3244-2010-0017

Detecting COVID-19 in Radiographic Images Using Deep Learning

Authors: Branson Ng Khin Swee; Fung Brian Pei-En; Ko Gi Hun; Lim Yu Hui; Tay Wei Hong, Allan; Tan Kok Joon
Email: {e0309295,e0310729,e0318604,e0324673,e0415683,e0309136}@u.nus.edu.

Abstract

The main goal of this project was to determine if an individual was infected with the 2019 Novel Coronavirus (COVID-19) by analyzing their chest X-rays (CXRs) with machine learning techniques. We sought first to determine whether or not it was possible to detect COVID-19 by analyzing CXR and, if it was possible, to explore how deep learning can improve upon traditional machine learning techniques in diagnosing COVID-19. These included investigating how deep learning could be used to extract features indicative of COVID-19 from CXRs, and if transfer learning could be utilized to retrain existing image classification networks for our purposes. In our results, the best performing model (transfer learning with DenseNet-121) achieved a high COVID-19 sensitivity score of 98.62%, and a weighted sensitivity score of 96.21%. This is comparable to the results of similar work done for pneumonia detection, and is better in performance than common COVID-19 detection tests like antigen tests.

1. Introduction

COVID-19 is an infectious virus that has killed over a million people globally[1]. With some experts predicting that vaccines could be widely available only by mid-2021[2], there is an urgent need to mitigate the spread of the virus. Infected individuals should be identified and quarantined early in order to contain the virus.

Currently, the testing of COVID-19 is mainly performed using the reverse transmission polymerase chain reaction (RT-PCR) or the antibiotics test methods.[3, 4] These tests, while widely used, are reported to have significant limitations, including their non-negligible false negative rates[8] (FNR) due to the long incubation period of diseases and also possible testing errors from the high flux of testing the limited number of medical staff have to conduct.[5, 6, 7] Coupled with the highly contagious nature of the virus, any failure in diagnosis of a COVID-19 carrier could result in an unwarranted surge in the number of infected.

Medical research into COVID-19 has identified several features within CXRs which could be indicative of COVID-19 infection.[8, 9] Some of these features include:
(1) pleural effusions[8] which is liquid build-up around the lung cavity (fig. 2), and
(2) regions of ground glass opacities[8] (GGO), or more opaque, which might be indicative of liquid build-up within the bronchial tubes (fig. 3).

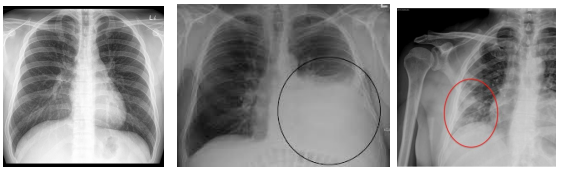


fig. 1-3 (left to right):
normal[a]; pleural effusion[b]; COVID-19 with GGOs[c]

However, some features such as GGOs may be overlooked, and requires trained radiologists to identify. This makes identifying COVID-19 in an accurate and timely fashion difficult, especially when hospitals are overloaded during a pandemic and manpower becomes a problem. Thus, deep learning methods, which have shown stable and outstanding results in detecting complex patterns in medical images, could be of crucial help in the task of detecting COVID-19 quickly, reliably, and precisely.[10 - 15]

In summary, we aim to develop a model which could help medical professionals identify whether a patient has COVID-19 by examining their CXRs as well as to explore how deep learning can improve upon traditional machine learning techniques in diagnosing COVID-19. In particular, we were looking to explore how deep learning can be used to identify features indicative of COVID-19, and how transfer learning can be used to re-train existing image classification networks for this purpose.

2. Background / Related Work

Extensive research has already been made in utilising deep learning in the field of medical imaging for lungs diseases.[16 - 24] While many of the models developed from those research are achieving high accuracy of over 90%, they face the common limitation of having small or unproportional datasets [21, 24, 25], mainly due to the lack of chest CT/X-rays of COVID-19 patients as they are not regarded as primary detection measures for COVID-19. Transfer learning models are thus often adopted for this task and commendable results are observed. Apostolopoulos and Mpesiana[24] experimented with various Convolutional Neural Networks (CNN) with transfer learning and achieved a high accuracy and specificity of 98.75% with VGG19. However, it is worth noting that their dataset was firstly unbalanced, with only a small sample (approx 10%) of COVID-19 CXR. Secondly, it was directly obtained from Cohen's github and used without initial pre-processing, which could potentially have led to duplicated images across training and validation sets, resulting in data leakage. Ruochi et. al[25] proposed using a two-step transfer learning framework known as the COVID19XrayNet, where ResNet34 is firstly trained and tuned using a dataset without COVID-19 cases, before being further transferred and tuned with a separate dataset with COVID-19 cases included. He achieved an overall accuracy of 91.92% using only 189 annotated COVID-19 X-ray images, stating that the model is likely to improve with more available annotated images. Ozturk et al.[21] with his proposed DarkCovidNet model inspired by the Darknet architecture, is able to achieve an accuracy of 98.08% and 87.02% for binary (COVID-19and no-findings) and multi-class (COVID, pneumonia, no-findings) classifications respectively, which he states is bottlenecked by the limited number of COVID-19 CXR images available. Thus, the unbalanced dataset is indeed a problem.

In this study, we experimented with traditional machine learning, deep-learning techniques to improve the precision and sensitivity of the models in detecting and differentiating COVID-19 CXRs, from viral pneumonia (VP) CXRs and normal CXR.
Our main contributions are listed as follows :
(1) Aggregating a balanced dataset with pre-processing to prevent possible data leakage through duplicated images. Experimented with regularization and image augmentation techniques to solve overfitting.
(2) Experimented with fusing metadata with labels by creating subclasses like COVID-Male, results were however not promising as we could not find sufficient data
(3) Attempted to use principal component analysis (PCA) to reduce dimensionality of images. However, upon attempting to recreate the images from the reduced features, there was visible artifacting and some known x-ray features such as GGOs became hard to distinguish.
(4) Using t-stochastic neighbor embeddings (TSNE) on the pre-trained image net extracted feature space, we found that there were some subclusters within each class. One in particular correlated

to CT scans, which we removed the aggregated dataset as our focus was on CXRs.
(5) Using our test subset, we managed to achieve a high accuracy score of 96.35% and weighted sensitivity score of 96.35% with our ensemble model and a COVID-19sensitivity score of 98.62% with Densenet121.

The structure of the remaining paper is presented as follows:
Section 3 Introduces the dataset we created and used for training the models. It also describes the various techniques we applied in the project.
Section 4 and 5 provide the microscopic and macroscopic analysis of our model's performance respectively.
Section 6 concludes the report.

---

3. Methods

In our project, we compare traditional and deep learning methods. Traditional machine learning models used include logistic regression and support vector machines (SVM). For deep learning, we experimented with pre-trained convolutional neural network (CNN) models such as ResNet18, VGG16 and DenseNet121 and ensembles. We've also tried infusing metadata and unsupervised learning to generate label subclasses.

3.1 Deep Learning

Given the medium of images, we were naturally inclined to use the tried and tested method of applying CNNs to our dataset to detect the presence of COVID-19. Given our relatively small dataset size (~3000 images), we have decided to use transfer learning with pre-trained ImageNet weights. We also experimented with various architectures of different complexities to find the best bias-variance fit. (i.e. the model that is complex enough without overfitting). We evaluated the models using our test subset results to avoid any positive bias from the validation subset.

3.2.1 Data

The dataset used initially was from the COVID-19 Radiography Database[26]. Unfortunately, the dataset is unbalanced and samples for COVID-19 were under-represented, only 7% of the dataset. (Appendix A, fig.1)

To rectify this issue, we sampled more COVID-19 images from another dataset[27] collected by the University of Montreal. This helped us to improve the balance of the dataset considerably (Appendix A, fig. 2).

To prevent any repeated data and potential data leakage, we resized all the images and applied an image difference hashing algorithm to calculate the image hashes and remove images with identical hashes. While we could have lost some data, this allowed us to ensure that there were no duplicate images included due to the aggregation of the two datasets..

3.2.1.1 ImageNet Normalization

As we have decided to approach it via transfer learning, our input data has to match the parameters of the ImageNet dataset. Hence, we reduced the image dimensions to 224x224 to fit the original image dimensions used in the ImageNet trained models.

3.2.1.2 Image Augmentation

Given the lack of data, image augmentation was also performed on the training data to increase its variety. We did random rotations of up to 5 degrees, random crops and slight random contrast and brightness adjustments (Appendix C). Having training data randomly augmented helps the model learn the true signal and generalize better to unseen CXR images. We also took caution to set seeds in torch, numpy and random when comparing between architectures trained with image augmentations.

3.2.1.3 Data split

We also took care to split the data and record the split in a csv so that all our models are benchmarked against the same split of data. Each split was also made to be representative of the original dataset (Appendix B).

3.3 Training Techniques

3.3.1 Transfer Learning

In the retraining process, we performed gradual fine-tuning of the weights. We first trained the head, and gradually made the rest of the model trainable from the head to the input layers. This gradual 'unfreezing' of the layers was used to avoid losing representations learnt from ImageNet weights by ensuring that the higher level features were adjusted first. This order was chosen because low level features are unlikely to change and high level features are more likely to be different for a new image classification task.

3.3.2 Regularization - Dropout, Weight Decay, Schedulers

Regularization was also used in the form of dropout layers, weight decay and learning rate schedulers. Given the immense complexity of these models, regularization was helpful in seeking out just that slight increase in performance. The learning rate scheduler employed reduced the learning rate by a factor of 0.1 every 2 epochs. We also used a weight decay parameter of 0.001 and Dropout at layers with probability 0.5.

3.3.3 Hyperparameters

We trained our models with the stochastic gradient descent (SGD) optimizer, categorical cross entropy loss function, 5 epochs and a learning rate of 0.003.

3.3.4 Weighted Sampler

Although our dataset is already quite balanced, we added a weighted sampler for our training data loader only. This was done so that our validation and test results would be consistent across comparisons regardless of random seed values set. Also, to ensure comparability between training runs for different models, we took care to set random seed values in torch so that each model gets exposed to the same weighted sample from the training subset.

3.3.5 Clustering

3.3.5.1 Unsupervised Label Subclasses

Given the various symptoms and stages of a respiratory illness like COVID-19, the appearance of the lungs in CXRs would be quite different depending on the stage of the illness the patient is in.

Given our lack of medical knowledge, we could only do this in an unsupervised manner. Our approach was to generate subclasses for COVID-19 and VP. We did this by using features extracted from either our pre-trained or trained CNNs. We used Principal Component Analysis (PCA) to reduce the dimensionality of the features while retaining 0.95 variance and then employed K-Means clustering.

We first tried setting our K value to the number of distinct stages a person could experience when diagnosed with COVID. This however led to vastly unbalanced datasets which made any training less meaningful, We thus resorted to determining K by measuring the sum of squared distances of each point from their centroids (J_{clust}).

3.3.5.2 Metadata - Age, Gender

Physiologically, lungs of different genders look different. Age also impacts the severity of COVID-19 in patients. Our hypothesis was thus to divide our data according to such metadata and perform classification on subsets of COVID-19 labels based on these metadata.

Gender was easy to separate but for age we used age bands of 0-30, 31-60, 60-100 to bin the age values into. The bins were chosen to maintain a more even class distribution but it was still fairly unbalanced. They were also chosen due to the significant differences in the impact of COVID-19 on these age groups.

Unfortunately, the dataset we had with such metadata[28] contained mainly COVID-19 data. In order to have some form of balance in our dataset, we could only divide the COVID-19 labels so that the VP and normal classes still had enough data. (Appendix E)

3.3.6 Ensembling

We ensembled ResNet34, DenseNet121 and VGG16 together with a final layer that took the concatenation of each of their outputs. We only trained the new heads of each individual CNN and the final layer of the ensemble.

3.3.7 Heat Maps - Class Activation Mappings (CAMs)

We employed Class Activation Mappings (CAM) to determine whether our models have learnt to look at the right parts of the CXR images after retraining.

To generate the CAMs, we extracted the input weights of the final linear layer of our model. By multiplying the weights of the final layer corresponding to the COVID-19 label together with the unpooled feature vectors, we then generated a matrix of numbers that corresponds to the CAM which shows us which parts of the images were most influential in the decision-making process of our model.[29]

3.3.8 Traditional Machine Learning

For our traditional learning approach, we also performed dimensionality reduction and feature extraction on the same set of images used for Deep Learning before they were used to train the models. Feature extracted data were then used to fit Support Vector Machine Classifier, using Scikit-Learn library.

3.3.8.1 Dimensionality Reduction

When working with images, the dimensionality of data can become an issue, which could negatively affect training speeds. As such, we attempted with several approaches such as using PCA, downsampling to reduce dimensionality of the images.

However, one issue we noticed when using PCA was that some features or details appear to be lost. By reconstructing the images from the PCA reduced data, we noticed that some of the reconstructed images included artifacts from other images, the shapes of the lungs became hard to distinguish (Appendix H).

We also performed feature extraction with pre-trained CNN models such as VGG16 so that we could get higher level features from the images for use in classification by our traditional machine learning models.

---

4. Evaluation of Model

4.1 Baseline

A previous study utilized transfer learning on patient CXRs using pre-trained CNN models including ResNet18, ResNet 50, SqueezeNet, and DenseNet-121, for detection of COVID-19 infection as well. However, their dataset, while larger, was skewed with few samples of COVID-19 (100 COVID-19 vs 3000 non-COVID-19)[30]. In order to attempt to improve upon their results, we attempted to improve the balance of the dataset by including more COVID-19 CXRs.

The evaluation metrics they used were sensitivity and specificity and all four models performed reasonably well. The results can be seen in this table, as referenced from their research paper[30] (fig. 4).

Model	Sensitivity	Specificity
ResNet18	98% ± 2.7%	90.7% ± 1.1%
ResNet50	98% ± 2.7%	89.6% ± 1.1%
SqueezeNet	98% ± 2.7%	92.9% ± 0.9%
Densenet-121	98% ± 2.7%	75.1% ± 1.5%

fig. 4: baseline result of similar research[30]

4.2 Results

While accuracy was one of the metrics we used to evaluate the performance of our models, we decided to focus more on other metrics such as precision, sensitivity and F1-score as these were more important when detecting COVID-19 as compared to plain accuracy (true positives + true negatives). Furthermore, as the classes are unbalanced, standard evaluation metrics that treat all classes as equally important might produce misleading conclusions, which is why we decided to use the precision-sensitivity metrics so that we can focus on the performance of the minority class, COVID-19.

Minimising false negatives is important so that infected individuals can be isolated quickly and the virus does not spread unchecked. This is why we focussed on the sensitivity metric.

We also placed emphasis on the precision metric as we do not want to misclassify negative cases as positive which might lead to unnecessary wastage of hospital resources.

Thus, the F1-score is suitable as the primary metric to evaluate our models as it considers both precision and sensitivity.

4.3 Deep Learning

4.3.1 Overview of Results on Test Subset

Architec-ture	Accuracy (test)	COVID19 Sensitivity (test)	Weighted Sensitivity (test)	Weighted F1-score (test)
ResNet34	0.9576	0.9726	0.9576	0.9577
DenseNet-121	0.9605	0.9862	0.9621	0.9605
VGG16	0.9342	0.9489	0.9326	0.9342
Ensemble	0.9635	0.9793	0.9635	0.9635

fig 5: Model performances of re-trained image networks.

Just a quick recap that the ensemble is a combination of ResNet34, Densenet121 and VGG16. While DenseNet121 performed best in COVID-19 Sensitivity, the Ensemble did better in the other 2 metrics. Overall, we still crowned DenseNet121 as the winner as it performed best at the most important metric which is COVID-19 sensitivity.

4.3.2 Heat Maps - Class Activation Mappings (CAMs)

In the set of CAMs below, we can see that the pre-trained model does not focus on the lung areas. The red regions (which indicates a higher interest in the region) in the CAMs are quite random and are not centered on the lungs (fig. 6). This is likely due to the weights being trained for processing of images which differ quite significantly from CXRs.

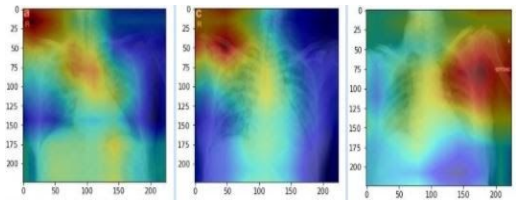


fig. 6: Pre-trained model regions of interest

After we trained the model on the X-ray dataset, we used the same procedure to generate the CAMs on the same images as above and the improvements over the previous set of CAMs are immediately obvious (fig. 7) with the model tending to focus much more on regions around the lungs, with some exceptions.

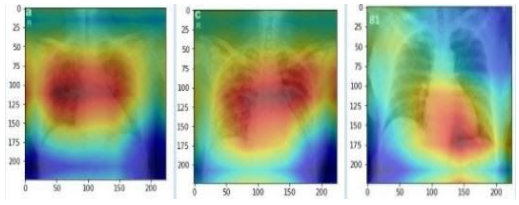


fig. 7: Re-trained model regions of interest showing a general improvement in the region of interest.

4.3.3 Metadata

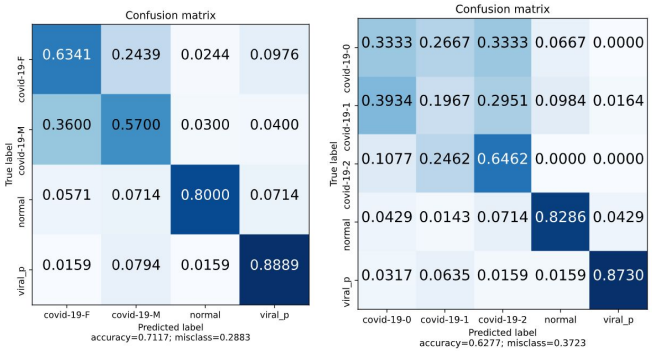


fig. 8: (Left) Gender labelled confusion matrix and,
fig. 9: (Right) Age labelled confusion matrix

For age, we divided our data into 0-30 (COVID-19-0), 31-60 (COVID-19-1), 61-100 (COVID-19-2).

From both confusion matrices (fig. 8), we can see that the splitting of COVID-19labels was not very effective. While we still believe that using metadata could still be useful, our lack of CXR images with metadata has not allowed us to fully explore this hypothesis.

4.3.4 Unsupervised Label Subclasses

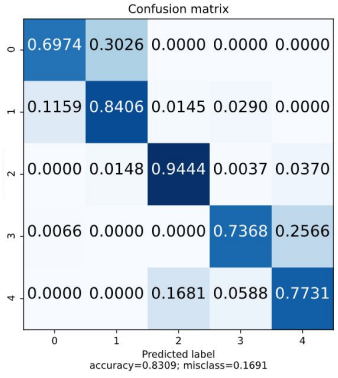


fig. 10: Unsupervised Labels

Here (in fig. 10) {0,1} refers to COVID-0 and COVID-1 labels, 2 refers to normal and {3,4} refers to VP-0 and VP-1 labels. These particular labels were generated by choosing K which had the lowest J_{clust}.

While it seems to have generated fairly distinct classes, without medical knowledge, we were not able to ascertain if the labels had meaningful representations or how we could have improved upon the generation apart from looking for K values and initialization states that had lower J_{clust}.

4.4 Traditional Machine Learning

4.4.1 Support Vector Machines (SVM)

We first attempted to classify the images using linear regression using the raw pixel intensity feature space. However, the data was not linearly separable in this feature space. We then pivoted our project scope to using SVMs instead.

As introduced in 3.2, we have performed 2 different dataset transformations before fitting the SVM classifier; feature extraction using VGG model and flattened pixel intensity values with size 224x224. We then constructed confusion matrices based on the two different datasets prepared.

The confusion matrices (Appendix J) display the normalised confusion matrix of the model using two different datasets. All 3 labels showed good accuracy of higher than 90%. However, we can observe for 'Normal' and VP labels, support vector machines that fit with VGG feature extracted data can classify 6% and 4% higher than naive pixel flattened data respectively.

In addition to our various dataset transformation attempts, we have fitted and evaluated our SVM classifier using 3 kernels, linear, polynomial (n=3) and radial basis function (rbf). The tables (fig. 11-12) below show the evaluation results of SVM using 3 different kernels and raw pixel flattened data.

kernel	Acc	CoVID-19 Sensitivity	Weighted Sensitivity	Weighted F1
linear	0.856	0.9261	0.8985	0.9014
rbf	0.944	0.9482	0.94237	0.9437
polynomial	0.905	0.9489	0.9326	0.9284

fig. 11: Raw Pixel Intensity Data

kernel	Acc	CoVID-19 Sensitivity	Weighted Sensitivity	Weighted F1
linear	0.861	0.9261	0.9063	0.9498
rbf	0.935	0.9766	0.9575	0.9645
polynomial	0.874	0.9815	0.9294	0.9645

fig. 12: PCA Dimensionality Reduction

We were able to observe that the radial basis function (rbf) kernel produces the best results for accuracy, COVID-19 sensitivity and weighted sensitivity metrics on the same feature space.

---

5. Discussion

5.1 Deep Learning

5.1.1 Overfitting

Given the complexity of our models, variance is high and overfitting is likely. Here are some comparisons of results with and without regularization (fig. 13).

Regular-ization	Val acc	Test acc	Val loss	Train loss
Without	0.9271	0.9211	0.2294	0.0160
With	0.9606	0.9518	0.1126	0.0610

fig. 13: accuracy and loss with and without regularization

We have also tracked our training and validation losses per batch. (Appendix D) densenet161_224_reg_4 indicates our training logs with regularization and densenet161_224_4 is training without regularization. The regularized model has a lower validation loss and a higher training loss compared to the unregularized model. This indicates that our regularization techniques have effectively discouraged the model from overfitting to the training data. It has also been able to better generalize to unseen data (validation data). This is further supported by our data from the above table.

Weight decay was chosen given the many parameters of our model. Learning rate scheduler was also employed to decay the learning rate by a factor of 0.1 every 2 epochs so that our model could converge to some minimum.

Image augmentation techniques were picked while glancing through our data. Rotations and RandomCrops were found to be effective in mimicking slight differences that hospitals had when taking their X-Rays and positioning patients. Not only that, we did slight ColorJitters to alter the contrast and brightness to mimic differences in X-Rays that could have resulted from the use of different machines. By augmenting our training dataset like so, we were hoping to get our model to ignore such noise from the data and focus on the actual signal such as the prevalence of GGO in the lung.

5.1.2 Deep Learning Model Comparisons

From section 4.3.1, we can see that DenseNet121 came out on top for non-ensembled models (fig. 5). While the ensembled model technically did better, it only did so by a small percentage while requiring much more computational resources to train. In the end, we thus chose DenseNet121 as our winner.

We also tried various depths to find a model of that architecture variant that suits the complexity of the problem at hand.

5.1.2.1 Model Depth

From the table of model comparisons(Appendix E), we can see that in any of the model architectures, shallower networks performed either better or equal to deeper networks. This suggests that features required to classify CXRs for COVID, VP or normal lungs only require simpler features and thus shallower networks. This has been corroborated by other research on CXRs[31]. This observation is made however with a small dataset in our project. The data constraint could thus have led to an equal or if not poorer performance in deeper networks as the deeper models did not have enough variance in the data to learn the true signal.

5.1.2.2 DenseNet121

DenseNets are arguably better at medical imaging given the need for observing simpler and smaller patterns like GGOs in CXRs. DenseNets could have excelled at this given the fact that every subsequent layer in DenseNet concatenates features from all previous layers. To fully test this hypothesis we could have made a replica of DenseNet121 without the concatenations. However, due to time constraints and our lack of knowledge, we did not test our hypothesis to that extent and were only able to observe the difference in performance of existing architectures.

5.1.2.3 Ensembling

We picked models that performed the best in their architecture type and thus chose ResNet34, DenseNet121 and VGG16. We felt that it was important to choose different architectures as each could be better at picking different features. We thus ended up with a model that topped the score with 0.3% greater test accuracy than DenseNet121. That being said, the ensemble was only slightly better and was in some sense not as effective given the computational costs of training and using an ensemble versus using a single DenseNet.

5.1.3 Metadata (Age, Gender)

The usage of metadata was unfortunately not very fruitful. This could have been due to the lack of images with metadata. Our model thus could not learn the differences between the labels as it did not have enough examples. That being said, given that it has had some level of success, it could also be indicative of its potential but just that we do not have enough data to prove it. (Appendix F).

5.1.4 Unsupervised Label Subclasses

The usage of the trained CNN to generate features to cluster via K-means was shown to be the most promising. This could possibly be because our trained CNN had internally separated the labels in their feature representations.

There also appears to be some learnable separation created from K-means clustering, but we were not able to determine if the subclasses generated were representative of anything in terms of severity of lung condition. But based on the confusion matrix, there appears to be some signal in the subclasses generated as our model was able to predict about 70% of the new labels accurately. The data here was also fairly balanced (Appendix G). This could be a useful method in generating meaningful subclasses with the help of radiographers to do manual checks.

5.2 Deep Learning V.S. Traditional

In the process of attempting both deep learning and traditional machine learning techniques, and comparing metrics such as accuracy, sensitivity score, etc. Both deep learning and traditional models appeared to perform well, achieving high sensitivity and accuracy rates. However, deep learning has several clear advantages over machine learning.

5.2.1 Feature Extraction

The main advantage of deep learning is that it allows models to learn to extract more meaningful features from images without requiring manual feature extraction, which results in better clustering of each class of image, and thus better performance of the models. Using TSNE (fig. 14), we can visualize these improvements by reducing the dimensionality of these feature spaces into 2 dimensions..

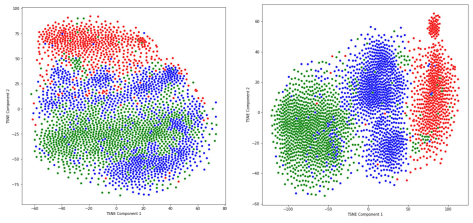


fig. 14: (Left) TSNE clustering in raw pixel intensity feature space
fig. 15: (Right) pre-trained vgg extracted features

We can see that before we even retrain the networks on our dataset, by using the upper layers of these networks as feature extractors, we can see that the three classes are already more distinctly clustered, and the clusters appear to be almost linearly separable (fig. 15). This also helped us to discover that the dataset included several COVID-19 CT scans, which correlates to the small dense red cluster of images (fig. 14; upper right), which we removed since our focus was on CXRs.

6. Conclusion

6.1 Summary

Overall, by improving the dataset we used through aggregation, and augmentation, and transfer learning, we managed to achieve similar results to existing studies on COVID-19 detection using deep learning[30].

And by comparing traditional machine learning and deep learning, we found that deep learning networks which utilize CNNs provide significant advantages over traditional machine learning.

Also, by analyzing our weights using CAM, we have shown that re-trained image networks might indeed be a step in the right direction.

6.2 Limitations

6.2.1 Metrics

Initially we did not properly consider what metric we should use to analyze our results. As such, the specificity metric is absent from our results.

6.2.2 Further Potential for Improvement (Learned Features)

One main limitation was that for a few samples, our model regions of interest did not lie around the lung regions within the CXR. Using the CAMs from previously, we generated the top-5 validation loss heatmap images (fig. 16) and found that the model may have picked out similar looking features which lie outside of the lungs.

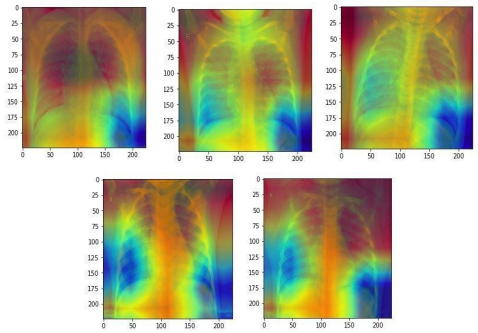


fig. 16: CAMs for High validation loss CXRs
(Left to right, Top) 8.55, 8.41, 8.40;
(Left to right, Bottom) 7,96, 7,82

This could possibly indicate that the model was not trained on enough CXR images and may need additional layers in order to build up a better internal representation of lungs within CXRs.

This also shows that there is potential to improve upon the already relatively high specificity, accuracy, and F1-score of the models.

6.2.3 Data Quality

The data we used were all gathered online and while reputable (Kaggle verified and from the University of Montreal), some samples were labelled by professional radiographers and were not confirmed diagnoses of COVID. This means that there could be some noise in our dataset or biases introduced.

Not only this but some of the CXRs could have been taken from the same patient on the same day. Such images while technically different might have leaked some information between our training and validation subsets.

6.2.4 Interpretability of Results.

Another flaw is the lack of interpretability of the model. For medical diagnosis, a simple binary classification of COVID-19 positive, or COVID-19 negative might not be suitable.

6.3 Possible Future Avenues for Improvement

Again, despite the flaw, considering that our results are quite high. We believe that there is still room for improvement.

6.3.1 Segmentation

One possible area that we can explore in the future would be to use lung segmentation to enhance the performance of our models. We could use image segmentation to first isolate the lung regions before feeding the processed CXR images to our CNN models. This could help the model to learn features only within the lung regions and ignore, for example, features which may appear similar to indicative features but are not within the lung.

6.3.2 Pre-processing

There is also another possible pre-processing step which we had not yet known previously which may improve detection of critical features within CXRs. One such technique is to apply Contrast-Limited Adaptive Histogram Equalization (CLAHE). This takes into account the fact that X-ray images are predominantly dark and hard to evaluate visually and as such, it adjusts image intensities to enhance the contrast by redistributing the pixel intensity. This might improve the training and testing performance of our models.

6.3.3 Dimensionality Reduction with Autoencoders

Due to our inexperience, we did not know of auto-encoders until late into this project. One area in which we could have explored to do more meaningful dimensionality reduction with regards to CXRs would be to train an auto-encoder on CXR images, instead of doing transfer learning since medical imaging analysis differs significantly from image classification.

References

1. Worldometer (2020, Nov 11). Coronavirus Update (Live). Retrieved November 11, 2020 from:https://www.worldometers.info/coronavirus/

2. Gallagher, J. (2020, October 27). Covid: How close are we to a vaccine? Retrieved November 11, 2020, from https://www.bbc.com/news/health-51665497

3. Brandon May (2020, Oct 15) Pros and Cons of the Common Types of COVID-19 Tests. Retrieved from: https://www.biospace.com/article/pros-and-cons-of-the-common-types-of-covid-19-tests/

4. Corman, V. M., Landt, O., Kaiser, M., Molenkamp, R., Meijer, A., Chu, D. K., Bleicker, T., Brünink, S., Schneider, J., Schmidt, M. L., Mulders, D. G., Haagmans, B. L., van der Veer, B., van den Brink, S., Wijsman, L., Goderski, G., Romette, J. L., Ellis, J., Zambon, M., Peiris, M., … Drosten, C. (2020). Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin, 25(3), 2000045. https://doi.org/10.2807/1560-7917.ES.2020.25.3.2000045

5. Jarrom D, Elston L, Washington J, et al (2020, Oct 01): Effectiveness of tests to detect the presence of SARS-CoV-2 virus, and antibodies to SARS-CoV-2, to inform COVID-19 diagnosis: a rapid systematic reviewBMJ Evidence-Based Medicine Published Online First.
doi: 10.1136/bmjebm-2020-111511

6. Xie, X., Zhong, Z., Zhao, W., Zheng, C., Wang, F., & Liu, J. (2020). Chest CT for Typical Coronavirus Disease 2019 (COVID-19) Pneumonia: Relationship to Negative RT-PCR Testing. Radiology,

296(2), E41–E45. https://doi.org/10.1148/radiol.2020200343

7. Robert H. (2020, Aug 10) Which test is best for COVID-19? Retrieved from: https://www.health.harvard.edu/blog/which-test-is-best-for-covid-19-2020081020734

8. An overview of COVID-19, with emphasis on radiological features (2020, Mar 27). Hong Kong College of Radiologists. Retrieved Nov 13, 2020 from: https://www.hkcr.org/news.php/events_list/cid,3/nid,176

9. Jain, G., Mittal, D., Thakur, D., & Mittal, M. K. (2020). A deep learning approach to detect Covid-19 coronavirus with X-Ray images. Biocybernetics and biomedical engineering, 40(4), 1391–1405. https://doi.org/10.1016/j.bbe.2020.08.008

10. Yoon SH, Lee KH, Kim JY, Lee YK, Ko H, Kim KH, Park CM, Kim YH (2020, Apr 21). Chest Radiographic and CT Findings of the 2019 Novel Coronavirus Disease (COVID-19): Analysis of Nine Patients Treated in Korea. Korean J Radiol. https://doi.org/10.3348/kjr.2020.0132

11. Sana Salehi, Aidin Abedi, Sudheer Balakrishnan, and Ali Gholamrezanezhad (2020) : Coronavirus Disease 2019 (COVID-19): A Systematic Review of Imaging Findings in 919 Patients. American Journal of Roentgenology 215:1, 87-93

12. 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 Pulm Med 20, 245. https://doi.org/10.1186/s12890-020-01286-5

13. Hong Kong College of Radiologists (2020, Mar 27). An overview of COVID-19, with emphasis on radiological features. Retrieved from: https://www.hkcr.org/lop.php/COVID19

14. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. (2018). Artificial intelligence in radiology. Nature reviews. Cancer, 18(8), 500–510. https://doi.org/10.1038/s41568-018-0016-5

15. Klang E. (2018). Deep learning and medical imaging. Journal of thoracic disease, 10(3), 1325–1328. https://doi.org/10.21037/jtd.2018.02.76

16. Rajaraman, S.; Candemir, S.; Kim, I.; Thoma, G.; Antani, S. (2018), Visualization and Interpretation of Convolutional Neural Network Predictions in Detecting Pneumonia in Pediatric Chest Radiographs.

17. Abiyev, R. H., & Ma'aitah, M. (2018). Deep Convolutional Neural Networks for Chest Diseases Detection. engineering,https://doi.org/10.1155/2018/4168538

18. Stephen, O., Sain, M., Maduh, U. J., & Jeong, D. U. (2019). An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare. Journal of healthcare engineering, https://doi.org/10.1155/2019/4180949

19. S. Xu, H. Wu and R. Bie (2019), "CXNet-m1: Anomaly Detection on Chest X-Rays With Image-Based Deep Learning," in IEEE Access, vol. 7, pp. 4466-4477 doi: 10.1109/ACCESS.2018.2885997.

20. Saraiva, A.A., Santos, D.B., Costa, N.J., Sousa, J.V., Ferreira, N.M., Valente, A., & Soares, S. (2019). Models of Learning to Classify X-ray Images for the Detection of Pneumonia using Neural Networks.

21. Tulin Ozturk, Muhammed Talo, Eylul Azra Yildirim, Ulas Baran Baloglu, Ozal Yildirim, U. Rajendra Acharya (2020) Automated detection of COVID-19 cases using deep neural networks with X-ray images. https://doi.org/10.1016/j.compbiomed.2020.103792

22. Mesut Toğaçar, Burhan Ergen, Zafer Cömert (2020). COVID-19 detection using deep learning models to exploit Social Mimic Optimization and structured chest X-ray images using fuzzy color and stacking approaches, https://doi.org/10.1016/j.compbiomed.2020.103805.

23. Harsh Panwar, P.K. Gupta, Mohammad Khubeb Siddiqui, Ruben Morales-Menendez, Vaishnavi Singh. (2020) Application of deep learning for fast detection of COVID-19 in X-Rays using nCOVnet, https://doi.org/10.1016/j.chaos.2020.109944.

24. Apostolopoulos, I.D., Mpesiana, T.A. (2020) Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. https://doi.org/10.1007/s13246-020-00865-4

25. Zhang, R., Guo, Z., Sun, Y., Lu, Q., Xu, Z., Yao, Z., Duan, M., Liu, S., Ren, Y., Huang, L., & Zhou, F. (2020). COVID19XrayNet: A Two-Step Transfer Learning Model for the COVID-19 Detecting Problem Based on a Limited Number of Chest X-Ray Images. https://doi.org/10.1007/s12539-020-00393-5

26. Rahman, T. (2020, Mar 28). COVID-19 Radiography Database. Retrieved Nov 11, 2020 from: https://www.kaggle.com/tawsifurrahman/covid19-radiography-database

27. GitHub - ieee8023. (2020, Oct 1).covid-chestxray-dataset. Retrieved (2020, Nov 11) from: https://github.com/ieee8023/covid-chestxray-dataset/

28. Abbas, A., Abdelsamea, M.M. & Gaber, M.M. (2020) Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network. https://doi.org/10.1007/s10489-020-01829-7

29. GitHub - ieee8023. (2020, Oct 1). covid-chestxray-dataset. retrieved from: https://github.com/ieee8023/covid-chestxray-dataset/

30. Minaee, S., Kafieh, R., Sonka, M., Yazdani, S., & Soufi, G. (2020, July 21). Deep-COVID: Predicting COVID-19 From Chest X-Ray Images Using Deep Transfer Learning. Retrieved November 11, 2020, from: https://arxiv.org/abs/2004.09363

31. Bressem, K.K., Adams, L.C., Erxleben, C. et al. Comparing different deep learning architectures for classification of chest radiographs. Sci Rep 10, 13590 (2020). https://doi.org/10.1038/s41598-020-70479-z

Picture sources

a. Normal CXR. Image credits: Assoc. Prof. Craig Hacking. Retrieved Nov 13, 2020 from: https://radiopaedia.org/cases/normal-cxr-and-lateral

b. CXR depicting pleural effusion. Image credits: James Heilman, MD, n.d. Retrieved Nov 13, 2020 from: https://www.medicalnewstoday.com/articles/318021

c. CXR of a patient with COVID-19 depicting ground glass opacities. Retrieved Nov 13, 2020 https://www.eurorad.org/case/16751

Appendix A - Data Gathering

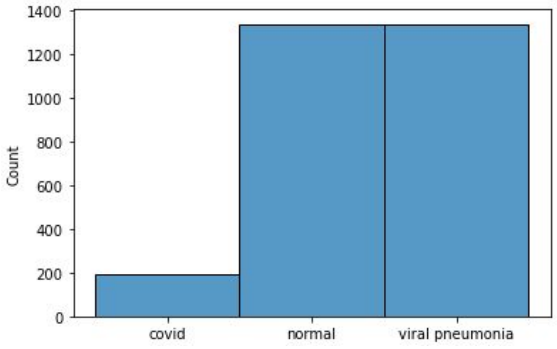


fig. 1: Original Dataset from Kaggle's COVID-19 Radiography Database

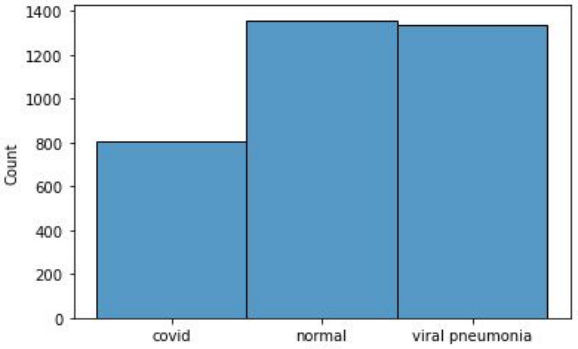
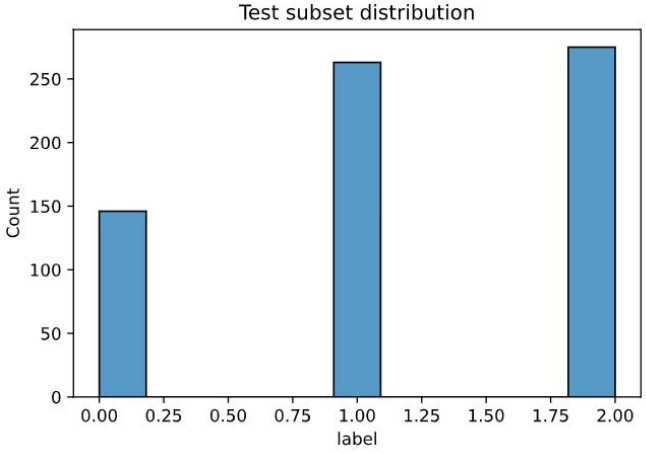
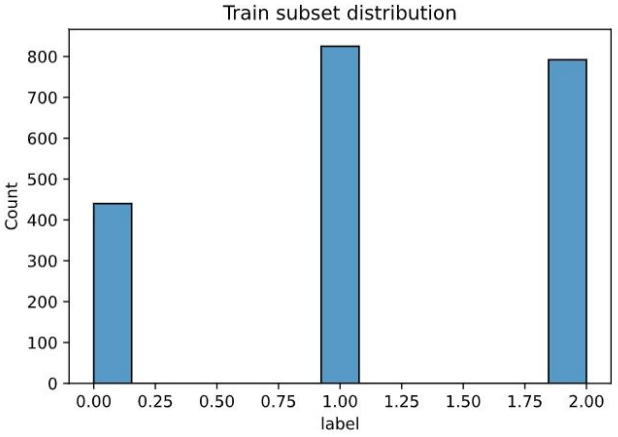
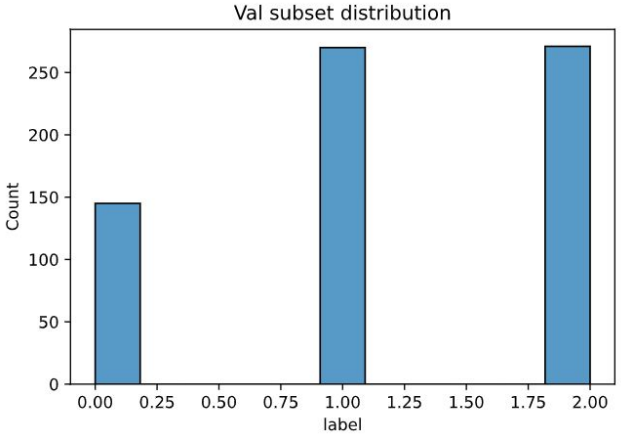


fig. 2: Class distribution of aggregated dataset using Kaggle and University of Montreal's Dataset

Appendix B - Distribution of subsets used for Deep Learning training







Appendix C - Random Crop illustration

fig. 1: Same image with different random crop

fig2. : Same image with different random crop

Appendix D - Overfitting and regularization

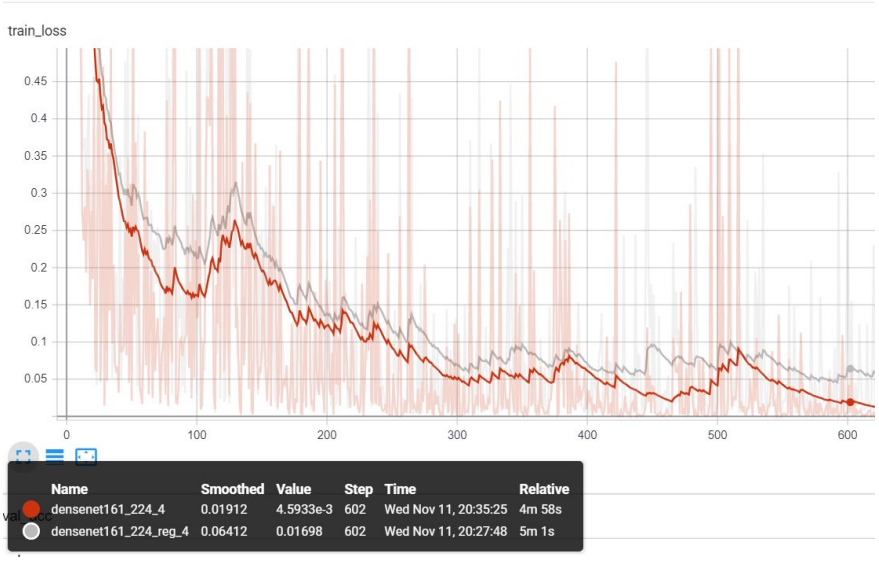


fig. 1: Training Loss Comparison with and without regularization (Grey is with regularization, brown is without)

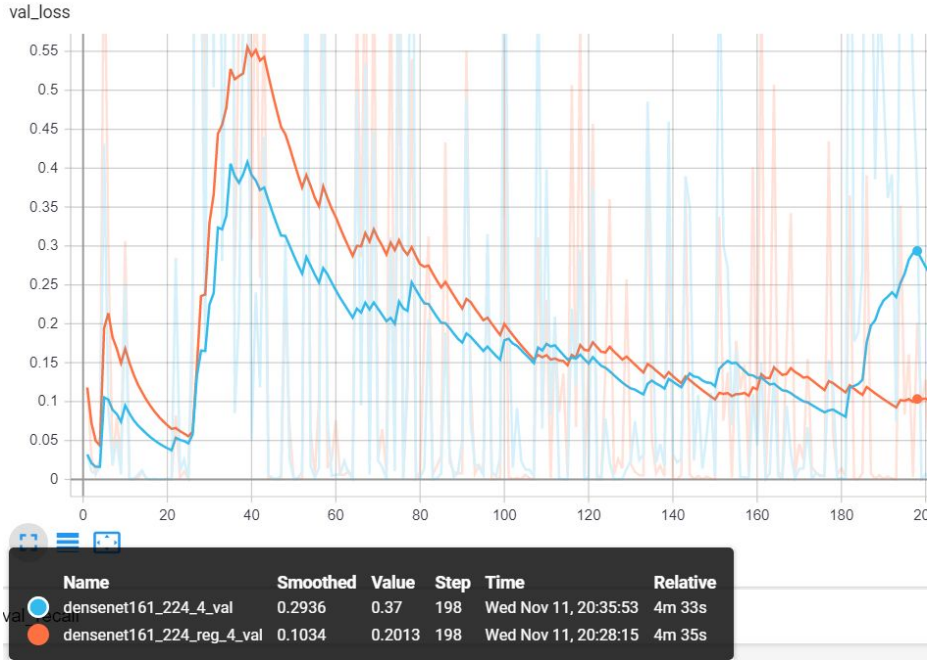


fig. 2: Validation Loss Comparison with and without regularization (Orange is with regularization, Blue is without)

Appendix E - Deep Learning CNN model performances

Trial ID	Show Metrics	val_acc	test_acc	val_recall	test_recall
densenet121_224...	☐	0.96210	0.96053	0.98621	0.98630
densenet161_224...	☐	0.95481	0.95614	0.98621	0.97260
densenet201_224...	☐	0.95627	0.95029	0.97931	0.97945
resnet101_224_0\...	☐	0.95190	0.95029	0.97931	0.97945
resnet34_224_0\1...	☐	0.95190	0.95760	0.97931	0.97260
resnet50_224_0\1...	☐	0.95044	0.95322	0.97931	0.97945
ensemble_224_0\...	☐	0.95918	0.96345	0.97241	0.97945
vgg16_224_0\160...	☐	0.94023	0.93421	0.97241	0.96575
resnet18_224_0\1...	☐	0.93440	0.92690	0.95862	0.95205

fig. 1: Tensorboard table logging for models trained

Appendix F - Metadata label distribution

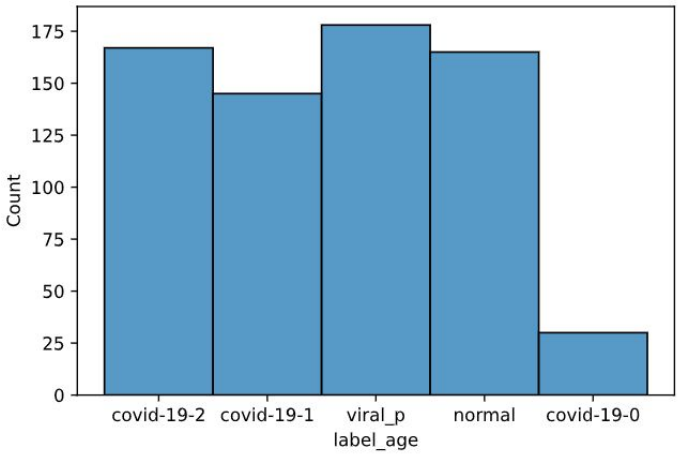


fig. 1: Metadata Age Label Spread
(The ranges below refer to ages)
0-30 (COVID-19-0), 31-60 (COVID-19-1), 61-100 (COVID-19-2).

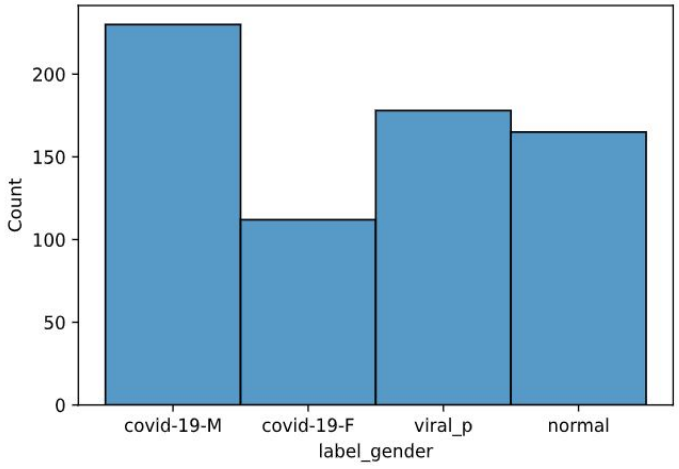


fig 2: Metadata Gender Label Spread
(M suffix for male, F suffix for female)

Appendix G - Unsupervised label generation distribution

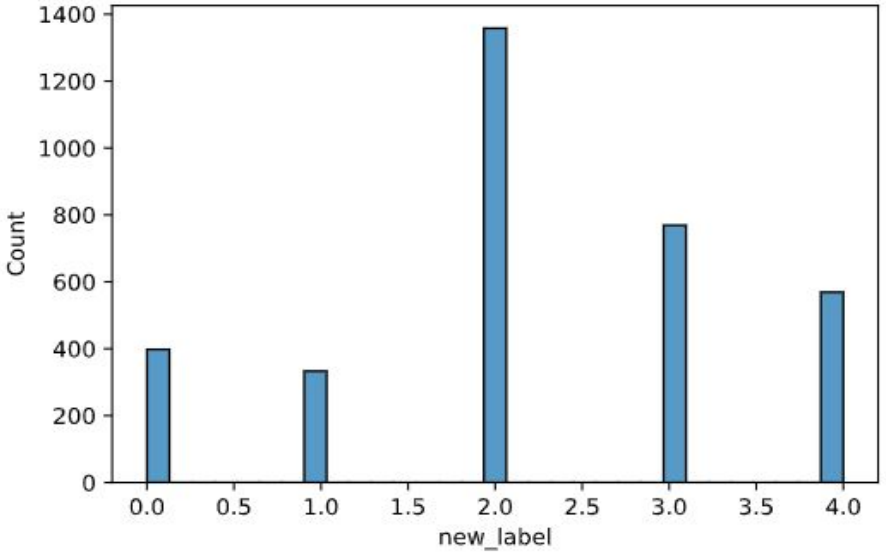


fig.1:
0.0 and 1.0 refers to subclasses of COVID-19;
2.0 is the Normal class;
3.0 and 4.0 are subclasses of VP;

Appendix H: Potential detail/feature loss in PCA reduced dimensionality data

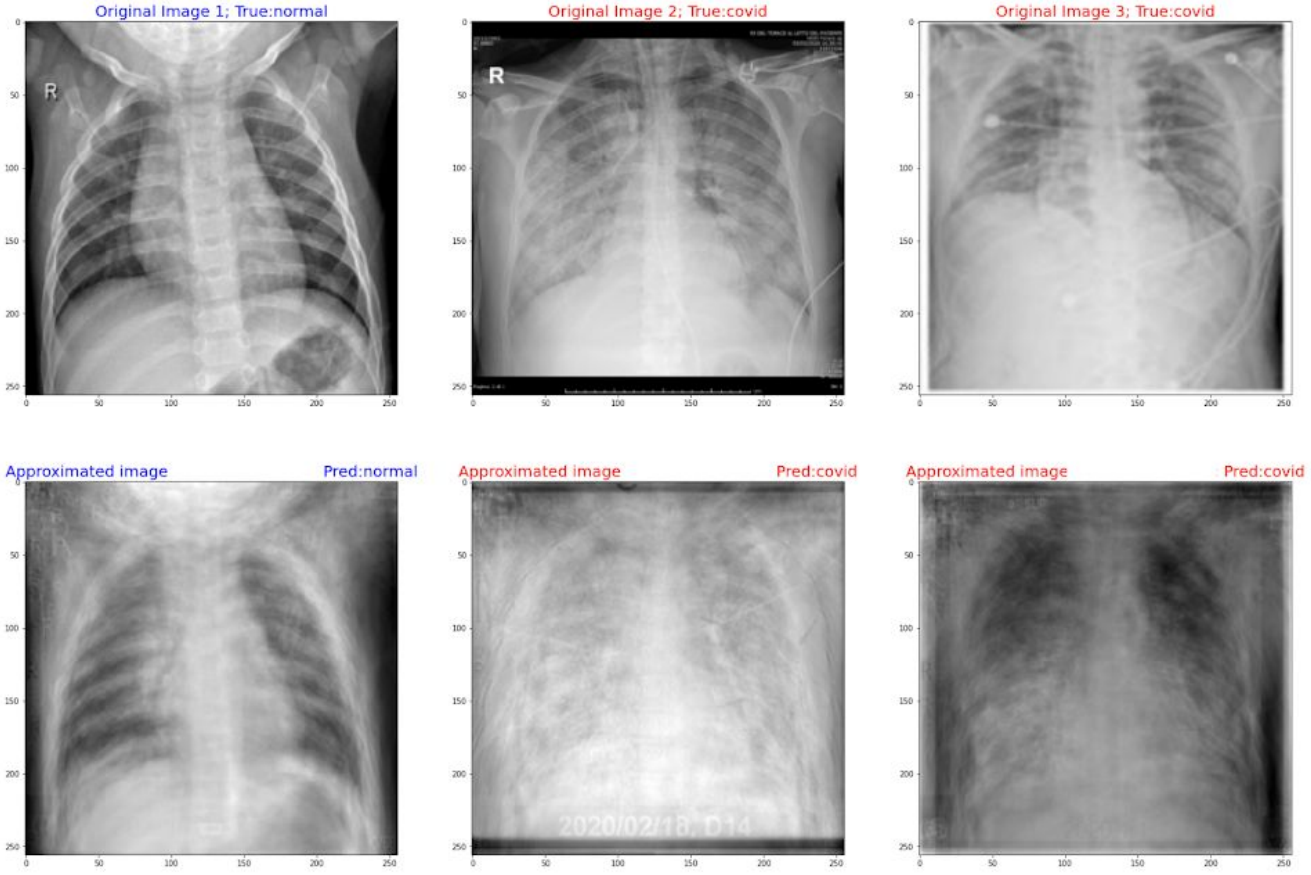


fig. 1-3 (Top, left to right) Original image
fig. 4-6:(Bottom, left to right) Reconstructed images from PCA reduced data.

Appendix J: Confusion Matrix for Support Vector Machine Classifier (fitting with VGG and flattened data)