Noname manuscript No.
(will be inserted by the editor)

COVIDScreen: Explainable deep learning framework for differential diagnosis of COVID-19 using chest X-Rays

Rajeev Kumar Singh · Rohan Pandey · Rishie Nadhan Babu

Received: date / Accepted: date

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 delay in obtaining test results calls for alternative methods of rapid and reliable diagnosis. In this article, we propose a novel Deep Learning based solution to rapidly classify COVID -19 patient using chest X-Ray. 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-Ray into three classes viz. COVID-19, Pneumonia and Normal. An effective pruning strategy as introduced in the proposed framework results in increased model performance, generalisability, and decreased model complexity. We incorporate explainability in our article by using Grad-CAM visualisation 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

Rajeev Kumar Singh
Shiv Nadar University
E-mail: rajeev.kumar@snu.edu.in

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 6,48,000 humans and infecting close to 16.2 million people as on 27th July 2020 [76]. With almost 188 countries getting affected and with an estimated financial loss of 8.8 trillion USD where the economy is expected to contract by 5.2 percent, 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 [55].

The basic reproduction number R_0, represents the average number of people who could be infected by an infected person and this value indicates 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 average R_0 is 3.28 [74,43]. Fig 1 gives an overview of how the COVID-19 cases have increased in the last few months.

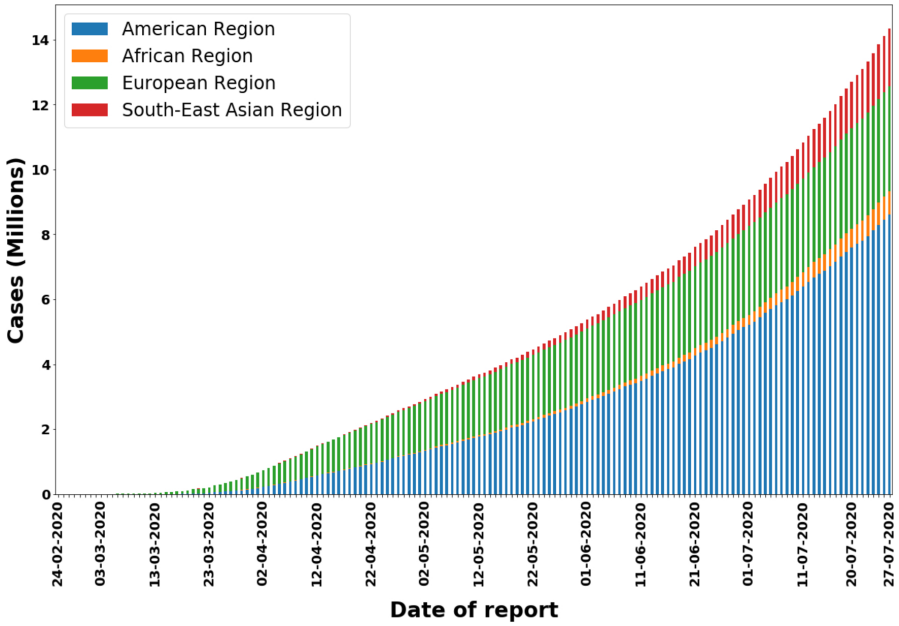

Fig. 1 Number of COVID-19 cases reported worldwide, as per WHO (21July202). These numbers are divided according to WHO regions [19].

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

sion may also happen due to touching of surfaces or objects contaminated with this virus [77]. The ease of transmission of the virus along with the high density of population in many countries forced nations to shut down industries and restrict the movement of people. Countries followed WHO guidelines of rigorous tracking, contact tracing, rapid diagnosis and immediate isolation of cases. These approaches faced roadblocks due to the gradual reopening of economic activities, which saw social distancing norms getting eased while testing kits were in short supply. The fact that the incubation period of this novel virus in an average 5-6 days and can be as long as 14-21 days makes the whole process of testing more complicated since one can infect many before they themselves becomes ill [82, 41, 45, 34, 3, 82, 70]. Modelling study by [27] estimates that many cases are asymptotic and upto 44 % of transmission may have already happened before the symptoms appear in a person. Thereby, early diagnosis of COVID-19 is crucial for timely referral of the patients to quarantine, rapid incubation of serious cases in specialised hospitals and containing the spread of this disease. [83] 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. 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 lower number of people getting tested daily [1]. The test may take up to 2 days to produce results [49]. 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, pre-clinical stage, clinical trials, regulatory review, manufacturing and quality control [58]. 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 [18].

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 use of Chest radiographs in case RT-PCR is not available or results are delayed [78]. Chest X-Rays are less-resource intensive and is 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 minimise the risk of infection while travelling for getting tested. Patients with 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 which 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 [36, 56, 12, 42, 21, 66]. Considering the success and potential promise 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 generalisable results. The research community has therefore been pooling a lot of data to further enhance the knowledge bank. Motivated from the recent progress made by the scientific community, we propose to explore the use of chest X-Rays for 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 propose to preprocess the image by using noise attenuation, contrast enhancement along with using 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 [64].

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 generalisability 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 organised as given. Section 2 gives a basic introduction to related work in this domain. Section 3 which is named as Materials and Methods has been divided into multiple subsections. Subsection 3.1 describes the data sets used for training validation and testing whereas subsection 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 [61]. 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 setting, 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 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, Cariomegaly, Abnormal Mediastinum, Pulmonary Edema, Tuberculosis, etc [32, 60, 5, 6, 30].

Chest radiography can potentially be the first-line imaging modality used for patients with suspected COVID-19 [80]. Chest radiography is a fast and relatively inexpensive imaging modality which 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 [38, 37]. The existence of X-Ray laboratories across the globe coupled with reliable imaging methodologies can potentially ease the pressure of the front-line COVID-19 warriors. [2] evaluated the performance of state-of-the-art convolutional neural networks including MobileNet V2, VGG-19, Inception, Xception and Inception ResNet V2. The

work by [72] proposes a SqueezeNet based architecture tuned for the COVID-19 diagnosis with Bayes optimization along with the validation phase. [75] have proposed a Deep Convolutional Neural Network design named COVID-Net using a lightweight residual projection-expansion projection-extension design pattern. The work by [54] 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.

Many recent studies have thus highlighted the significance of deep learning for detection of patients with COVID-19 using chest X-Rays. Majority of these studies have focused on proposing new deep learning architectures and exploring the feasibility of existing architectures for COVID-19 detection. The proposed framework in this paper explores various hypotheses testings to justify decisions taken during model training. This study also aims to reduce computational cost and modelling complexity while simultaneously improving model efficiency through the implementation of U-Net based segmentation and a proposed pruned ensemble framework. This research also addresses the issue of explanability by means of Grad-CAM visualization. The usefulness of GANs for the task of image augmentation is suitably explored to Improve the model performance.

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 following 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 overlap of images in the publically available COVID-19 dataset collections, we provide the number of unique samples in each class of these datasets.

Class	Sources	Samples
COVID-19	COVID-19 image data collection[17]	422
	Figure 1 COVID-19 Chest X-Rays[14]	35
	Actualmed COVID-19 Chest X-Rays[13]	58
	COVID-19 Radiography Database [69]	58
Pneumonia	RSNA Pneumonia Detection Challenge[63]	6041
Normal	RSNA Pneumonia Detection Challenge[63]	8851

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

Unbalanced data is a common problem in the image classification task wherein some classes have fewer samples as compared to others. This issue has

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

Table 2 Dataset splits used for classification task in this study.

the potential to make deep CNNs profoundly biased against the less frequent class [35]. 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 balanced distribution of all training classes by undersampling Pneumonia and Normal class [81], In dataset B we upsample the COVID-19 class using Random rotation of 25%, Horizontal flipping and Gaussian Blur [23]. 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 [39]. 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 the Shenzhen and Montgomery County datasets consisting of 662 and 138 chest X-Ray samples respectively. Both the datasets include manifestation of Tuberculosis and normal cases along with their respective masks [31].

Weight Initialization: In order to test the effect of weight initialization, we use the CheXpert dataset [29]. 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 initial weights for the classification task.

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 seg-

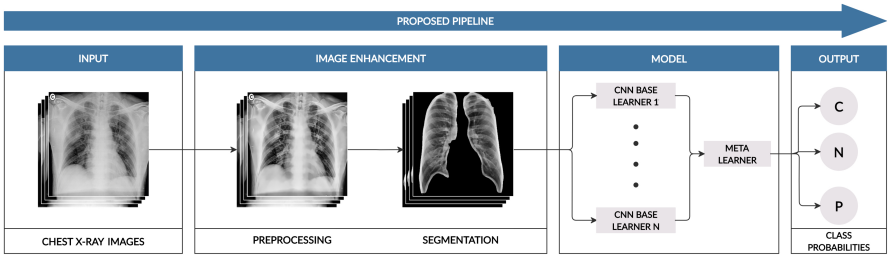

Fig. 2 Graphical Abstract: The proposed pipeline works as follows: The input images (chest X-Rays) are passed through preprocessing stage followed by segmentation. These images are then fed simultaneously into multiple base learners to generate class probabilities which are then passed into the meta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

mentation and classification [62,44,68,40]. 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 [16,52,15,25].

Thanks to the excellent efficiency of CNN architectures in medical imaging activities, we use the following state of the art standard architectures as baseline for the proposed pipeline: VGG-19, VGG-16, ResNet-50, DenseNet-161 and DenseNet-169 [65,26,28]. 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 [20] :

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

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.

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 Epoch. Grid search builds a model for every combination of hyperparameters specified and evaluates each model accordingly. Grid search has thus, 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. The search ranges were $[1e-15, 1e-1]$, $[0.85, 0.99]$, $[0.1, 0.8]$, and $[1e-10, 1e-3]$ respectively. Model checkpoints has 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 Equalisation 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 [7,33].

The first step in the proposed preprocessing pipeline includes a variant of histogram equalisation referred to as CLAHE, Contrast Limited Adaptive Histogram Equalisation, which is frequently used to enhance different types of medical images [57]. This technique effectively spreads out the most frequent intensity values in the images. The more common Histogram Equalisation technique considers the global contrast of the image which can sometimes lead to a loss of information due to over-brightness [67]. CLAHE acts as an alternative which divides the image into smaller blocks called 'tiles' and each of these tiles are histogram equalised to confine the spread of intensity values to that particular region using the general histogram equalization formula:

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

here 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 emperically 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. Fig.3 suitably demonstrates the impact of CLAHE by showing the constrast using COVID-19 sample.

To prepare images for further processing such as segmentation and classification, certain image denoising filters capable of removing 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

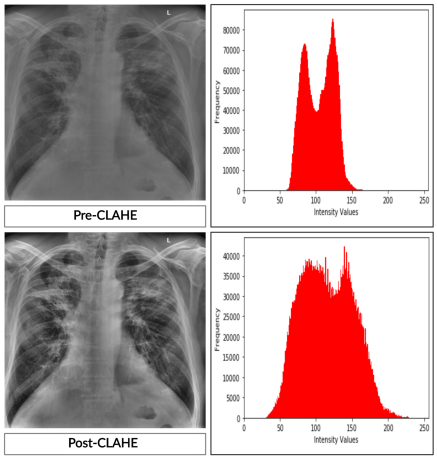

Fig. 3 Sample images of COVID-19 pre and post CLAHE along with histogram.

loss of information in the image compared with local mean algorithms [10,9, 8].

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 pixel 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 image samples [4]. This non-local means filter is characterized by the following function [10]:

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

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 weighted average for a given pixel and h, regulates the filter strength having a tradeoff between removal of noise and image detail.

The best fitting parameters have been empirically found to be *templateWindowSize=7*, *searchWindowSize=21* and h=7 post experimenting 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 two-fold benefit of superior model performance and reduced computational cost [16,11]. In this study, the dataset used for training has been derived from various data sources as mentioned in section 3.1, forming three classes viz 'COVID-19', 'normal', 'pneumonia'.

Qualitative exploration of data reveals that the image widths and heights across these 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 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 [62].

A brief overview of U-Net architecture is described to give insight to the reader. 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 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 section 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 optimised using Adam optimiser with 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:

$$PixelAccuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{4}$$

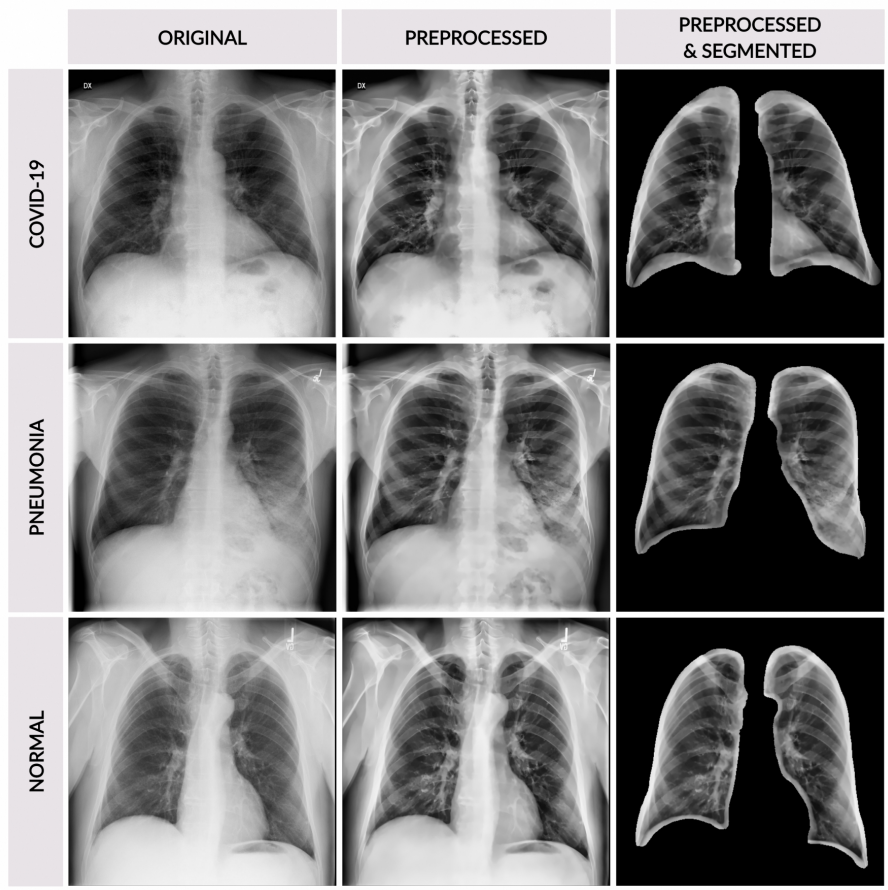

Fig. 4 Sample images of COVID-19, Normal and Pneumonia classes at various stages through the proposed pipeline.

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 dataset. 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 generalisability 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 generalisability. The accuracy of predictions made by a set of base learners is often better than a single best learner [24]. 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 noramalises the output scores into a probability distribution:

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

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 confidence or performance of the specific base learners. 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.

The meta-learner hypothesis function takes predictions made by the base-models as input and learns to combine them to make a more accurate, robust output prediction. This is referred to as the Stacked Generalisation ensemble technique and can result in improved predictive performance than any individual base-learner [71,79]. In this study, we propose a modified stacked generalisation 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 deploy 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

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 ← B - $Base_{remove}$, S - $C_{Base_{remove}}$

training of all base-learners b_i as mentioned in section 3.2.1 on $train_{base}$ which is the training set. By freezing the base-learner weight updates, these models then use the hold-out set $train_{meta}$ 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 [71]. The meta-learner models are thus effectively trained on this holdout set $train_{meta}$ to avoid overfitting. We now evaluate each meta-learner on the test set, and choose the meta-learner $META_{best}$ 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 $META_{best}$ and starting with all the N base-learners, we use a pruning approach as shown algo-

rithm 1 to remove redundant base-learners resulting in increased model performance, generalisability, 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 generalisability 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, 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 large-scale availability of labeled training data. Acquisition and labelling of medical image data is 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:

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

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

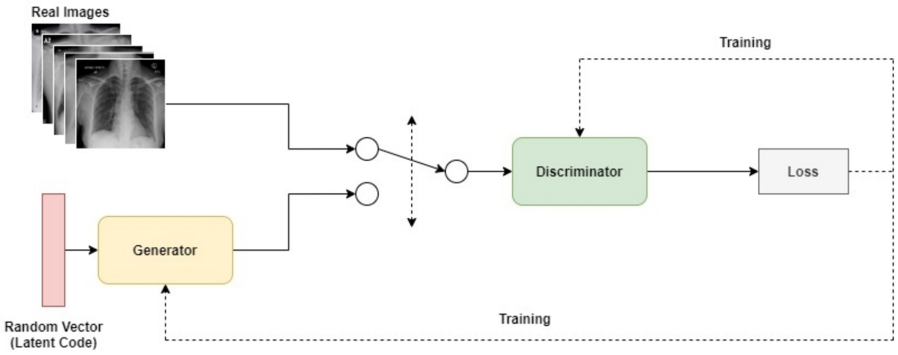

Fig. 5 General GAN Architecture. The real images and synthetic images generated using the Generator and fed into the discriminator. Using the WGAN-GP loss function backpropagation is done to improve both the networks.

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 the model and low generalizability of certain models [46]. 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. We are really not going to trust a forecast if we don't understand why it was made [51].

We deploy gradient-weighted class activation mapping known as Grad-CAM that is class-discriminatory and locates relevant regions of image. It is a gradient-based visualization method which calculates the scores in a trained model for a given image category using the feature maps of the deepest convolutional layer [64]. 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 huge chal-

lenge 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 proposed question of multi-class classification.

4.1.1 Overall Accuracy (OA)

Overall Accuracy is a metric for evaluating classification models [50]. Informally, overall accuracy is the fraction of predictions the model gets right. Formally, it is defined as:

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

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 [73]. A low precision will correspond to high false positives and would result in unwanted burden on the health care systems, catering to patients incorrectly classified as having a 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_p) [73]. 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:

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

4.1.4 F1-Score (F1)

F1- Score is defined as the harmonic mean of precision and recall [84]. 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 maximised. 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 [48]. 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 is estimated using a per-annotator empirical prior over the class labels.

4.2 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 has been specified in section 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.

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

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

Table 3 Performance metrics of base learners 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

Table 4 Performance metrics of VGG–19 using different training techniques, where C,N and P are COVID-19, Normal and Pneumonia respectively. Here TopHead refers to training just the custom head, and entire refers to training the entire network.

4.4 Hypothesis 3: Weight Initialization

An important hyper parameter tuning required during deep neural networks 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.

4.5 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 section 3.1.

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

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

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.

The dataset split as mentioned in table 2 use 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 indicate that upsampling and unbalanced training sets do not improve model improve performance on the hold-out test set. As a consequence, downsampled balanced class distribution, Dataset A, is used for further hypothesis testing.

4.6 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 in table 8. Observing the results in table 7, 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 expectated results. For further hypothesis testing, we have used the Preprocessed and Segmented dataset for evaluation.

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 7 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.

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

Table 8 Enhancement techniques used for hypothesis 5.

4.7 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 section 3.2.4. To evaluate the performance of various meta-learner algorithms, we have deployed all base learners, i.e VGG-19, VGG-16, DenseNet-121, DenseNet-169 and ResNet-50 retrained using the best obtained training methods as shown in table 9, and then iteratively used various meta-learners to evaluate their performance. It is clear from the table 10 that Naive Bayes outperformed all other Hypothesis functions with a high overall accuracy of 98 and a Kappa score of 0.97. The precision and recall for each class as summarised in table 10 reveal that Naive Bayes hypothesis function as a meta-learner offers a high degree of generalisability. 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 finalise the optimum hyperparameters for all Hypothesis functions before evaluating them as a meta-learner in the ensemble set-up.

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

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.

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

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.

The various hyperparameters used in the grid search for this hypothesis are: learning rate for XGBoost algorithm; kernel type, gamma for Support Vector Machines; regularisation 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 of layers for Neural Network. Naive Bayes algorithm does not have any hyper-parameters.

At the end of the pruning process as mentioned in section 3.2.4, among all base-learner functions in the member set, VGG-16, ResNet-50, DenseNet-169

Study	Accuracy	Kappa Score		Precision	Recall	F-1 Score
Pruned Naive	98.67	0.98	C	100	100	100
			N	98	98	98
			P	98	98	98
Wang & Wong	93.34	0.9	C	98.91	91	94.79026907
			N	90.47	95	92.67967865
			P	91.26	94	92.60973767
Oh, Park & Ye	88.9	-	C	76.9	100	86.94177501
			N	95.7	90	92.76252019
			P	90.3	93	91.63011457
Khan, Shah & Bhat	90.21	0.83	C	97	89	92.82795699
			N	92	85	88.36158192
			P	87	95	90.82417582
Ozturk et al.	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

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

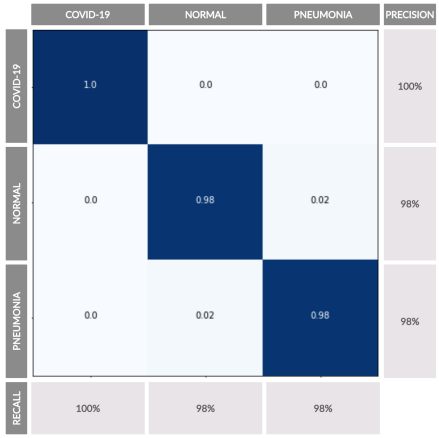

Fig. 6 The final confusion matrix obtained on test set, with X-axis representing the actual class labels and Y-axis representing the predicted class labels from the final pruned ensemble model.

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 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 the best performing model are shown in Fig.7 and Fig.6 respectively. The Grad-CAM visualisation by four final base-learners for COVID-19 sample has been shown in Fig.8.

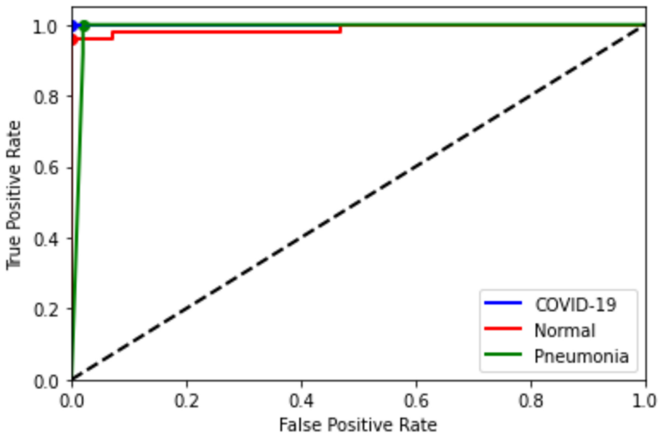

Fig. 7 Receiver operating characteristic curve for the final pruned ensemble model.

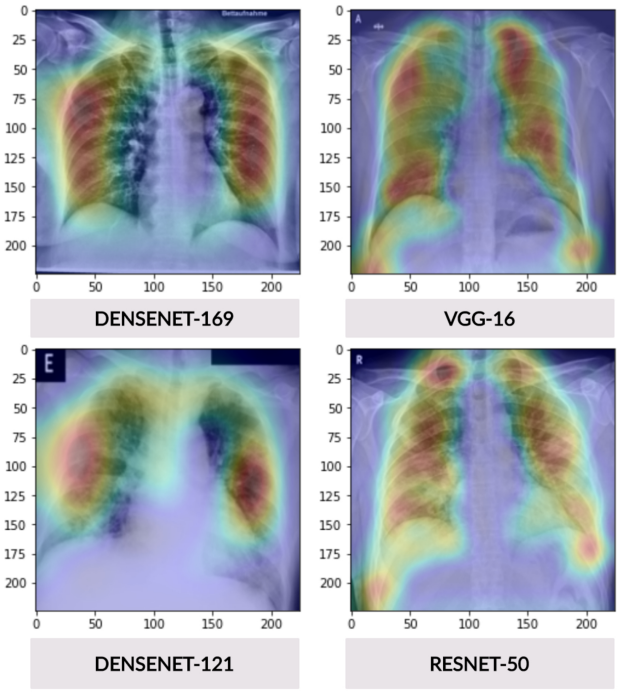

Fig. 8 COVID-19 Grad-CAM visualization for the base learners used in the final model.

4.8 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 have been trained and images generated using each of these methods are shown in Fig 9 [22,53,47,59]. 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 have been used for training the WGAN with $\lambda = 10$, ncritic $= 5$, α $=0.0001$, $\beta_1 = 0$, $\beta_2 = 0.9$ [22], adam optimizer with learning rate of 0.0002, a latent vector of 100 dimensions and a batch size of 64 has been used [22].

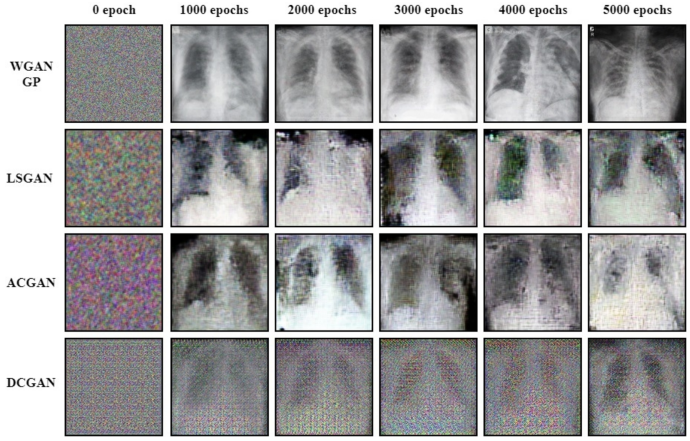

Fig. 9 Images Generated by various GAN Models. The models explored include Wasserstein GAN (WGAN), Least Squares GAN (LSGAN), Auxiliary Classifier GAN (ACGAN), and Deep Convolution GAN. The synthetic images generated using WGAN are extremely realistic.

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 summarised 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 analyzing 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 have been minimal, which has hindered fast progress of research studies. To address this issue, we have deployed multiple Generative Adversarial Networks and qualitatively evaluated these samples through visual inspection. 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, false negative error rate is far more expensive than false positive error rate since 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 manual segmentation of some of the COVID-19 chest X-Rays highlighting the importance of collaboration between AI and medical community.

The proposed model does have some limitations. One major limitation of this study is the small sample size. Despite the promising results of using the AI model to screen patients with COVID-19, further data collection is required to test the generalisability of the AI model to other patient populations. 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.

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 is also reducing monotonous procedural elements such as examining chest X-Ray images thus, allowing doctors to focus

on more demanding tasks. In conclusion, these results illustrate the potential role for a highly accurate AI algorithm for the rapid identification of COVID-19 patients, which could be helpful in combating the current disease outbreak. We believe the AI model proposed could be a useful screening tool to quickly diagnose infectious diseases such as COVID-19 that does not require radiologist input or physical tests.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Adams, H.J., Kwee, T.C., Yakar, D., Hope, M.D., Kwee, R.M.: Chest ct imaging signature of covid-19 infection: in pursuit of the scientific evidence. Chest (2020)
2. Apostolopoulos, I.D., Mpesiana, T.A.: Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. Physical and Engineering Sciences in Medicine p. 1 (2020)
3. Arons, M.M., Hatfield, K.M., Reddy, S.C., Kimball, A., James, A., Jacobs, J.R., Taylor, J., Spicer, K., Bardossy, A.C., Oakley, L.P., et al.: Presymptomatic sars-cov-2 infections and transmission in a skilled nursing facility. New England journal of medicine (2020)
4. Baozhong, L., Jianbin, L.: Overview of image noise reduction based on non-local mean algorithm. In: MATEC Web of Conferences, vol. 232, p. 03029. EDP Sciences (2018)
5. Bar, Y., Diamant, I., Wolf, L., Greenspan, H.: Deep learning with non-medical training used for chest pathology identification. In: Medical Imaging 2015: Computer-Aided Diagnosis, vol. 9414, p. 94140V. International Society for Optics and Photonics (2015)
6. Bar, Y., Diamant, I., Wolf, L., Lieberman, S., Konen, E., Greenspan, H.: Chest pathology detection using deep learning with non-medical training. In: 2015 IEEE 12th international symposium on biomedical imaging (ISBI), pp. 294–297. IEEE (2015)
7. Bieniecki, W., Grabowski, S., Rozenberg, W.: Image preprocessing for improving ocr accuracy. In: 2007 International Conference on Perspective Technologies and Methods in MEMS Design, pp. 75–80. IEEE (2007)
8. Buades, A.: A non-local algorithm for image denoising computer vision and pattern recognition, 2005. 2: 60–65. Google Scholar Google Scholar Digital Library Digital Library (2005)
9. Buades, A., Coll, B., Morel, J.M.: On image denoising methods. CMLA Preprint **5** (2004)
10. Buades, A., Coll, B., Morel, J.M.: Non-local means denoising. Image Processing On Line **1**, 208–212 (2011)
11. Chen, H., Dou, Q., Yu, L., Heng, P.A.: Voxresnet: Deep voxelwise residual networks for volumetric brain segmentation. arXiv preprint arXiv:1608.05895 (2016)
12. Cheng, J.Z., Ni, D., Chou, Y.H., Qin, J., Tiu, C.M., Chang, Y.C., Huang, C.S., Shen, D., Chen, C.M.: Computer-aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in ct scans. Scientific reports **6**(1), 1–13 (2016)
13. Chung, A.: Actualmed covid-19 chest x-ray data initiative. `https://github.com/agchung/Actualmed-COVID-chestxray-dataset` (2020)
14. Chung, A.: Figure 1 covid-19 chest x-ray data initiative. `https://github.com/agchung/Figure1-COVID-chestxray-dataset` (2020)
15. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention, pp. 424–432. Springer (2016)

16. Ciresan, D., Giusti, A., Gambardella, L.M., Schmidhuber, J.: Deep neural networks segment neuronal membranes in electron microscopy images. In: Advances in neural information processing systems, pp. 2843–2851 (2012)
17. Cohen, J.P., Morrison, P., Dao, L.: Covid-19 image data collection. arXiv 2003.11597 (2020). URL https://github.com/ieee8023/covid-chestxray-dataset
18. Deming, M.E., Michael, N.L., Robb, M., Cohen, M.S., Neuzil, K.M.: Accelerating development of sars-cov-2 vaccines—the role for controlled human infection models. New England Journal of Medicine (2020)
19. Geneva: World Health Organization, .: WHO coronavirus disease (COVID-19) dashboard. (2020 (accessed July 27, 2020)). URL https://covid19.who.int/
20. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)
21. Gorantla, R., Singh, R.K., Pandey, R., Jain, M.: Cervical cancer diagnosis using cervixnet-a deep learning approach. In: BIBE, pp. 397–404 (2019)
22. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Advances in neural information processing systems, pp. 5767–5777 (2017)
23. Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G.: Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications **73**, 220–239 (2017)
24. Hansen, L.K., Salamon, P.: Neural network ensembles. IEEE transactions on pattern analysis and machine intelligence **12**(10), 993–1001 (1990)
25. Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.M., Larochelle, H.: Brain tumor segmentation with deep neural networks. Medical image analysis **35**, 18–31 (2017)
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)
27. He, X., Lau, E.H., Wu, P., Deng, X., Wang, J., Hao, X., Lau, Y.C., Wong, J.Y., Guan, Y., Tan, X., et al.: Temporal dynamics in viral shedding and transmissibility of covid-19. Nature medicine **26**(5), 672–675 (2020)
28. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708 (2017)
29. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 590–597 (2019)
30. Islam, M.T., Aowal, M.A., Minhaz, A.T., Ashraf, K.: Abnormality detection and localization in chest x-rays using deep convolutional neural networks. arXiv preprint arXiv:1705.09850 (2017)
31. Jaeger, S., Candemir, S., Antani, S., Wáng, Y.X.J., Lu, P.X., Thoma, G.: Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. Quantitative imaging in medicine and surgery **4**(6), 475 (2014)
32. Jaiswal, A.K., Tiwari, P., Kumar, S., Gupta, D., Khanna, A., Rodrigues, J.J.: Identifying pneumonia in chest x-rays: A deep learning approach. Measurement **145**, 511–518 (2019)
33. Jamal, I., Akram, M.U., Tariq, A.: Retinal image preprocessing: background and noise segmentation. Telkomnika **10**(3), 537–544 (2012)
34. Jang, S., Han, S.H., Rhee, J.Y.: Cluster of coronavirus disease associated with fitness dance classes, south korea. Emerging infectious diseases **26**(8) (2020)
35. Johnson, J.M., Khoshgoftaar, T.M.: Survey on deep learning with class imbalance. Journal of Big Data **6**(1), 27 (2019)
36. Kallenberg, M., Petersen, K., Nielsen, M., Ng, A.Y., Diao, P., Igel, C., Vachon, C.M., Holland, K., Winkel, R.R., Karssemeijer, N., et al.: Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. IEEE transactions on medical imaging **35**(5), 1322–1331 (2016)
37. Kanne, J.P., Little, B.P., Chung, J.H., Elicker, B.M., Ketai, L.H.: Essentials for radiologists on covid-19: an update—radiology scientific expert panel (2020)

38. Kooraki, S., Hosseiny, M., Myers, L., Gholamrezanezhad, A.: Coronavirus (covid-19) outbreak: what the department of radiology should know. Journal of the American college of radiology (2020)
39. Krawczyk, B.: Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence **5**(4), 221–232 (2016)
40. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp. 1097–1105 (2012)
41. Lauer, S.A., Grantz, K.H., Bi, Q., Jones, F.K., Zheng, Q., Meredith, H.R., Azman, A.S., Reich, N.G., Lessler, J.: The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: estimation and application. Annals of internal medicine **172**(9), 577–582 (2020)
42. Liu, S., Liu, S., Cai, W., Pujol, S., Kikinis, R., Feng, D.: Early diagnosis of alzheimer's disease with deep learning. In: 2014 IEEE 11th international symposium on biomedical imaging (ISBI), pp. 1015–1018. IEEE (2014)
43. Liu, Y., Gayle, A.A., Wilder-Smith, A., Rocklöv, J.: The reproductive number of covid-19 is higher compared to sars coronavirus. Journal of travel medicine (2020)
44. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440 (2015)
45. Luo, L., Liu, D., Liao, X.l., Wu, X.b., Jing, Q.l., Zheng, J.z., Liu, F.h., Yang, S.g., Bi, B., Li, Z.h., et al.: Modes of contact and risk of transmission in covid-19 among close contacts. medRxiv (2020)
46. Maguolo, G., Nanni, L.: A critic evaluation of methods for covid-19 automatic detection from x-ray images. arXiv preprint arXiv:2004.12823 (2020)
47. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp. 2794–2802 (2017)
48. McHugh, M.L.: Interrater reliability: the kappa statistic. Biochemia medica: Biochemia medica **22**(3), 276–282 (2012)
49. Mei, X., Lee, H.C., Diao, K.y., Huang, M., Lin, B., Liu, C., Xie, Z., Ma, Y., Robson, P.M., Chung, M., et al.: Artificial intelligence–enabled rapid diagnosis of patients with covid-19. Nature Medicine pp. 1–5 (2020)
50. Metz, C.E.: Basic principles of roc analysis. In: Seminars in nuclear medicine, vol. 8, pp. 283–298. WB Saunders (1978)
51. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1–38 (2019)
52. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV), pp. 565–571. IEEE (2016)
53. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier gans. In: International conference on machine learning, pp. 2642–2651 (2017)
54. Oh, Y., Park, S., Ye, J.C.: Deep learning covid-19 features on cxr using limited training data sets. IEEE Transactions on Medical Imaging (2020)
55. Park, C.Y., Villafuerte, J., Abiad, A.: Updated assessment of the potential economic impact of covid-19 (2020)
56. Pereira, S., Pinto, A., Alves, V., Silva, C.A.: Brain tumor segmentation using convolutional neural networks in mri images. IEEE transactions on medical imaging **35**(5), 1240–1251 (2016)
57. Pizer, S.M.: Contrast-limited adaptive histogram equalization: Speed and effectiveness stephen m. pizer, r. eugene johnston, james p. ericksen, bonnie c. yankaskas, keith e. muller medical image display research group. In: Proceedings of the First Conference on Visualization in Biomedical Computing, Atlanta, Georgia, vol. 337 (1990)
58. Pronker, E.S., Weenen, T.C., Commandeur, H., Claassen, E.H., Osterhaus, A.D.: Risk in vaccine research and development quantified. PloS one **8**(3), e57755 (2013)
59. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)

60. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al.: Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225 (2017)
61. Ranney, M.L., Griffeth, V., Jha, A.K.: Critical supply shortages—the need for ventilators and personal protective equipment during the covid-19 pandemic. New England Journal of Medicine **382**(18), e41 (2020)
62. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, pp. 234–241. Springer (2015)
63. RSNA: RSNA Pneumonia Detection Challenge (2019). URL `https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data`
64. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp. 618–626 (2017)
65. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
66. Singh, R.K., Gorantla, R.: Dmenet: Diabetic macular edema diagnosis using hierarchical ensemble of cnns. Plos one **15**(2), e0220677 (2020)
67. Stark, J.A.: Adaptive image contrast enhancement using generalizations of histogram equalization. IEEE Transactions on image processing **9**(5), 889–896 (2000)
68. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9 (2015)
69. Tawsifur, R.: COVID-19 radiography database. (2019). URL `https://www.kaggle.com/tawsifurrahman/covid19-radiography-database`
70. Tong, Z.D., Tang, A., Li, K.F., Li, P., Wang, H.L., Yi, J.P., Zhang, Y.L., Yan, J.B.: Potential presymptomatic transmission of sars-cov-2, zhejiang province, china, 2020. Emerging infectious diseases **26**(5), 1052 (2020)
71. Tsai, C.F.: Stacked generalisation: a novel solution to bridge the semantic gap for content-based image retrieval. Online Information Review (2003)
72. Ucar, F., Korkmaz, D.: Covidiagnosis-net: Deep bayes-squeezenet based diagnostic of the coronavirus disease 2019 (covid-19) from x-ray images. Medical Hypotheses p. 109761 (2020)
73. Van Rijsbergen, C., Van Rijsbergen, C.: Information Retrieval. Butterworths (1979). URL `https://books.google.co.in/books?id=t-pTAAAAMAAJ`
74. Vynnycky, E., Trindall, A., Mangtani, P.: Estimates of the reproduction numbers of spanish influenza using morbidity data. International Journal of Epidemiology **36**(4), 881–889 (2007)
75. Wang, L., Wong, A.: Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. arXiv preprint arXiv:2003.09871 (2020)
76. WHO, et al.: Coronavirus disease 2019 (COVID-19): situation report, 189 (2020)
77. WHO, et al.: Modes of transmission of virus causing covid-19: implications for ipc precaution recommendations: scientific brief, 27 march 2020. Tech. rep., World Health Organization (2020)
78. WHO, et al.: Use of chest imaging in covid-19: a rapid advice guide, 11 june 2020. Tech. rep., World Health Organization (2020)
79. Wolpert, D.H.: Stacked generalization. Neural networks **5**(2), 241–259 (1992)
80. Wong, H.Y.F., Lam, H.Y.S., Fong, A.H.T., Leung, S.T., Chin, T.W.Y., Lo, C.S.Y., Lui, M.M.S., Lee, J.C.Y., Chiu, K.W.H., Chung, T., et al.: Frequency and distribution of chest radiographic findings in covid-19 positive patients. Radiology p. 201160 (2020)
81. Yen, S.J., Lee, Y.S.: Cluster-based under-sampling approaches for imbalanced data distributions. Expert Systems with Applications **36**(3), 5718–5727 (2009)
82. Yu, P., Zhu, J., Zhang, Z., Han, Y.: A familial cluster of infection associated with the 2019 novel coronavirus indicating possible person-to-person transmission during the incubation period. The Journal of infectious diseases **221**(11), 1757–1761 (2020)
83. Zhou, Q., Gao, Y., Wang, X., Liu, R., Du, P., Wang, X., Zhang, X., Lu, S., Wang, Z., Shi, Q., et al.: Nosocomial infections among patients with covid-19, sars and mers: A rapid review and meta-analysis. medRxiv (2020)

84. Zijdenbos, A.P., Dawant, B.M., Margolin, R.A., Palmer, A.C.: Morphometric analysis of white matter lesions in mr images: method and validation. IEEE transactions on medical imaging **13**(4), 716–724 (1994)