajgopal H, Rangarajan K, Namboodiri V, Banerjee S, Arora C (2020) CovidAID: COVID-19 detection using chest X-ray. arXiv preprint arXiv:200409803
54. Mao X, Li Q, Xie H, Lau RYK, Wang Z, Smolley SP (2017) Least squares generative adversarial networks. In: 2017 IEEE international conference on computer vision (ICCV), pp 2813–2821. https://doi.org/10.1109/ICCV.2017.304
55. McHugh ML (2012) Interrater reliability: the kappa statistic. Biochem Med 22(3):276–282
56. Mei X, Lee HC, Ky Diao, Huang M, Lin B, Liu C, Xie Z, Ma Y, Robson PM, Chung M, Bernheim A, Mani V, Calcagno C, Li K, Li S, Shan H, Lv J, Zhao T, Xia J, Long Q, Steinberger S, Jacobi A, Deyer T, Luksza M, Liu F, Little BP, Fayad ZA, Yang Y (2020) Artificial intelligence-enabled rapid diagnosis of patients with covid-19. Nat Med 26(8):1224–1228. https://doi.org/10.1038/s41591-020-0931-3
57. Metz CE (1978) Basic principles of ROC analysis. Semin Nucl Med 8(4):283–298. https://doi.org/10.1016/S0001-2998(78)80014-2
58. Miller T (2019) Explanation in artificial intelligence: insights from the social sciences. Artif Intell 267:1–38. https://doi.org/10.1016/j.artint.2018.07.007
59. Milletari F, Navab N, Ahmadi S (2016) V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV), pp 565–571. https://doi.org/10.1109/3DV.2016.79
60. Narayan Das N, Kumar N, Kaur M, Kumar V, Singh D (2020) Automated deep transfer learning-based approach for detection of covid-19 infection in chest X-rays. IRBM. https://doi.org/10.1016/j.irbm.2020.07.001
61. Narin A, Kaya C, Pamuk Z (2020) Automatic detection of coronavirus disease (covid-19) using X-ray images and deep convolutional neural networks. arXiv preprint arXiv:200310849
62. Odena A, Olah C, Shlens J (2017) Conditional image synthesis with auxiliary classifier GANs. In: PMLR, international convention centre, Sydney, Australia, proceedings of machine learning research, vol 70, pp 2642–2651
63. Oh Y, Park S, Ye JC (2020) Deep learning covid-19 features on CXR using limited training data sets. IEEE Trans Med Imaging 39(8):2688–2700. https://doi.org/10.1109/TMI.2020.2993291
64. Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Rajendra Acharya U (2020) Automated detection of covid-19 cases using deep neural networks with X-ray images. Comput Biol Med 121:103792. https://doi.org/10.1016/j.compbiomed.2020.103792
65. Park CY, Villafuerte J, Abiad A (2020) Updated assessment of the potential economic impact of covid-19. https://doi.org/10.22617/BRF200144-2
66. Pereira S, Pinto A, Alves V, Silva CA (2016) Brain tumor segmentation using convolutional neural networks in MRI images. IEEE Trans Med Imaging 35(5):1240–1251. https://doi.org/10.1109/TMI.2016.2538465
67. Pizer SM, Johnston RE, Ericksen JP, Yankaskas BC, Muller KE (1990) Contrast-limited adaptive histogram equalization: speed and effectiveness. In: [1990] Proceedings of the first conference on visualization in biomedical computing, pp 337–345. https://doi.org/10.1109/VBC.1990.109340
68. Pronker ES, Weenen TC, Commandeur H, Claassen EHJHM, Osterhaus ADME (2013) Risk in vaccine research and development quantified. PLOS ONE 8(3):1–7. https://doi.org/10.1371/journal.pone.0057755
69. Radford A, Metz L, Chintala S (2015) Unsupervised representation learning with deep convolutional generative adversarial networks. In: International conference on learning representation (ICLR), pp 1–16
70. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz C, Shpanskaya K, et al. (2017) Chexnet: radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:171105225
71. Ranney ML, Griffeth V, Jha AK (2020) Critical supply shortages: the need for ventilators and personal protective equipment during

the covid-19 pandemic. New Engl J Med 382(18):e41. https://doi.org/10.1056/NEJMp2006141

72. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF (eds) Medical image computing and computer-assisted intervention: MICCAI 2015. Springer, Cham, pp 234–241
73. RSNA (2019) RSNA pneumonia detection challenge. https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data
74. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision (ICCV)
75. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: International conference on learning representation (ICLR)
76. Singh RK, Gorantla R (2020) Dmenet: diabetic macular edema diagnosis using hierarchical ensemble of CNNS. PLOS ONE 15(2):1–22. https://doi.org/10.1371/journal.pone.0220677
77. Stark JA (2000) Adaptive image contrast enhancement using generalizations of histogram equalization. IEEE Trans Image Process 9(5):889–896. https://doi.org/10.1109/83.841534
78. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)
79. Tan M, Le QV (2019) Efficientnet: rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:190511946
80. Tawsifur R (2019) COVID-19 radiography database. https://www.kaggle.com/tawsifurrahman/covid19-radiography-database
81. Tong ZD, Tang A, Li KF, Li P, Wang HL, Yi JP, Zhang YL, Yan JB (2020) Potential presymptomatic transmission of sars-cov-2, Zhejiang province, China, 2020. Emerg Infect Dis 26(5):1052
82. Tsai CF (2003) Stacked generalisation: a novel solution to bridge the semantic gap for content-based image retrieval. Online Inform Rev. https://doi.org/10.1108/14684520310510091
83. Tsiknakis N, Trivizakis E, Vassalou EE, Papadakis GZ, Spandidos DA, Tsatsakis A, Sánchez-García J, López-González R, Papanikolaou N, Karantanas AH et al (2020) Interpretable artificial intelligence framework for covid-19 screening on chest X-rays. Exp Ther Med 20(2):727–735
84. Ucar F, Korkmaz D (2020) Covidiagnosis-net: deep bayes-squeezenet based diagnosis of the coronavirus disease 2019 (covid-19) from X-ray images. Med Hypotheses 140:109761. https://doi.org/10.1016/j.mehy.2020.109761
85. Van Rijsbergen C (1979) Information retrieval. Butterworth-Heinemann, MA, USA
86. Vynnycky E, Trindall A, Mangtani P (2007) Estimates of the reproduction numbers of Spanish influenza using morbidity data. Int J Epidemiol 36(4):881–889. https://doi.org/10.1093/ije/dym071
87. Wang L, Wong A (2020) Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest X-ray images. arXiv preprint arXiv:200309871
88. WHO, et al. (2020a) Modes of transmission of virus causing covid-19: implications for ipc precaution recommendations: scientific brief, 27 march 2020. WHO/2019-nCoV/Sci_Brief/Transmission_modes/2020.2
89. WHO, et al. (2020b) Use of chest imaging in covid-19: a rapid advice guide, 11 june 2020. WHO/2019-nCoV/Clinical/Radiology_imaging/2020.1
90. Wolpert DH (1992) Stacked generalization. Neural Netw 5(2):241–259. https://doi.org/10.1016/S0893-6080(05)80023-1
91. Wong HYF, Lam HYS, Fong AHT, Leung ST, Chin TWY, Lo CSY, Lui MMS, Lee JCY, Chiu KWH, Chung T, et al. (2020) Frequency and distribution of chest radiographic findings in covid-19 positive patients. Radiology. https://doi.org/10.1148/radiol.2020201160
92. Yen SJ, Lee YS (2009) Cluster-based under-sampling approaches for imbalanced data distributions. Expert Syst Appl 36(3, Part 1):5718–5727. https://doi.org/10.1016/j.eswa.2008.06.108
93. Yu P, Zhu J, Zhang Z, Han Y (2020) A familial cluster of infection associated with the 2019 novel coronavirus indicating possible person-to-person transmission during the incubation period. J Infect Dis 221(11):1757–1761. https://doi.org/10.1093/infdis/jiaa077
94. Zhang J, Xie Y, Liao Z, Pang G, Verjans J, Li W, Sun Z, He J, Li Y, Shen C, et al. (2020) Viral pneumonia screening on chest X-ray images using confidence-aware anomaly detection. arXiv preprint arXiv:200312338
95. Zhou Q, Gao Y, Wang X, Liu R, Du P, Wang X, Zhang X, Lu S, Wang Z, Shi Q, Li W, Ma Y, Luo X, Fukuoka T, Ahn HS, Lee MS, Liu E, Chen Y, Luo Z, Yang K (2020) Nosocomial infections among patients with covid-19, sars and mers: a rapid review and meta-analysis. medRxiv https://doi.org/10.1101/2020.04.14.20065730
96. Zijdenbos AP, Dawant BM, Margolin RA, Palmer AC (1994) Morphometric analysis of white matter lesions in MR images: method and validation. IEEE Trans Med Imaging 13(4):716–724. https://doi.org/10.1109/42.363096

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