CT manifestation of COVID-19

Introduction

Background: Chest computerized tomography (CT) has been used as a preliminary diagnostic tool for COVID-19 in a few countries where availability of reverse transcription polymerase chain reaction (RT-PCR) tests was limited. However, current CT-based diagnosis is largely based on review of the imaging result by a trained healthcare professional which may be inefficient or infeasible for triaging suspected cases during an outbreak. Automated diagnosis by a machine learning algorithm could be useful in these settings. One recent publication utilized a deep learning network (COVID-Net) performed under an enhanced processor to distinguish COVID-19 and community acquired pneumonia patients using their CT scan data. Compared to physician review, which has a sensitivity of 97% and specificity of 75%, this algorithm achieved higher specificity (96%) but lower sensitivity (90%) against the gold standard RT-PCR test [Li et al, 2020].

Aim: We aim to evaluate whether machine learning can improve prediction accuracy in classifying COVID-19 cases vs non-COVID-19 cases using lung CT images. We are especially interested in whether it's possible to increase sensitivity above 90% given the same specificity of 96%, with the understanding that our model could underperform COVID-Net.

Methods

Data source: We utilized the CT-scan dataset assembled by researchers at University of California, San Diego (UCSD) [https://github.com/UCSD-AI4H/COVID-CT]. This is an open-source dataset with CT images extracted from peer-reviewed or pre-print publications, with confirmatory RT-PCR test for COVID-19. The raw data contained 349 CT scans from 214 patients with confirmed COVID-19 and 397 CT scans from 91 patients with non-COVID-19 diagnosis. Some patients had multiple scans at different stages of the disease. Age and gender information are available for only a limited number of patients with confirmed COVID-19.

Data preprocessing: The first dataset included all CT images from the UCSD collection without augmentation, including multiple scans per patient. The advantage of using all scans is that we capture more variation in images and disease progression as well as it gives us a much bigger sample size than if we only include one picture for each patient. However, the disadvantage of this is that some hospitals tend to take more scans for the same patient and, by including all images, we are likely to introduce batch effects. We also considered the possibility of exploiting the multiple images taken for the same patient across time, but unfortunately no image contains specific timeline information.

To further enlarge the training sample size and increase variation in CT images, we augmented each image into three believable-looking images with the following stochastic distortion to the

original image: (i) rotate the image 5 degrees with probability 80%, (ii) flip from left to right with probability 50%, (iii) randomly zoom by 0.8 with probability 50% and (iv) flip from top to bottom with probability 50% [Zhao et al, 2020][Augmentor package]. By doing so we were able to increase training sample size from 608 images to 1824 images. This second dataset was generated with the intention to avoid overfitting since we have a relatively small dataset. However, the input images are still heavily correlated, as we can only remix existing information. Thus, this might still be inadequate to get rid of overfitting completely. Note that this manipulation applied only to training images; test images were not augmented.

For both original and augmented images, we generated the corresponding predictor dataset by (i) transforming all images to landscape orientation, (ii) resizing images to the largest image possible to avoid information loss, (iii) cropping the image to 200*400 out of computational necessity, (iv) transforming the images from RGB scale to grayscale, (v) generating gradient field and (vi) extracting feature vectors through histogram of oriented gradients (HOG) transformation with 12 partitions for height and width and 9 intervals for angles. All data were standardized before being analyzed.

Batch effect evaluation: Batch effects might exist due to the fact that (1) sample patients might concentrate in several facilities and (2) multiple images for the same patient were likely taken at the same facility. Such batch effects might hinder classification performance if not included as a feature. To evaluate the presence of a batch effect, 2-group K means was performed to cluster pre-augmented processed images; we assumed there would be a strong batch effect if, among patients who had more than 1 CT scan images, there was a high probability that all of their images were classified into the same group. Two clusters were used because, if the same patient's images were to be classified into 2 clusters, his/her images were more likely to be classified into at least 2 clusters when 3 or more clusters were specified.

Classification: For the dataset without augmentation, we randomly divided the processed data into training and test sets based on a 80-20 split. All models were fit and optimized using the training set generated from unaugmented and augmented images and then their performance was compared via the test set. We considered the following classification algorithms:

- Naive Bayes
- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- Logistic Regression (with L1 penalty)
- Support Vector Machine (SVM)
- Random Forest
- Gradient Boosted Classification Trees
- Ensemble (voting classifier)

We used 10-fold cross-validation to find optimal tuning parameters for each model. The final parameter values are recorded in Table 2. For models with high-dimensional parameter spaces, we conducted a cross-validated randomized grid search prior to a more targeted search to help narrow down the feasible parameter space for searches.

We also considered whether models fit better after dimensional reduction. We use principal components analysis (PCA) to identify the orthogonal components explaining maximum variation and then selected components such that at least 85% of the total variation was explained.

We evaluated the performance of final models in the test set using ROC curves. To determine the final model and prediction cutoffs, we also considered the trade-off between test sensitivity and specificity. Since we intend to use machine learning-based classification of CT images as a screening tool for COVID-19, we favor models that yield the highest specificity in the region with near-perfect sensitivity.

We also attempted to understand what were the most important (HOG-transformed) features by (1) visualizing the features and (2) manually reviewing whether there were distinct features that contributed to the misclassified labels compared with the correctly classified images.

Prediction accuracy expectation: Li et al were able to collect CT scans from 3,322 eligible patients, and all CT examinations were performed under a standard imaging protocol, which largely eliminates batch effects. By using a convolutional neural network with enhanced processing units, Li et al were able to learn local patterns from more than 1K slices of each CT scan, and were also able to extract both 2D and 3D features from the images. Before being fed into the COVNet, the neural network used by Li et al, all 3D CT images were preprocessed in the following way: (1) lung region was extracted using U-net based segmentation method (2) all images were down-sampled 5 times in the z-direction (3) all images were resized to # of slices * 224*224 (4) a maximum intensity projection (MIP) was applied to all images to intensity the lesion (5) then each training image was rotated randomly and flipped horizontally or vertically with probability 0.5 (6) finally, random Gaussian noise with a maximum variance of 0.015 was added to the image. In addition, COVNet was also pre-trained for classification of SARS vs non-SARS cases prior to adapting it to classification of COVID-19 vs non COVID-19. Thus, given the limited number of publicly available images sourced, it's unrealistic to think that we would achieve the same accuracy level of 92% as in Li et al as our less sophisticated algorithm is more likely to overfit in such a small sample of data. Thus, we expect our models to underperform previous deep-trained classifier(s).

Results

Among the 214 COVID-19 patients in the data, 73 had data on age and 59 had data on gender. Out of the 73 patients with some kind of age data, 7 were under 20 years old, 23 were between age 20-40, 26 were between 41-64 and 16 were greater than 65 years old. Out of 59 patients with reported gender, 27 were female and 32 were male.

Augmented images for an original image are displayed in Figure 1 to illustrate the variation created in the training data after augmentation.

Clustering analysis output indicated the possibility of a moderate batch effect. Out of 55 patients who had multiple images, and thus whose images were more prone to batch effects, 32

patients' images were classified into 2 clusters at the same time. The complete image cluster assignment result is shown in Table 1.

Figure 2 displays the ROC curves in the test set using models trained on the image data without augmentation. For LDA and QDA only, we present ROC curves based on prediction models that were trained on the first 153 orthogonal PCs. This is because the models based on the original training data had significantly worse performance (data not shown). For all other models, the test ROC curves were based on the models trained on the original training set. Overall, all classification algorithms performed well, yielding area-under-the-curve (AUC) ranging from 0.76 to 0.82. Among all models considered, the Random Forest model had the highest AUC (0.82).

We also assessed the predictive performance of all models that were trained on the augmented image data. The ROC curves in the test set are shown in Figure 3. Similarly, for LDA and QDA only, we present the ROC curves based on the models trained on the first 224 orthogonal PCs. Data augmentation significantly improved the predictive performance of all models. Gradient boosting yielded the highest AUC (0.86) and achieved a 77% test accuracy. The SVM model slightly underperformed gradient boosting on AUC (0.85), but managed to achieve a higher test accuracy of 79% (Figure 4). In addition, the SVM model also outperformed the other models in the tradeoff between test sensitivity and specificity. Specifically, when a sensitivity level of 0.90 or above is considered, the SVM model yields the highest specificity among all models considered (Figure 3). Hence, we select the SVM model with RBF kernel as our final model. Based on the principle of favoring high sensitivity, we suggest the following decision rule for our SVM model: classify as COVID-19 case if the predicted score exceeds -0.607. This yields a test sensitivity and specificity of 94% and 50%, respectively (Table 3). Such decision rule can be recalibrated to reflect users' preference with regard to sensitivity-specificity trade-off, those of clinical experts and practitioners.

We were able to plot the most important HOG-transformed features but we found the visualization difficult to interpret thus did not present the results here (they are however part of the output in our data analysis code). We were not able to identify a method of visualizing the most important features in predicting COVID-19 cases on the original RGB scale within the deadline of this task. Table 4 (1) presents a list of misclassified test images from COVID patients, and table 4 (2) has the same information from test non-COVID patients. Carefully looking at these images, we again found it hard to eye-ball signals that differ between misclassified and correctly classified images, without sufficient medical knowledge.

Discussion

Using publicly available data, we trained a machine learning model to classify images suspected of having pneumonia caused by COVID-19 based on CT lung scans. We compared several supervised learning algorithms, ultimately selecting a model based on a support vector machine with a radial basis function kernel. While we were able to improve model performance by using data augmentation and cross-validation of tuning parameters, our model was not able to achieve the classification accuracy of a previously reported model which used transfer learning and neural networks.

There were several important limitations which may have limited model performance. First, it is possible that the pooling of images from different machines or radiologists, each with their own preferred examination settings and styles, could produce “batch effects” in which the model uses some of these non-disease specific metrics to classify images. Indeed, we found some evidence that this may exist in our data by using k-means to separate images into two clusters. Ultimately, we did not adjust for batch effects because we believed that if our prediction algorithm were put into practice for field work by the general public, we don’t want them to worry about the quality of their CT scans or the particularity of the scans before uploading. In fact, the CT scans obtained by Li et al’s cohort were collected under a strict universal protocol in China and were centrally managed. But such practice would restrict the generalizability of the prediction model to any test set outside of their radiology protocol.

Another limitation is that the number of publicly available images is still quite low. While we were fortunate to be able to use data that others had pulled together from multiple sources, we still only had 746 images from 305 patients (mostly from China and Italy). Therefore, it is possible that our model will not extrapolate well outside these contexts.

Furthermore, our model is probably sensitive to the eccentricities of the case and control images that were included. For instance, our model may detect general pneumonia rather than features that are particularly specific to COVID-19. This could cause issues in areas in which there are many non-COVID pneumonia cases. In addition, some of our images are from later stages of the disease and therefore our model may not be fully prepared to identify the earlier stages, which would possibly be most useful to clinicians for early identification/triage.

It is also possible that our model fit might have been improved through further feature engineering. While the histogram of oriented gradients (HOG) transformation generally extracts informative features from images, it can be sensitive to the input parameters (i.e. block size, orientation, etc). We varied a few of these inline with previous literature but it is possible that through further refinement we could have come up with better feature vectors. Additionally, there are other feature extraction transformations, like local binary patterns (LBP) that may work better in this setting where texture may be important given that CT images are of human tissue [7].

Our result is consistent with the existing literature which documents the superior performance of neural networks in image recognition and classification tasks. Indeed, this fact is evidenced by the dominance of these methods in ImageNet and Kaggle competitions involving image recognition and classification [8, 9]. This is likely because the unstructured nature of images requires substantial pre-processing to extract the relevant features for accurate computer vision. Convolutional neural networks and other “deep learning” algorithms are able to use their many convolutional layers to efficiently identify these features from the data and then use them to perform classification tasks with high accuracy and greater generalizability from case to case [10]. Most importantly, by using convolutional neural networks, correlated pixel regions (i.e local 3*3 or 5*5 pixel regions) are identified and looked at together to avoid overfitting and to capture more relevant features.

Despite the fact that we were unable to improve on the performance of existing methods, we still maintain that a machine learning model to assist with COVID-19 case identification would be a convenient tool to aid frontline healthcare workers during the pandemic. Due to its convenient set-up (such as an online webpage where doctors can upload a patient's CT scan result), a machine-learning based screening tool could be especially useful in low-resource settings. For example, low-income countries with few infectious disease specialists or primary care physicians could use the predicted result as a temporary surrogate diagnosis for the purpose of infectious disease control as well as triage for treatment. Additionally, hospitals or medical care facilities in hard hit areas that are overwhelmed by a surge of patients presenting with COVID-like symptoms but without immediate access to adequate PCR tests could also use our machine learning prediction for triage during high acuity period. Finally, another benefit of a remote tool is that it could minimize interaction between healthcare workers and suspected COVID cases, which is important for keeping frontline workers safe and limiting disease spread.

Reference

[1] Ai T, Yang Z et al. Correlation of Chest CT and RT-PCR Testing in Coronavirus Disease 2019 (COVID-19) in China: A report of 1014 cases. Radiology. 0(0). Feb 26, 2020. https://doi.org/10.1148/radiol.2020200642

[2] Li L, Qin L et al. Artificial Intelligence Distinguishes COVID-19 from Community Acquired Pneumonia on Chest CT. Radiology.0(0). Mar 19, 2020. https://doi.org/10.1148/radiol.2020200905

[3] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly supervised classification and localization of common thorax diseases. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2097–2106, 2017

[4] Joseph Paul Cohen and Paul Morrison and Lan Dao, COVID-19 Image Data Collection arXiv 2020

[5] Zhao J, Y Zhang et al. COVID-CT-Dataset: A CT Scan Dataset about COVID-19 .2020. arXiv preprint: https://arxiv.org/pdf/2003.13865.pdf

[6] Augmentor package. Last accessed May 12th, 2020. https://github.com/mdbloice/Augmentor/blob/master/README.md

[7] Alhindi, T. J., Kalra, S., Ng, K. H., Afrin, A., & Tizhoosh, H. R. (2018, July). Comparing LBP, HOG and deep features for classification of histopathology images. In 2018 international joint conference on neural networks (IJCNN) (pp. 1-7). IEEE.

[8] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 2015

[9] List of Kaggle winners for Image Classification tasks. https://github.com/ShuaiW/kaggle-image

[10] A. A. M. Al-Saffar, H. Tao and M. A. Talab, "Review of deep convolution neural network in image classification," *2017 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*, Jakarta, 2017, pp. 26-31, doi: 10.1109/ICRAMET.2017.8253139.

Figure 1. An example of original image and augmented images

	Original image	Augmented images X 3
Patient 190		  

Table 1. Selected clustering results for patients who have multiple images (with 2 clusters)

Patient ID	# of images assigned cluster 1	# of images assigned cluster 1	Total # of images
1	3	1	4
2	11	5	16
13	5	0	5
18	2	1	3
19	1	2	3
20	3	0	3
21	1	1	2
22	2	0	2
23	5	1	6
62	2	4	6
63	5	0	5
78	2	0	2
81	2	0	2
82	1	2	3
83	1	2	3
93	3	0	3
94	3	0	3
105	1	1	2
106	2	2	4
115	2	0	2
116	1	1	2
117	1	1	2
118	1	1	2
119	1	1	2
140	1	3	4
141	2	0	2
142	2	0	2
143	3	1	4
144	4	2	6
145	5	1	6
150	4	4	8
152	2	0	2
153	2	1	3
170	3	0	3
171	2	1	3
176	8	0	8
177	4	2	6
178	4	0	4
179	2	0	2
182	5	3	8
183	2	0	2
184	1	1	2
185	2	0	2
186	2	0	2
187	1	1	2
188	2	0	2
189	1	1	2
197	4	1	5
199	2	0	2
200	2	0	2
215	2	0	2
216	2	2	4

Table 2. Model parameter tuning results

Model	Parameter	Optimal Value	Description
Naive Bayes	-	-	-
Linear Discriminant Analysis	-	-	-
Quadratic Discriminant Analysis	-	-	-
Penalized Logistic Regression	C	0.042	Inverse of regularization strength
Support Vector Machine	C	0.901	Regularization parameter. The strength of the regularization is inversely proportional to C.
	kernel	RBF	
	gamma	0.001	Kernel coefficient for ‘rbf’, ‘poly’ and ‘sigmoid’.
Random Forest	max_depth	10	The maximum depth of the tree.
	min_samples_leaf	3	The minimum number of samples required to split an internal node
	min_samples_split	10	The minimum number of samples required to be at a leaf node
	n_estimators	50	The number of trees in the forest.
Gradient Boosted Classification Trees	learning_rate	0.04	Shrinkage of the contribution of each tree
	max_depth	5	Maximum depth of the individual regression estimators
	min_samples_leaf	3	The minimum number of samples required to be at a leaf node
	min_samples_split	8	The minimum number of samples required to split an internal node
	n_estimators	752	The number of boosting stages to perform.

Figure 2. Receiver operating characteristic (ROC) curve comparing performance of image classifiers in the test set using data without augmentation

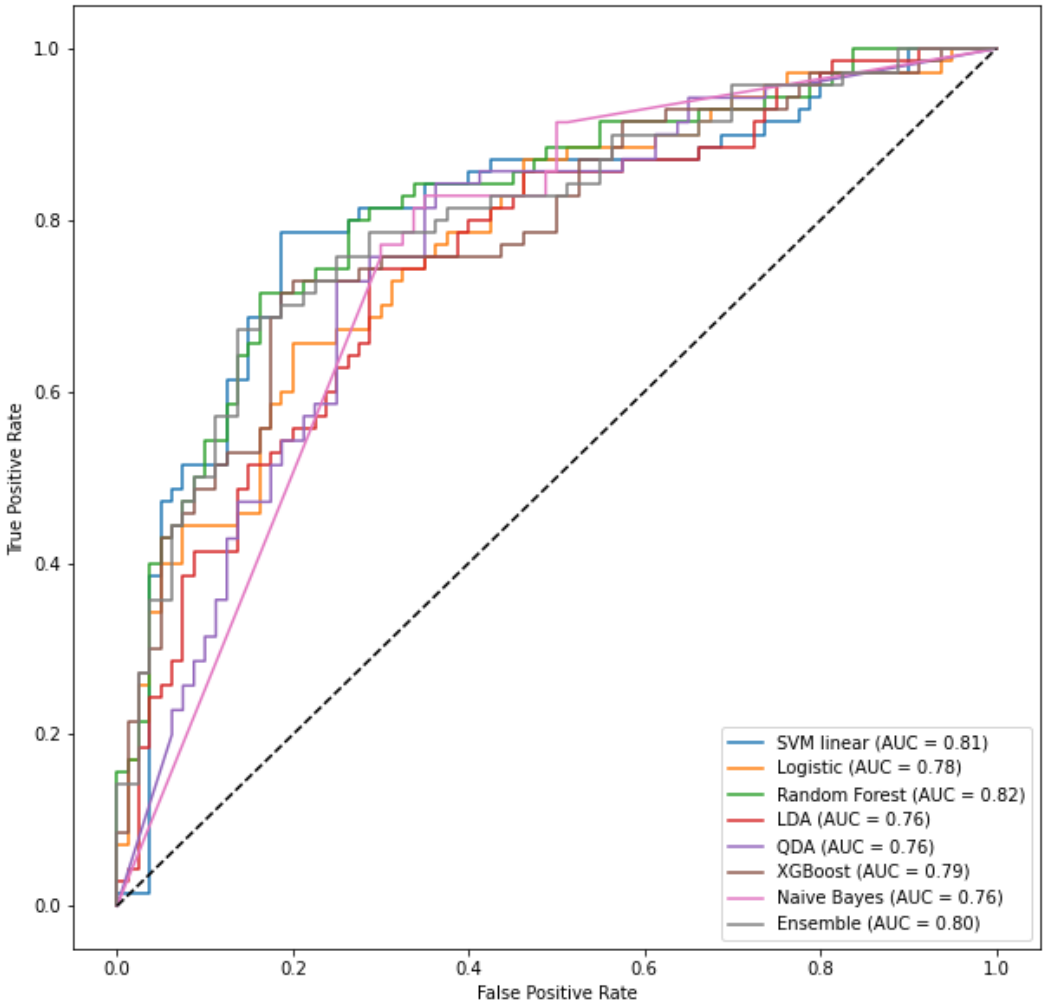


Figure 3. Receiver operating characteristic (ROC) curve comparing performance of image classifiers in the test set using augmented data

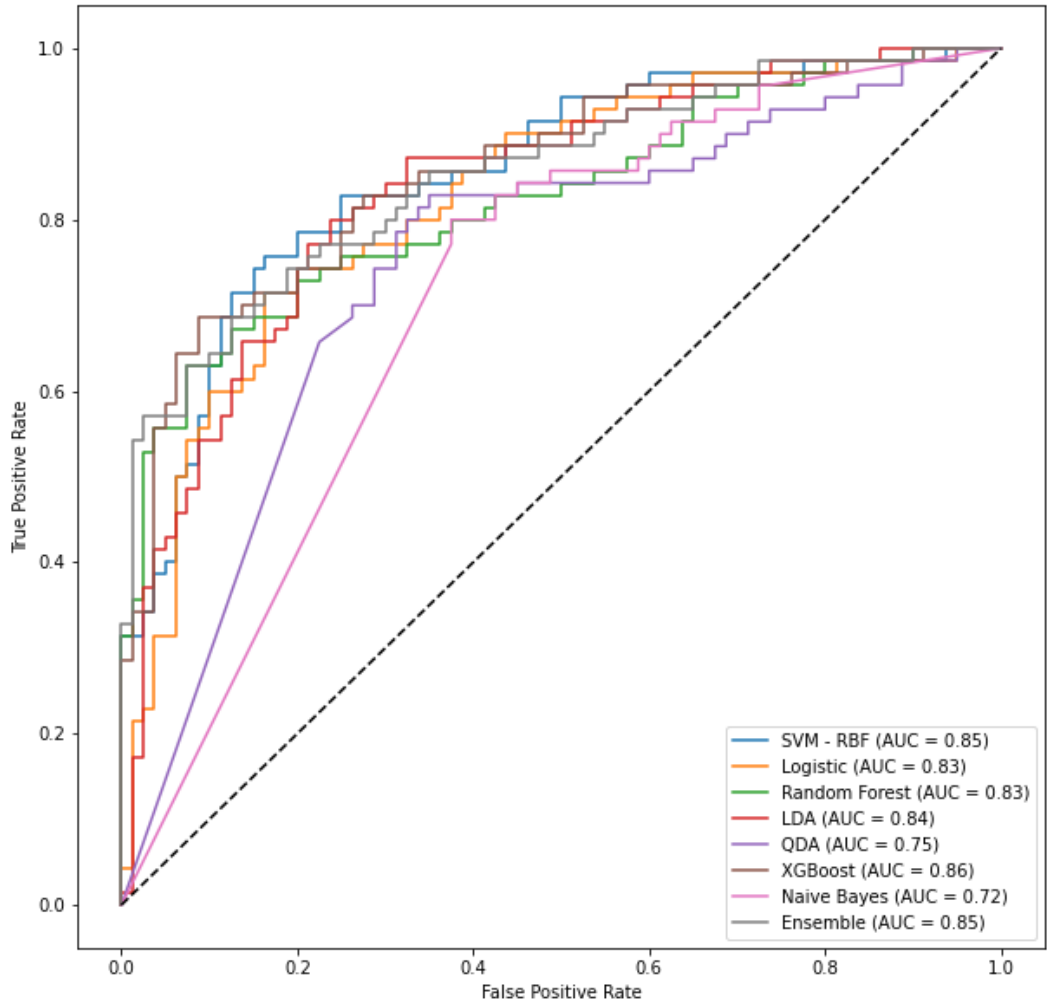


Figure 4. Confusion matrix in the test set for the support vector machine model based on the augmented image data

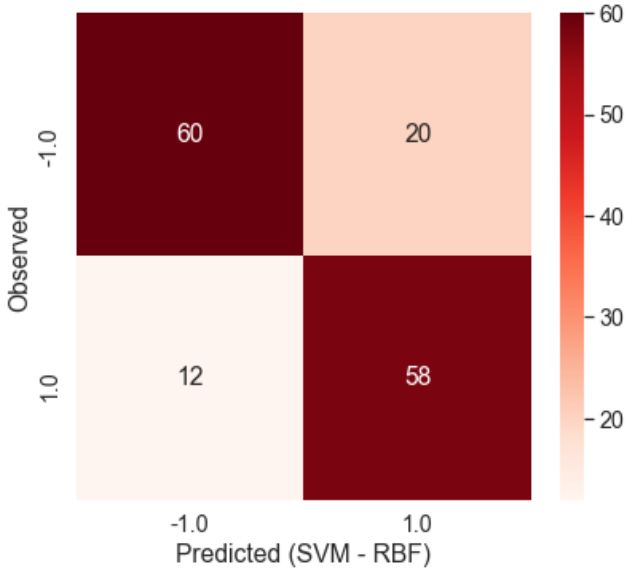


Table 3. Test sensitivity and specificity at various cutoffs based on the support vector machine model with rbf kernel trained on augmented image data

Sensitivity	Specificity	Score
0.914	0.538	-0.559
0.914	0.500	-0.582
0.943	0.500	-0.607
0.943	0.425	-0.682
0.957	0.425	-0.703
0.957	0.400	-0.711
0.971	0.400	-0.717
0.971	0.225	-0.924
0.986	0.225	-0.932
0.986	0.063	-1.144

Table 4 (1) Misclassified COVID-19 patients CT scans from SVM model with RBF kernel

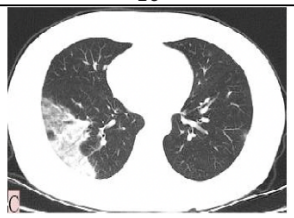
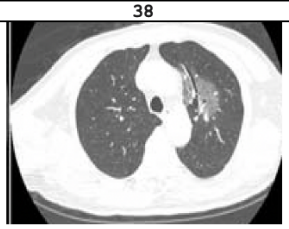
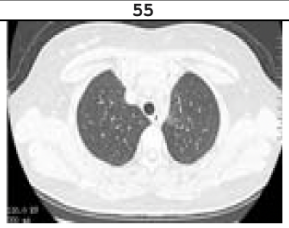
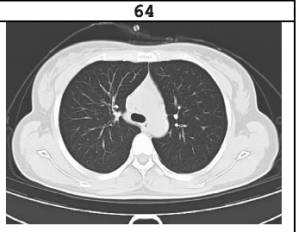
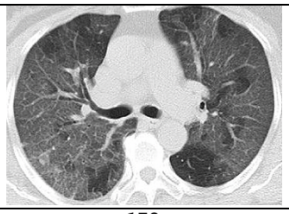
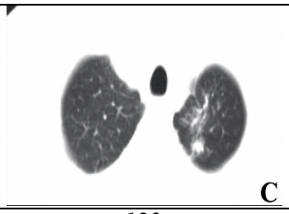
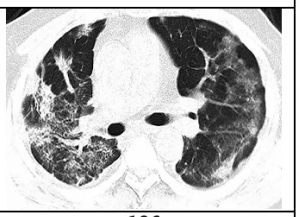
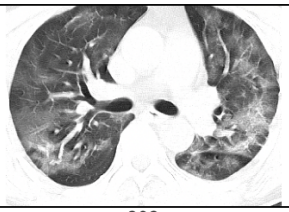
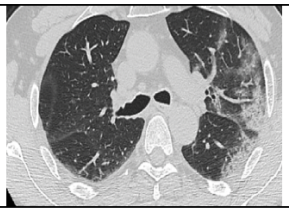
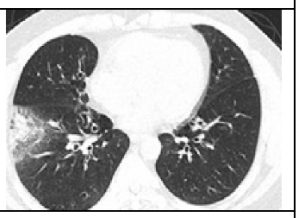
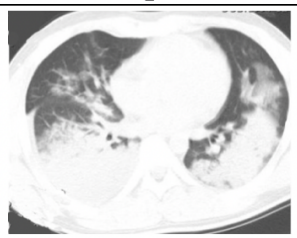
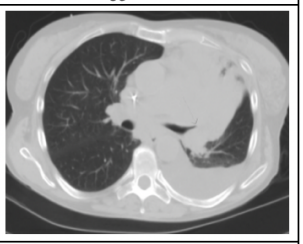
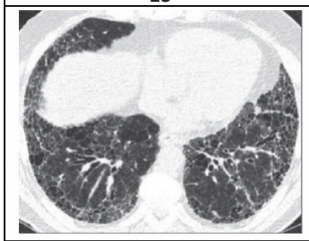
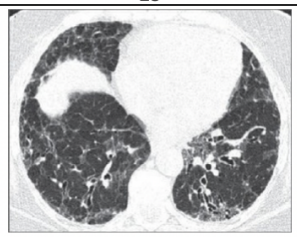
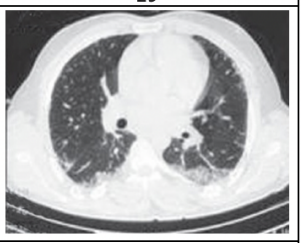
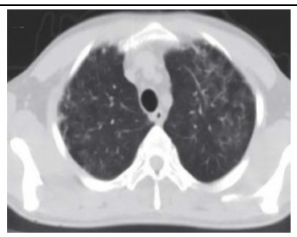
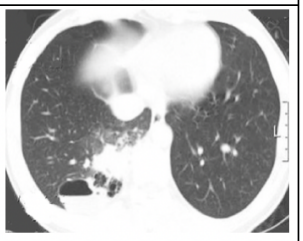
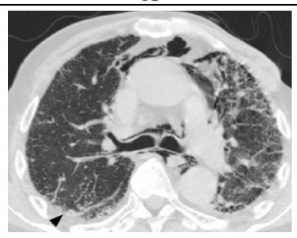
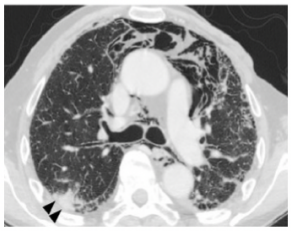
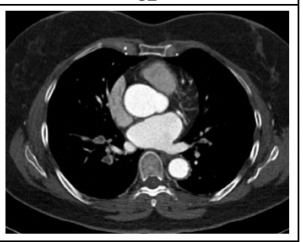
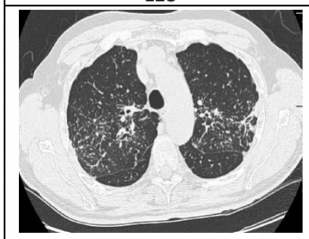
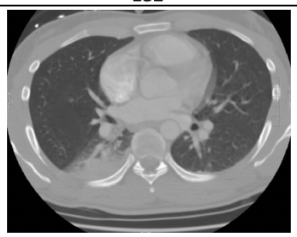
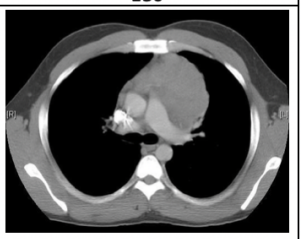
16	38	55	64
			
68	**73**	**140**	**142**
			
147	**178**	**186**	**186**
			
202			
			

Table 4 (2) Misclassified Non COVID-19 patients CT scans from SVM model with RBF kernel

1	2	3	69
			
25	**25**	**69**	**29**
			
38	**39**	**59**	**63**
			

76	83	83	82
			
125	**131**	**132**	**138**
			
139	**146**