CHAPTER 1

Deep learning-based hybrid models for prediction of COVID-19 using chest X-ray

R. Shree Charran[a] **and Rahul Kumar Dubey**[b]
[a]Department of Management Studies, Indian Institute of Science, Bengaluru, Karnataka, India
[b]Robert Bosch Engineering and Business Solutions Private Limited, Bengaluru, Karnataka, India

1 Introduction

The novel coronavirus disease-2019 pandemic is the biggest public health epidemic faced by mankind. The virus has spread to every habited continent since its arrival in Asia in late 2019. Across all developed and developing nations, the cases are rising daily. The Exponential spread of the infection has led to a severe shortage of accurate testing kits as they can't be manufactured fast enough, creating panic amongst the citizens of several countries. This has resulted in the selling of bogus COVID-19 test kits and other fake vaccines to the public. The limited availability of accurate diagnostic test kits has resulted in an urgent need to focus on other methods for diagnosis. As COVID-19 attacks the epithelial cells which line our respiratory tract, we can use X-rays to examine the health of the lungs of a patient. Furthermore, provided that all major hospitals have access to X-ray imaging equipment, without the special test sets, X-rays could be used to monitor for COVID-19.

Currently, the only complication lies with the fact that the chest X-rays of COVID-19 patients have similar abnormalities with a Pneumonia Infected patient. Exploration is in progress to completely understand how COVID-19 pneumonia contrasts with different sorts of pneumonia. Data from these investigations can conceivably help find and facilitate our comprehension of how SARS-CoV-2 influences the lungs. So far, scientists have found that individuals with COVID-19 pneumonia were bound to have: (1) pneumonia that influences the two lungs rather than only one (2) lungs that had a trademark "ground-glass" appearance by means of CT check (3) abnormalities in some research tests, especially those evaluating liver

Novel AI and Data Science Advancements for Sustainability in the Era of COVID-19
https://doi.org/10.1016/B978-0-323-90054-6.00001-5

capacity. This clearly indicates that there is considerable room for the use of AI in diagnosing COVID-19 and differentiating it from viral pneumonia.

The Computer Vision groups across the globe have made huge efforts over the last decade and made many State of the Art models open to the public. These State-of-the-art models are conditioned on various data types and can be fine-tuned for certain typical tasks and purposes. For this analysis want to harness the capabilities and predictive power of pre-trained models to classify between COVID-19, non-COVID Pneumonia, and Normal.

2 Related work

Rousan, Elobeid, Karrar, et al. (2020) studied that chest CT scans and chest X-rays show characteristic radiographic findings in patients with COVID-19 pneumonia. The study aims at describing the chest X-ray findings and temporal radiographic changes in COVID-19 patients. The authors studied the X-rays of 88 COVID-19 confirmed patients. A total of 190 chest X-rays were obtained for the 88 patients. Thirty-one percent of the X-rays showed visible abnormalities. The most common finding on chest X-rays was peripheral ground glass opacities affecting the lower lobes. In the course of illness, the opacities progressed into consolidations peaking around 6–11 days. Thus they conclude that Chest X-ray can be used in the diagnosis and follow Yee and Raymond (2020) developed a pneumonia predictor using feature extraction from Transfer Learning. InceptionV3 was used as the feature extractor. K-Nearest Neighbor, Neural Network, and Support Vector Machines were used to classify the extracted features. The Neural Network model achieved the highest sensitivity of 84.1%, followed by Support vector machines and K-Nearest Neighbor Algorithm. Among all the classification models, the support vector machines model achieved the highest AUC of 93.1% for patients with COVID-19 pneumonia. Barstugan, Ozkaya, and Ozturk (2020b) used machine learning algorithms to classify between COVID-19 and non-COVID-19 images. The authors considered feature extraction techniques like gray-level size zone matrix and discrete wavelet transform. The extracted features were classified using a support vector machine and 2-, 5-, and 10-fold cross-validation. The authors achieved 99.68% of accuracy for the SVM trained using the GLSZM feature extraction method. Wang, Zha, Li, et al. (2020) proposed the use of deep learning to distinguish COVID-19 and other pneumonia types. The authors segmented and eliminated irrelevant areas. DenseNet121-FPN was implemented for lung segmentation, and COVID19-Net that had a

DenseNet-like structure, was proposed for classification purposes. The authors reported 0.87 ROC and 0.88 AUC scores for the validation sets. Kassani, Kassasni, Wesolowski, Schneider, and Deters (2020) introduced a feature extractor-based multi-method ensemble approach for computer-aided analysis of COVID-19 pneumonia. Six Machine learning algorithms were trained on the features extracted by CNNs to find the best combination of features and learners. Considering the high visual complexity of image data, proper deep feature extraction is considered a critical step in developing deep CNN models. The experimental results on the chest X-ray datasets showed that the features extracted by Dense-Net-121 and finally trained using the Bagging tree classifier generate the best predictions with 99.00% classification accuracy. Wang and Wong (2020) introduced COVID-Net, to detect COVID-19 from X-ray images of the chest. The COVID-Net architecture was designed from a mixture of 1 × 1 convolutions, depth-wise convolution, and residual modules to allow for deeper system design and prevent the issue of gradient disappearing. The dataset given was a mix of the COVID chest X-ray dataset provided by Cohen, Morrison, and Dao (2020b), and Kaggle chest X-ray images dataset (Kaggle, 2020) for multi-class classification of multi-class classification of normal vs bacterial vs COVID-19 infection dataset. The obtained accuracy of this study was 83.5%. Khan, Shah, and Bhat (2020) proposed CoroNet, to automatically detect COVID-19 from chest X-ray images. Coronet was built using the Xception architecture with ImageNet weights. CoroNet achieved an overall accuracy of 89%, precision of 93% and recall of 98.2% for 4-class cases being COVID-19, Viral and bacterial Pneumonia and Healthy. The same model achieved 95% accuracy for 3-class classification i.e., COVID-19, Pneumonia and Healthy. Chouhan et al. (2020) proposed a deep learning approach to classify pneumonia from chest X-rays using State of the art pre-trained models. They tested the performances of State of the art pre-trained models like AlexNet, DenseNet, and Inception V3 etc. to extract features. The extracted features were passed through individual classifiers and the predictions of individual architectures were obtained. An overall ensemble of all five pretrained models was observed to outperform all other models. Rajaraman et al. (2020) studied and found that performing Reiterative pruning and selecting the best pruned model improved the prediction accuracies and further helped minimize parameter numbers as redundant parameters which do not help improve the prediction performance are eliminated. Further they were able to better the performance by use of ensembles of pruned models. Awarding weights based on their predictions, the authors observed that the weighted averaging ensemble of the pruned models outperformed the other ensemble methods. Overall it was identified that combinations of iterative pruning of models and ensembles of models helped reduce prediction variance, model complexity. In this chapter, we evaluate four different approaches/hybrids using State of the art pre trained models so to achieve maximum Accuracy and have low False Negatives.

3 Modeling

3.1 PCA-feature ensembles

The baseline models are initialized with ImageNet weights and are used to extract the image features. To act as a feature extractor, the final softmax layer is removed. The features extracted for all the baseline models are combined and reduced by using PCA. The number of PCA components is selected so as to explain 90% of the total variance. These PCA features are finally passed through a dense 256 layer and a softmax for final predictions. The architecture of PCA-Feature Ensembles for the baseline model is depicted in Fig. 1.

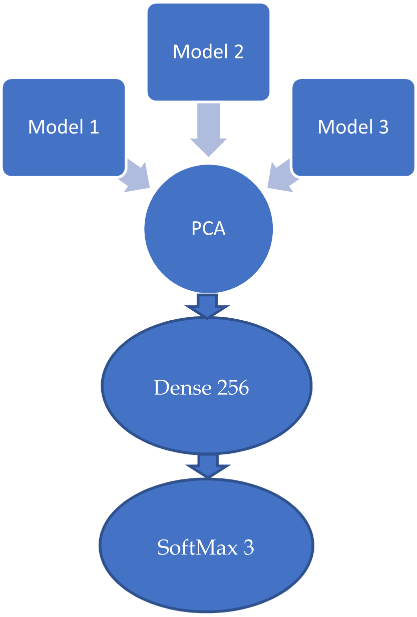

Fig. 1 Proposed feature ensemble for the baseline model.

3.2 Optimally weighted majority voting

This is a naïve but effective approach. The main Baseline models are individually assessed on the dataset and the probabilities prediction for all the classes are made. The prediction vector is a weighted average of the individual probabilities across all classes. The final prediction Y is the maximum probable class.

$$Y = \operatorname{argmax} \sum_{1}^{m} W_j \cdot P_{ij} \tag{1}$$

where W_j is the weight that can be assigned to the jth classifier.

The weights, W_j are calculated by a grid search so as to find best linear combination for most accuracy. Fig. 2 depicts the Weighted Majority Weighting ensemble.

3.3 Feature extraction

Feature extraction consists of using the representations learned by a previous network to extract distinguishing features from new samples. These features are then classified. The methodology involves (i) extracting the image features from the images (ii) The extracted features are then trained using a machine learning classification algorithm. The Feature extraction task is

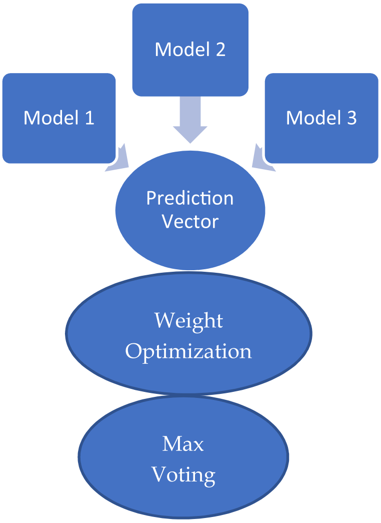

Fig. 2 Proposed weighted max voting model.

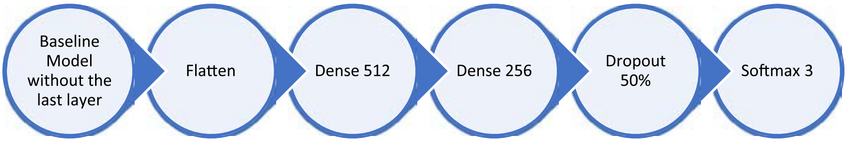

Fig. 3 Layer modification of baseline model.

performed using the various baseline models for comparison. For the process of classifying the features, we shall utilize the following three classification:

(i) Support Vector Machine (Cristianini, Shawe-Taylor, et al., 2000),
(ii) Bagging Classifier (Barstugan, Ozkaya, & Ozturk, 2020a) and
(iii) ADABOOST (Rosebrock, 2020) as previous works prove them to be constantly performing well for similar tasks.

3.4 Layer modification

The baseline networks are initialized with the weights from the ImageNet. The convolutional and max-pooling layers are frozen so that we don't modify their weights. The final softmax layer, mapping to 3 output classes, was replaced with 2 dense layers, 50% dropout layer, and softmax layer mapping to the X-ray labels. These layers were introduced to maximize baseline model classification accuracy during the transfer learning process. Once this is done, we would start retraining. In this way, we manage to take advantage of the feature extraction stage of our network and only tune the new additional layers to work better with our dataset.

Transfer learning by retraining the layers at all is not always a good idea. If the destination task is based on a small dataset that is very similar to the one the network was trained on, leaving the weights frozen and putting a classifier on top of the output probabilities is likely to be more useful, yielding largely similar results without risking overfitting. The architecture of layer modification for the baseline model is depicted in Fig. 3.

4 Experimental setup

4.1 Baseline models

In this section, we explain in brief about the selected pre-trained models which we will use as baseline models for our experiments.

4.1.1 VGG-16 (Simonyan & Zisserman, 2015)

VGG16 is a convolution neural net (CNN) network that was utilized to win ImageNet competition in 2014. Most remarkable thing about VGG16 is that as opposed to having countless hyper-parameter they concentrated on having convolution layers of 3 × 3 channel with a step 1 and consistently utilized same padding space and maxpool layer of 2 × 2 channel of stride 2. At last it has 2 fully associated layers with a softmax for final output. The 16 in VGG16 refers to it has 16 layers that have the weights. This system is a truly huge system and it has around 138 million parameters. VGG-16, although based off of AlexNet (Krizhevsky, Sutskever, & Hinton, 2017), it has the following key differences:

(a) It has replaced the large receptive fields of AlexNet's (11 × 11 with a stride of 4), with very small receptive fields (3 × 3 with a stride of 1). This introduces three ReLU units instead of just one, making the decision function to be more discriminative. Further this reduces the parameters (27 times the number of channels) instead of AlexNet's (49 times the number of channels).

(b) VGG-16 incorporates 1 × 1 convolutional layers to make the decision function more non-linear without changing the receptive fields.

(c) The small-size convolution filters allows VGG-16 to have a large number of weight layers; of course, more layers leads to improved performance.

4.1.2 ResNet 50 (He et al., 2016)

ResNet, short for Residual Networks is a classic neural network used as a backbone for many computer vision tasks. This model was the winner of ImageNet challenge in 2015. The key breakthrough with ResNet was it allowed training extremely deep neural networks with 150 + layers successfully. Prior to ResNet training very deep neural networks was difficult due to the problem of vanishing gradients. There are numerous variations of ResNet, for example same idea yet with a different number of layers. We have ResNet-50, ResNet-101, ResNet-110, ResNet-152 and so forth. The name ResNet followed by a two or more digit number basically suggests the ResNet design with a specific number of neural layers. ResNet-50 is one of the most compact and vibrant networks. The architecture of ResNet50 has 4 stages. The network can take the input image having height, width as multiples of 32 and 3 as channel width. Every ResNet architecture performs the initial convolution and max-pooling using 7 × 7 and 3 × 3 kernel sizes, respectively. Afterward, Stage 1 of the network starts and it has 3 Residual blocks containing 3 layers each. The size of kernels used to perform the convolution operation in all 3 layers of the block of stage 1 are 64, 64 and 128, respectively. The convolution operation in the Residual Block is performed with stride 2. Hence, the size of input will be reduced to half in terms of height and width but the channel width will be doubled. As we progress from one stage to another, the channel width is doubled, and the input size is reduced to half. For deeper networks like ResNet50, ResNet152, etc., bottleneck design is used. For each residual function F, 3 layers are stacked one over the other. The three layers are 1 × 1, 3 × 3, 1 × 1 convolutions. The 1 × 1 convolution layers are responsible for reducing and then restoring the dimensions. The 3 × 3 layer is left as a bottleneck with smaller input/output dimensions. Finally, the network has an Average Pooling layer followed by a fully connected layer having 1000 neurons.

4.1.3 Inception V3 (Szegedy et al., 2015)

Inception V1 was the winner of the ImageNet Competition 2014. It created the record lowest error rate at ImageNet dataset. The model is continuously improved so as to enhance the accuracy and decrease the complexity of the model. Inception V3 network stacks 11 inception modules where each module consists of pooling layers and convolutional filters with rectified linear units as activation function. The input of the model is two-dimensional images of 16 horizontal sections of the brain placed on 4 3 4 grids as produced by the preprocessing step. Three fully connected layers of size 1024, 512, and 3 are added to the final concatenation layer. A dropout with rate of 0.6 is applied before the fully connected layers as means of regularization. The model is pre-trained on ImageNet dataset and further fine-tuned with a batch size of 8 and learning rate of 0.0001. Inception V3 has the following changes compared to its previous models:

(a) Uses RMSProp optimizer instead of SGD l.
(b) Added Batch Normalization to the dense layer of the Auxiliary classifier.
(c) Uses of 7 × 7 factorized Convolution
(d) Label Smoothing Regularization: Regularizes the classifier by calculating the influence of label dropout during training. It penalizes and prevents the classifier from predicting very high probabilities for any single class. This improved the error rate by 0.2%.

We shortlisted these three architectures as our baseline as they have consistently shown good performance in regular image classification tasks and medical image classification tasks (Choi, 2015; Margeta, Criminisi, Lozoya, Lee, & Ayache, 2016; Tajbakhsh et al., 2016). Table 1 highlights the connection type, parameters and total floating-point operations in the three baseline models.

Table 1 Key points of the baseline models.

Model	Connection type	Parameters	Floating point operations
VGG-16	Fixed-kernel	138 M	19.6 B
ResNet-50	Shortcut	23 M	11 B
Inception V3	Wider-parallel	24 M	2 B

4.2 Dataset

The dataset used comprises of labeled chest X-ray images of (i) COVID-19 infected (ii) Pneumonia infected and (iii) healthy people obtained from the following public sources;

(i) Kaggle Pneumonia dataset (1583 normal X-ray + 4273 pneumonia X-ray) (Shih et al., 2019).

(ii) Kaggle Covid-chest Dataset (150 COVID-19) (Kaggle, n.d.).

(iii) GitHub UCSD-AI4H/COVID-CT (288 COVID-19 X-ray) (GitHub, 2020).

(iv) SIIM.org (60 COVID-19 X-ray) (SIIM.org, n.d.).

(v) University of Montreal (684 COVID-19 X-ray) (Cohen, Morrison, & Dao, 2020a).

Fig. 4 shows the imbalance in the classes of X-ray images in the dataset. Pneumonia Infected X-rays constitute 61%, Healthy (Non-Pneumonia and Non-COVID-19) X-rays constitute 22% and the rest 17% are COVID-19 X-rays. Many classification algorithms have low predictive accuracy for the infrequent class. Thus we treat this imbalance by making use of data augmentation strategy to partially rectify this skew in data. Figs. 5–7 show random samples from the dataset for the 3 Classes. It can be noticed that for an untrained eye it's nearly impossible to predict and point out the opacities in chest X-ray.

4.3 Data augmentation

It must be noted that X-ray images are usually of high resolution i.e. usually 1024 pixels × 1024 pixels and are single-channel images and not RGB, unlike normal images. The most common data augmentation technique i.e., cropping of the images, will not be performed on X-ray images to ensure abnormalities within the images is not cropped out. Therefore we perform the following augmentation strategies:

(a) **Flipping**: We perform separate horizontal and vertical flips for each image dataset.

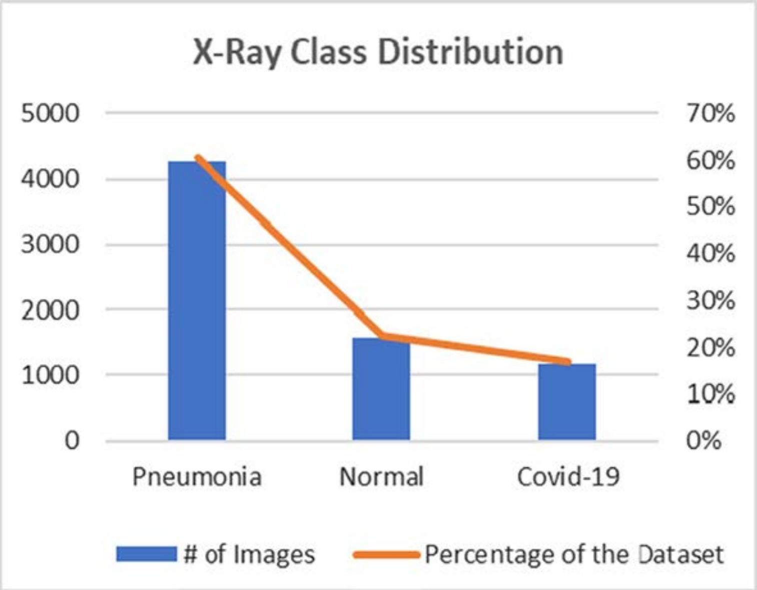

Fig. 4 Distribution of chest X-ray classes.

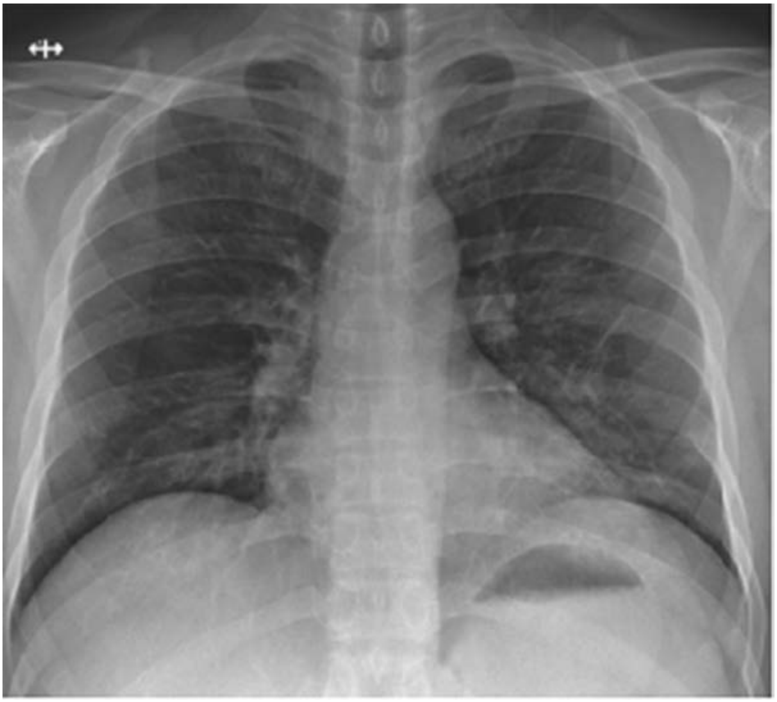

Fig. 5 COVID-19 positive chest X-ray.

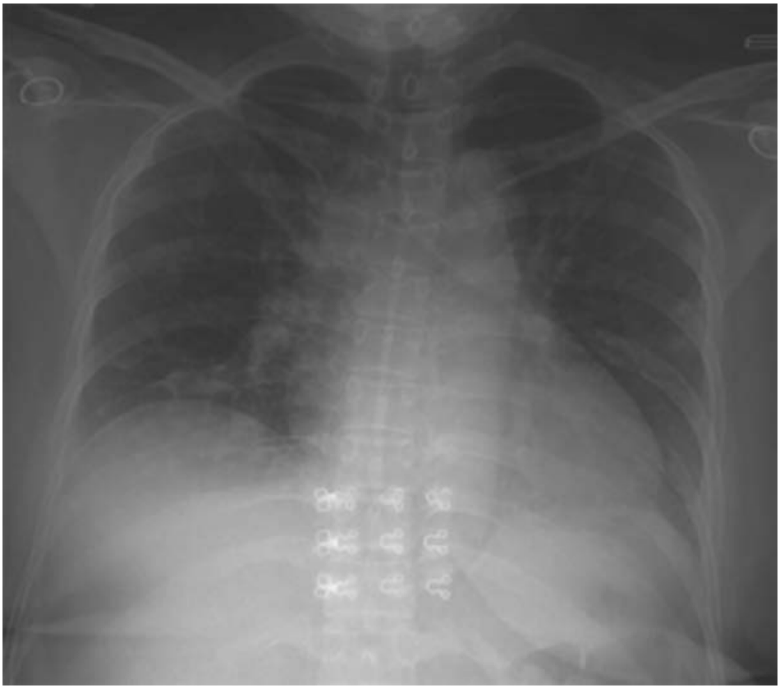

Fig. 6 Pneumonia (positive) chest X-ray.

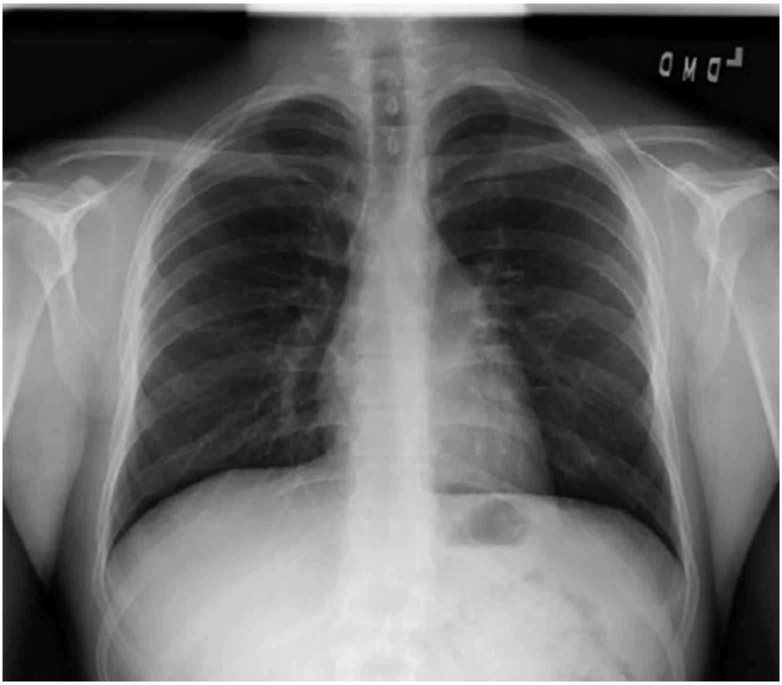

Fig. 7 Healthy chest X-ray.

(b) Rotation: Rotation of images is done using the following transformation,

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

where θ is between 10 and 90 degrees, is applied.

(c) Gaussian Noise: An array, A, is generated where each element in the array is a sample from a Gaussian distribution with $\mu = 0$ and with $\sigma 2$ in the range of [0.1, 0.9]. For each image X in the dataset, we obtain a noisy image, $\mathrm{X}' = \mathrm{X} + \mathrm{A}$.

(d) Jitter: For each image in the dataset, we add a small amount of contrast (±1–5 intensity values).

(e) Power: For each image in the dataset, we take it to power. The power, p, is given by:

$$p = n \times r + 1$$

where n is a number taken from a Gaussian distribution with mean 0 and variance 1 while r is a number <1. Then, the augmented image, X_a, is given by,

$$X_a = \operatorname{sign}(X) \times (|X|^p)$$

The sign and power are each taken elementwise.

(f) Gaussian Blur: A function defined by the variance between 0.1 and 0.9. ($r = 3\sigma$) is applied to blur the images

(g) Shearing: For each image in the dataset, the following transformation is done,

$$A = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}$$

S is the amount that image is to be sheared, and it is in the range of [0.1, 0.35].

4.4 Other preprocessing

The images vary in quality and dimension, ranging from 1215 × 759 pixels to 1024 × 1024 pixels due to multiple sources. To handle this issue we brought all the images to the size of 778 × 778 pixels to obtain a constant dimension for across all the input images.

4.5 Evaluation metrics

The Evaluation metrics are derived from the confusion matrix. Confusion Matrix is performance measure for most classification problems where output can be two or more classes. It is a table with four different combinations of predicted and actual values. (Confusion Matrix of N^2 combination can also be used to note the predictions v/s actual of all N classes). Table 2 shows a standard Confusion Matrix for a 2 Class case.

4.5.1 Accuracy

Classification accuracy is a naïve metric. It is the number of correct predictions made divided by the total number of predictions made. Accuracy in confusion metric terms is given by:

$$\text{Accuracy} = \frac{\text{True positive} + \text{True negative}}{\text{True positive} + \text{True negative} + \text{False positive} + \text{False negative}}$$

4.5.2 Precision

Precision can be thought of as a measure of a classifiers exactness. A low precision indicates a large number of False Positives. Precision in confusion metric terms is given by:

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}}$$

4.5.3 Recall

Recall calculates how many of the Actual Positives our model capture through labeling it as Positive (True Positive). Recall is the model metric we use to select our best model when there is a high cost associated with False Negative. Thus in Covid patient detection, If a Covid patient (Actual Positive) goes through the test and predicted as not sick

Table 2 Confusion matrix.

	Predicted	
Actual	**Negative**	**Positive**
Negative	True negative	False positive
Positive	False negative	True positive

(Predicted Negative). The cost associated with False Negative will be extremely high if the sickness is contagious. The recall in confusion metric terms is given by:

$$\text{Recall} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}}$$

4.5.4 F-1 score

F1 Score is a good measure to use if we need to seek a balance between Precision and Recall and since there is an uneven class distribution of the COVID samples.

F1-Score in confusion metric terms is given by:

$$\text{F1} - \text{Score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

4.6 Experimental details

The primary goal of our experiment is to utilize the power of altered transfer learning approaches to correctly diagnose COVID infection against Pneumonia infection and Normal-No infection using chest X-ray images. As discussed in Section 3, we have prepared 17 different models and studied them separately. For training, we used RMSProp optimizer and the cross-entropy loss function. The learning rate is started from the value of 0.001 and is reduced by 1 after every 5 epochs. The early stopping function takes care of the epoch number. The total images after augmentation processes and duplication removal was 211142 and 10% of this was held for testing.

5 Results and discussion

The pre-trained models are taken in their bare form as suggested by their respective papers for image classification without any alterations to get a benchmark. We have conducted the experiments using the methodology discussed in Section 3. Additional details to the methodology are as below:

(i) Hybrid 1: The feature ensemble model, the features are extracted individually from VGG-16, ResNet-50, and Inception V3 and combined to form a 4048 features. The new feature vector is reduced with PCA for 90% variance explained and passed through a dense layer and softmax.

(ii) Hybrid 2: The probability predictions of VGG-16, ResNet-50, and Inception V3 is passed through a weighted voting system to determine the final predictions. The weights are determined using a solver to ensure the three weights predict produce the best accuracy on the validation set.

(iii) Hybrid 3: Modified Architecture of the three Models are trained individually. This architecture allows us to take advantage of the feature extraction stage of our network and only tune the new additional layers to work better with our dataset.

(iv) Hybrid 4: The features extracted from VGG-16, ResNet-50, Inception V3 are passed separately through the three machine learning classifiers. This results in 3×3 combinations. This helps to identify which models produce the most distinguishable feature representation

The Accuracy, Precision, Recall and F1 score for all the hybrid models are reported in Table 3. It can be seen that VGG-16 although was the simplest of the 3 baseline models still outperforms the other 2 considerably for the Chest X-ray dataset. It achieves an F1-Score of 94.14. Fig. 8 shows the comparative results of the 3 baseline models.

In Hybrid 1: 1000 VGG-16 features, 2048 ResNet-50 features and 1000 Inception V3 features are individually extracted. These 4048 features are fed to a PCA to perform feature selection and develop a union feature set from them. The new feature set comprises of 1257 features explaining 92.64% of the total variance. The predictions after this enhanced feature set is passed through the dense and softmax layer produces a F1 score of 95.74 which outperforms all the 3 individual baseline models. This is mainly due to the fact that combined features has more representational capacity than the features from any single model. Additionally it can be noted the new feature set is smaller than features extracted from Resnet-50.

In Hybrid 2: The final voting weights were 0.43, 0.18, and 0.39 to attain the best F1 score. The Higher weight for VGG-16 can be explained from its performance in the baseline study. The combined prediction power of multiple models clearly outperform the baseline models as they achieve a F1 score of 96.19. This can be expected from an ensemble model as it helps utilize the power of individual model features. Fig. 9 shows the results of Hybrid 1 and Hybrid 2. It can be seen both have better performance than the Baseline models. This highlights the advantage and the power of model ensembles.

In Hybrid 3: The modified architecture has significantly improved the individual score of the models by an average of 9.8% as seen in Fig. 10. This is because of the extended architecture could take advantage of the feature

Table 3 Summary of results.

Type	Models	Accuracy (%)	Precision (%)	Recall (%)	F1-Score
Baseline	VGG-16	90.19	94.12	94.16	94.14
	ResNet-50	87.28	92.12	91.18	91.65
	Inception V3	82.22	85.12	85.24	85.18
Hybrid 1	Feature Ensembles	91.76	95.88	95.60	95.74
Hybrid 2	Majority Voting	92.19	96.33	96.06	96.19
Hybrid 3	VGG-16 Modified	**99.52**	**99.77**	**97.93**	**98.84**
	ResNet-50 Modified	97.75	95.80	94.83	95.31
	Inception V3 Modified	92.09	95.33	95.47	95.40
Hybrid 4	VGG-16 SVM	91.19	94.12	93.15	93.63
	VGG-16 Bagging	90.19	92.22	92.16	92.19
	VGG-16 AdaBoost	90.19	89.16	90.10	89.63
	ResNet-50 SVM	96.11	89.12	89.12	89.12
	ResNet-50 Bagging	95.12	95.12	95.07	95.09
	ResNet-50 AdaBoost	90.18	88.12	88.12	88.12
	Inception V3 SVM	84.29	85.12	85.20	85.16
	Inception V3 Bagging	**99.36**	**99.36**	**99.12**	**99.24**
	Inception V3 AdaBoost	80.12	85.00	86.12	85.56

Bold stands for best model.

extraction stage of our network and only tune the new additional layers to work better with our dataset. Modified VGG-16 achieved an accuracy of 99.52% It also is observed that the F-1 score of the Inception V3 model beats the ResNet-50 despite the accuracies of ResNet-50 is higher. Fig.10 depict the performance of Hybrid 3.

In Hybrid 4: The Bagging classifier performs best across all three models. The Inception V3-Bagging variant performs outstandingly with 99.36% accuracy (135/21114 misclassified). Fig.11 Compares the performance of various feature extractor and classifier combinations.

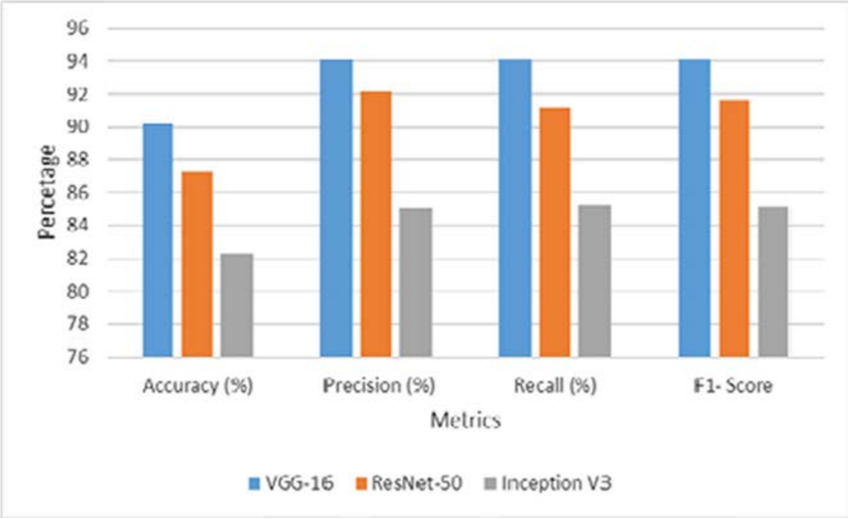

Fig. 8 Results of baseline models.

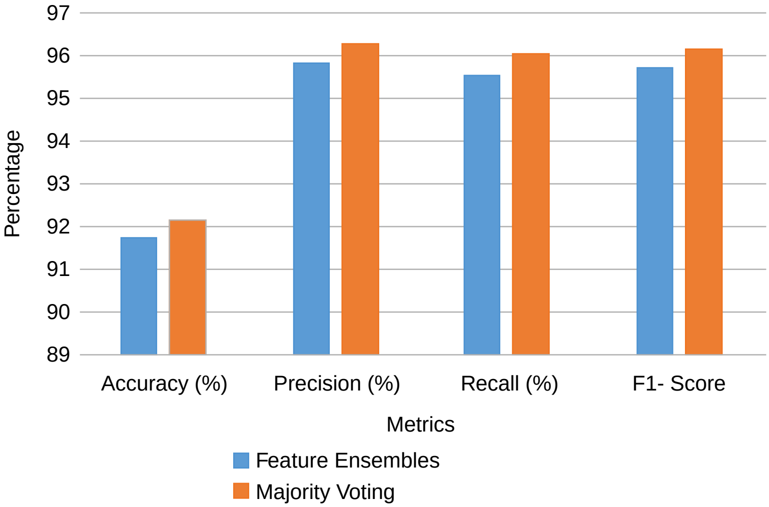

Fig. 9 Results of Hybrid 1(feature ensemble) and Hybrid 2 (max voting).

Overall best performer is the Hybrid 3-VGG-16 with modified layers with an accuracy of 99.52% and F1-score of 98.84. The confusion matrix of the same is shown in Table 4. It can be seen that out of 21,114 images, only 101 were misclassified. The achieved accuracy to 99.52% is far higher than any testing kit available in the market. Another breakthrough is the fact

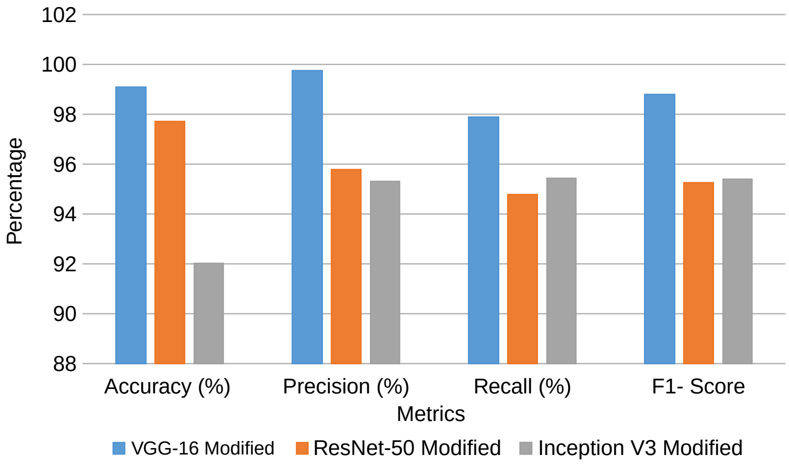

Fig. 10 Hybrid 3 model.

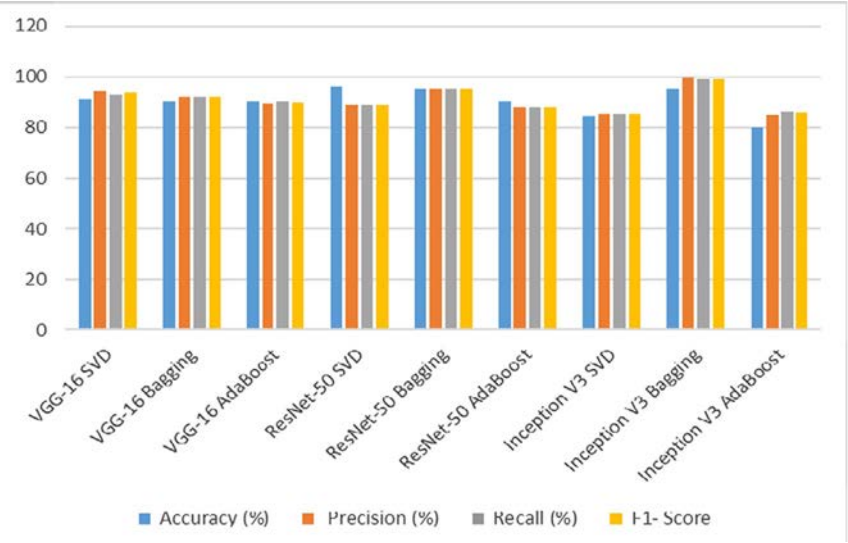

Fig. 11 Hybrid 4 model.

Table 4 Confusion matrix modified VGG-16.

	True		
Predicted	**COVID-19**	**Pneumonia**	**Normal**
COVID-19	3523	25	0
Pneumonia	76	12,765	0
Normal	0	0	4725

that there were ZERO cases where the infection was recorded as normal (False Negative Normal).

6 Conclusions

In this chapter, we propose a quick diagnostic tool using ensemble/hybrid approaches to classify COVID-19 and pneumonia from chest X-ray images using pre-trained models. We explored 4 possible hybrid methods incorporating pre-trained architectures like Inception V3, VGG-16, and ResNet18 trained on the ImageNet dataset. We used the 3 architectures for Feature extraction and ensemble prediction. We found that the modified VGG-16 and the Inception v3 + Bagging achieved accuracies of 99.52% and 99.36%, respectively. For future study, we propose increasing the dataset size and using hand-crafted features. Our findings support the notion that deep learning—AI approaches can be used to improve and ease the diagnostic process and improve disease management.

References

Barstugan, M., Ozkaya, U., & Ozturk, S. (2020a). *Coronavirus (COVID-19) classification using CT images by machine learning methods*. arXiv preprint arXiv:2003.09424.

Barstugan, M., Ozkaya, U., & Ozturk, S. (2020b). *Coronavirus (COVID-19) classification using CT images by machine learning methods*. ArXiv https://arXiv:2003.09424.

Choi, S. (2015). X-ray image body part clustering using deep convolutional neural network: SNUMedinfo at imageCLEF 2015 medical clustering task. In *Proc. workshop CLEF 2015 working notes*.

Chouhan, V., Singh, S. K., Khamparia, A., Gupta, D., Tiwari, P., Moreira, C., et al. (2020). A novel transfer learning based approach for pneumonia detection in chest X-ray images. *Applied Sciences*, *10*(2), 559.

Cohen, J. P., Morrison, P., & Dao, L. (2020a). *COVID-19 image data collection*. [Online] Available: https://arxiv.org/abs/2003.11597.

Cohen, J. P., Morrison, P., & Dao, L. (2020b). *Covid-19 image data collection*. arXiv 2003.11597 https://github.com/ieee8023/covid-chestxray-dataset.

Cristianini, N., Shawe-Taylor, J., et al. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press.

GitHub. (2020). https://github.com/ieee8023/covid-chestxray-dataset. from April 16, 2020-May 5, 2020.

He, K., et al. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Kaggle. (2020). *Kaggle's chest X-ray images (pneumonia) dataset*. https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia.

Kaggle. Available: https://www.kaggle.com/andrewmvd/convid19-xrays.

Kassani, S. H., Kassasni, P. H., Wesolowski, M. J., Schneider, K. A., & Deters, R. (2020). *Automatic detection of coronavirus disease (covid-19) in X-ray and CT images: A machine learning-based approach*. arXiv preprint arXiv: 2004.10641.

Khan, A. I., Shah, J. L., & Bhat, M. M. (2020). CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest X-ray images. *Computer Methods and Programs in Biomedicine*, *196*, 105581. Crossref. Web.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90.

Margeta, J., Criminisi, A., Lozoya, R. C., Lee, D., & Ayache, N. (2016). Finetuned convolutional neural nets for cardiac MRI acquisition plane recognition. *Computer Methods in Biomechanics and Biomedical Engineering Imaging and Visualization*. https://doi.org/10.1080/21681163.2015.1061448.

Rajaraman, S., Siegelman, J., Alderson, P. O., Folio, L. S., Folio, L. R., & Antani, S. K. (2020). *Iteratively pruned deep learning ensembles for covid-19 detection in chest X-rays*. arXiv preprint arXiv:2004.08379.

Rosebrock, A. (2020). *Detecting COVID-19 in X-ray images with Keras, tensor flow, and deep learning*. https://www.pyimagesearch.com/2020/03/16/detecting-covid-19-in-x-ray-images-with-keras-tensorflow-and-deep-learning/.

Rousan, L. A., Elobeid, E., Karrar, M., et al. (2020). Chest x-ray findings and temporal lung changes in patients with COVID-19 pneumonia. *BMC Pulmonary Medicine*, *20*, 245. https://doi.org/10.1186/s12890-020-01286-5.

Shih, G., et al. (2019). Augmenting the National Institutes of Health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology. Artificial Intelligence*, *1*(1), 1–5.

SIIM.org. https://SIIM.org/covid-chestxray-dataset.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International conference on learning representations*.

Szegedy, C., et al. (Jun. 2015). Going deeper with convolutions. In *Proc. IEEE conf. comput. vis. pattern recognit* (pp. 1–9).

Tajbakhsh, N., et al. (2016). Convolutional neural networks for medical image analysis: Full training or fine-tuning? *IEEE Transactions on Medical Imaging*, *35*(5), 1299–1312.

Wang, L., & Wong, A. (2020). *COVID-net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images*. arXiv preprint arXiv:2003.09871.

Wang, S., Zha, Y., Li, W., et al. (2020). A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis. *European Respiratory Journal*, *56*(2), 2000775.

Yee, S. L. K., & Raymond, W. J. K. (2020). Pneumonia diagnosis using chest X-ray images and machine learning. In *Proceedings of the 2020 10th international conference on biomedical engineering and technology (ICBET 2020)* (pp. 101–105). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3397391.3397412.