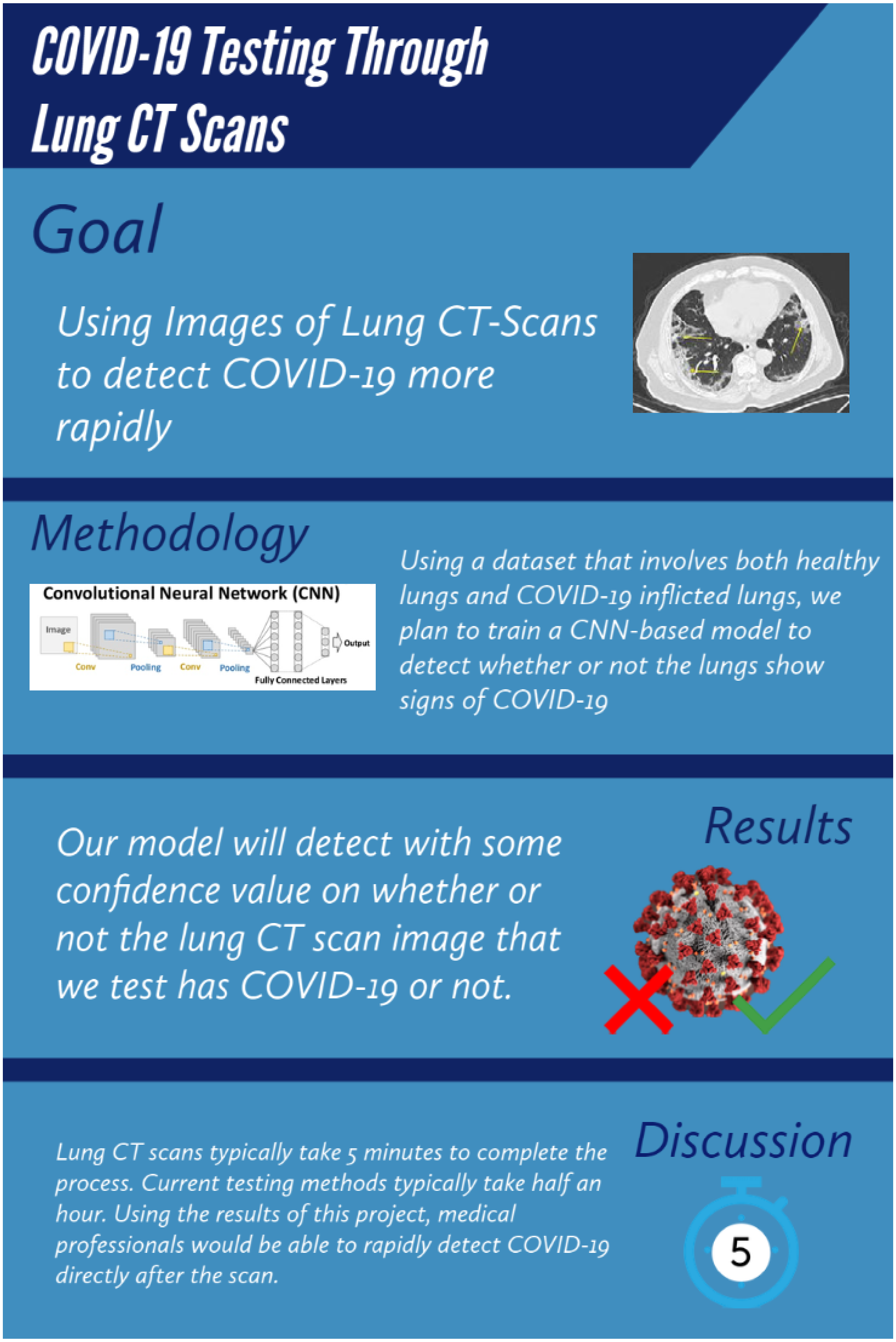
CT Scan COVID Diagnosis

CS 4641 Fall 2020: Team 39's Machine Learning Project

View Cs4641-group39.GitHub.io on GitHub

Summary



Introduction

Our goal is to create a model that will learn from the lung CT scan dataset to determine whether or not new patients are at risk for pneumonia as a result of being COVID-19 positive. With the new knowledge surrounding the virus, it becomes much more crucial that patients seek special attention if the virus has infected the lungs. With this model, doctors only need to upload the CT scan from the patient and the model will give vital, potentially life-saving information on the next step of treatment.

Methods

We plan to use lung CT Scans provided on kaggle to determine whether a patient has COVID-19. We're also planning to use a CNN-based model through Pytorch in order to train our model to identify COVID-19 through CT lung scans.

Results

Given a set of lung images (some patients positive for Covid-19, some patients exposed to other conditions such as pneumonia, and some healthy patients) our model will determine (to some % of confidence) whether or not a given patient's CT lung scan is Covid positive.

Discussion

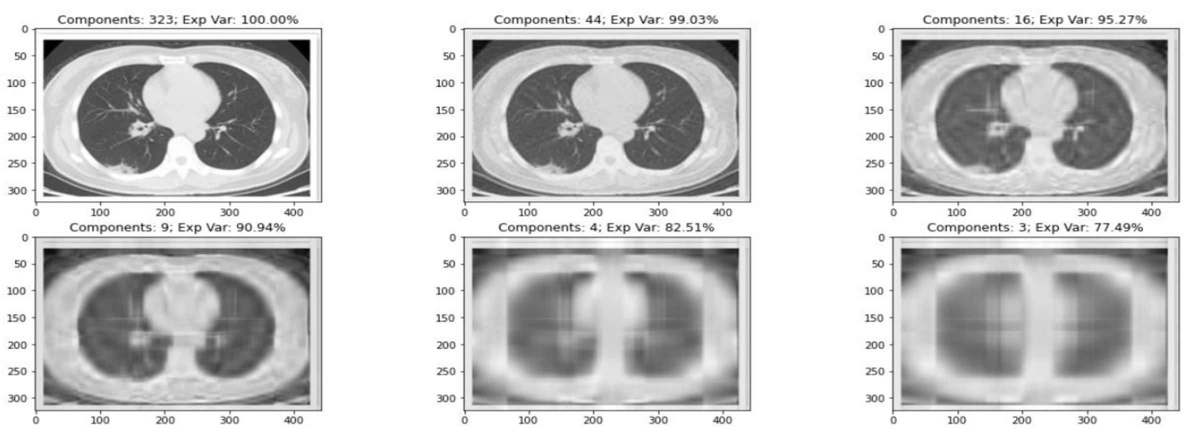
The significance of this relates back to the ease with which doctors will be able to rapidly confirm patients' condition, and subsequently be able to use this information for deciding which route to take with treatment. This will be especially significant for at-risk patients if the CT scan can reveal the results faster than a mouth or nose swab test.

Unsupervised Learning

PCA

PCA allows us to remove features in our data that contribute the least to the variance in the data. This can help quicken the training of our supervised models. In addition, it can reveal the most significant features within our data. Here is an example of applying PCA to a CT scan image:

Applying PCA (Negative Patient)

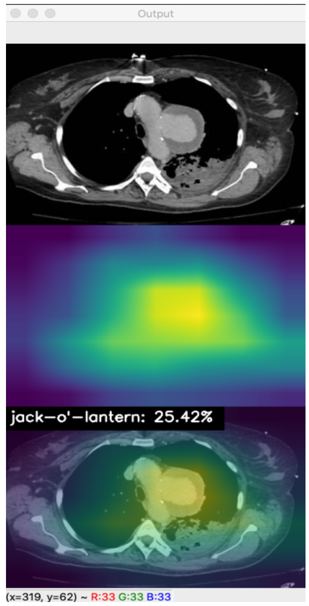


PCA was successfully implemented, but no use cases have been explored yet. We are potentially considering using it for our CNN, which will be implemented in the near future for supervised learning. For most test cases, 99% of variance was able to be recovered while dropping a majority of the components. This indicates that we can apply PCA to reduce the number of components to be analyzed; as a result, fewer computations have to be made to still yield an accurate model.

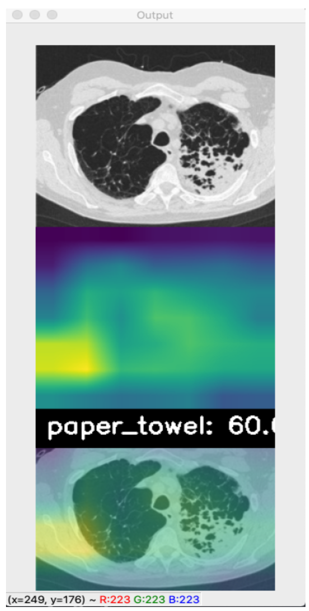
We can make some observations from the PCA result however. We can immediately tell that important features look like white spaces within the lung scan. It could be picking up on ground glass opacities, which are indicitive of illness in the lung. We will talk more about ground glass opacity in the next few sections.

CAM

Class Activation Maps highlight the parts of the image that the Neural Network focuses on so we can better understand how our neural network works in supervised learning when we reach that stage. CAMs gives us an idea of what part of the data has to be focused on by building a heat map. We pasted this outputted heat map over our original image to understand important visual features.
When we worked with The VGG16 architecture, the output seemed to focus more on the heart area than the rest of the lung scan. The model also didn't have the best accuracy.



When we worked with The resnet architecture, our output highlighted the ground glass opacity in the lungs with a much better accuracy.

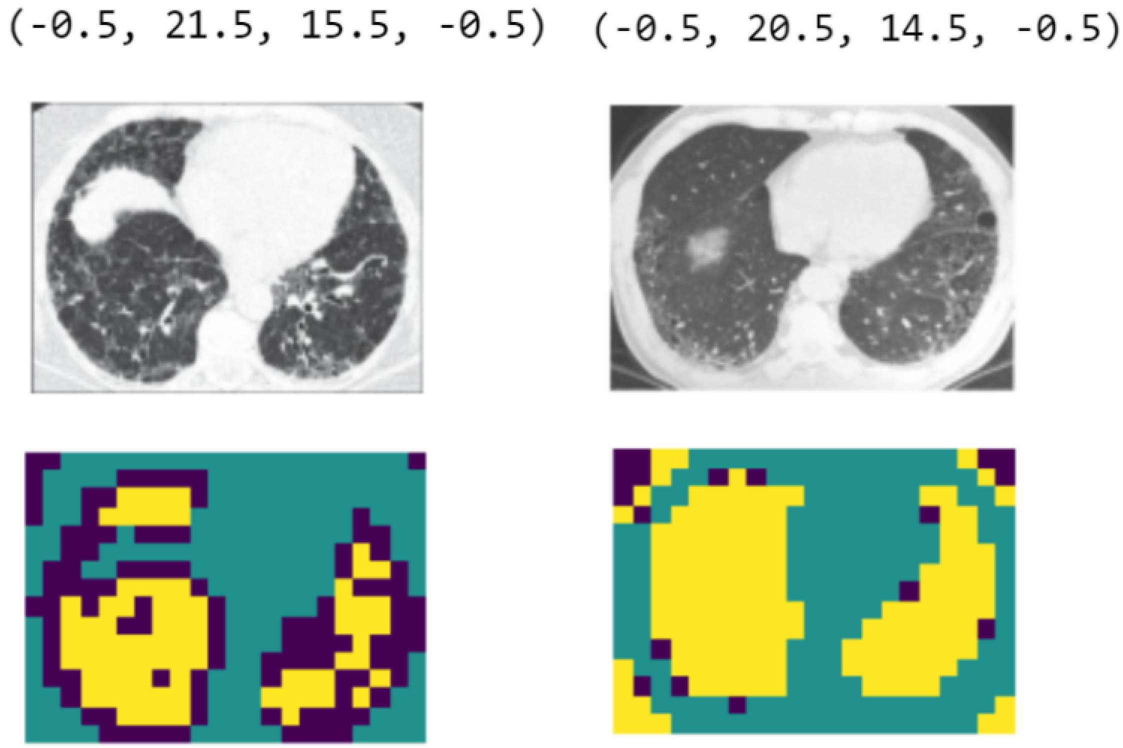


In conclusion, the CAM model tells us that ground glass opacity and the heart are possibly important features, a trend we observe with GMM and DBSCAN as well.

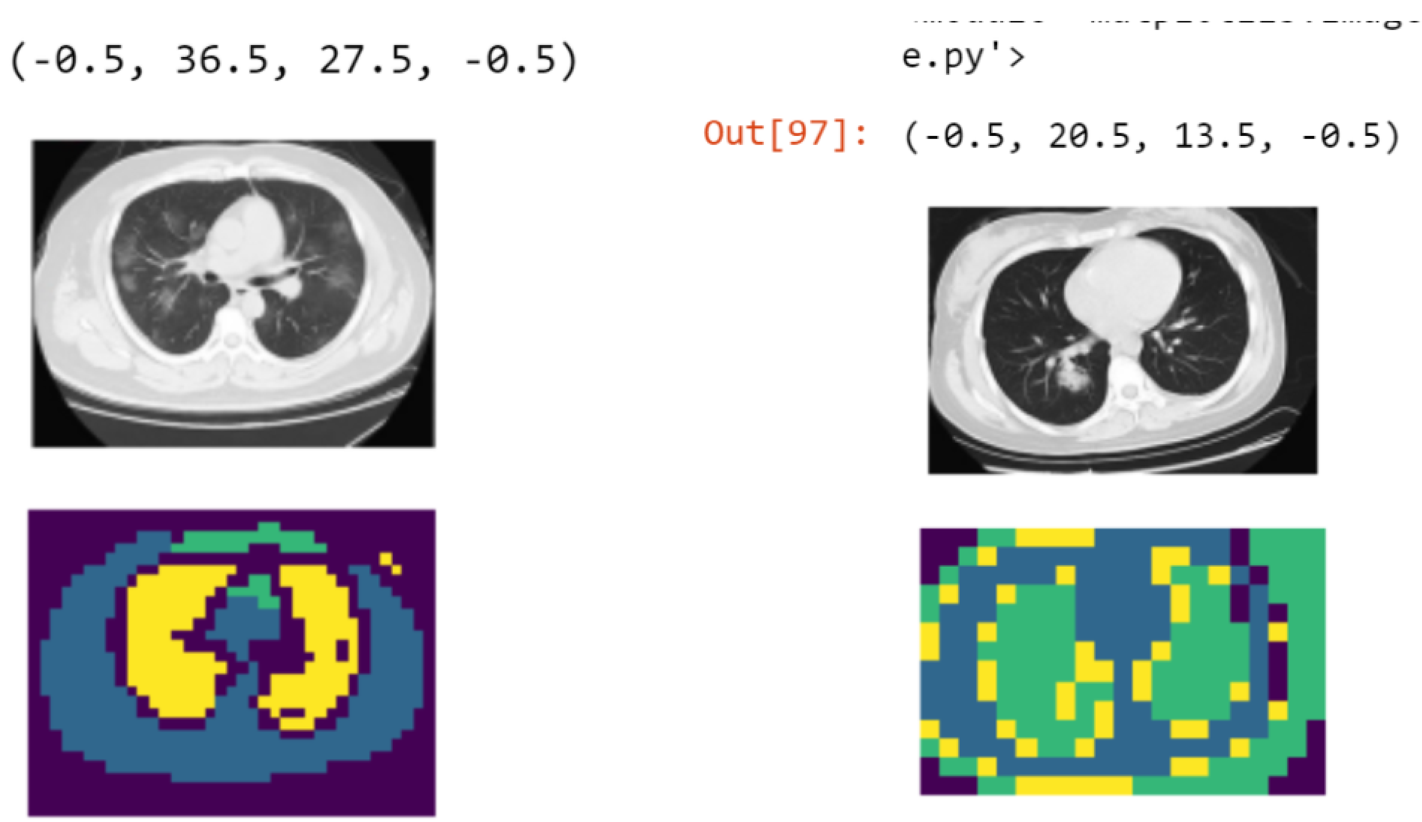
DBSCAN

DBSCAN is probably not the best way to go for image segmentation because the epsilon parameter is too sensitive and would need to take time to find the perfect value to segment properly. Each image that we ran through DBSCAN needed a drastically different epsilon in order to show results. Even though the resulting images are interesting and helped us realize that there are differences between COVID-positive and negative, it's not the best to reduce noise and portions of the images. Some additional understandings that we found through DBSCAN on the other hand was that many of the CT scans clustered the center area much more larger for negative cases and the center area was smaller for positive COVID cases. This could potentially related to the methods on how medical professionals take COVID CT-Scans but it was one of the observations that we made through our clustering.

COVID Negative DBSCAN Clustering



COVID Positive DBSCAN Clustering



GMM

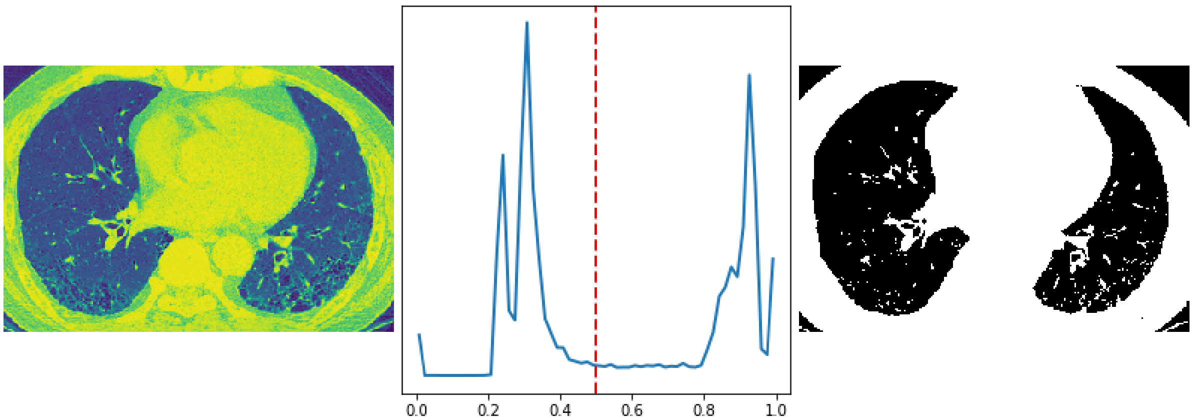
The results found from GMM has several implications towards further understanding what happens to a human lung when coming in contact with the Covid-19. In order to understand the findings from GMM, we must first discuss the capabilities and limitations of GMM in regards to clustering grayscale images.

In terms of capabilities, GMM has the ability to pick up ground glass opacity, which is abnormal findings on a CT scan which result in a hazy, obscured view of the underlying structure of the lungs. Because GMM can recognize these subtle changes, the algorithm bypasses the loss of information from this phenomenon, allowing it to draw focus towards anomalies in the lungs. Now, we certainly are no medical experts, but the ability of the algorithm to filter out the "noise" associated with the CT scans certainly has positive implications in the medical field, not just in terms of Covid-19, but CT scans in general.

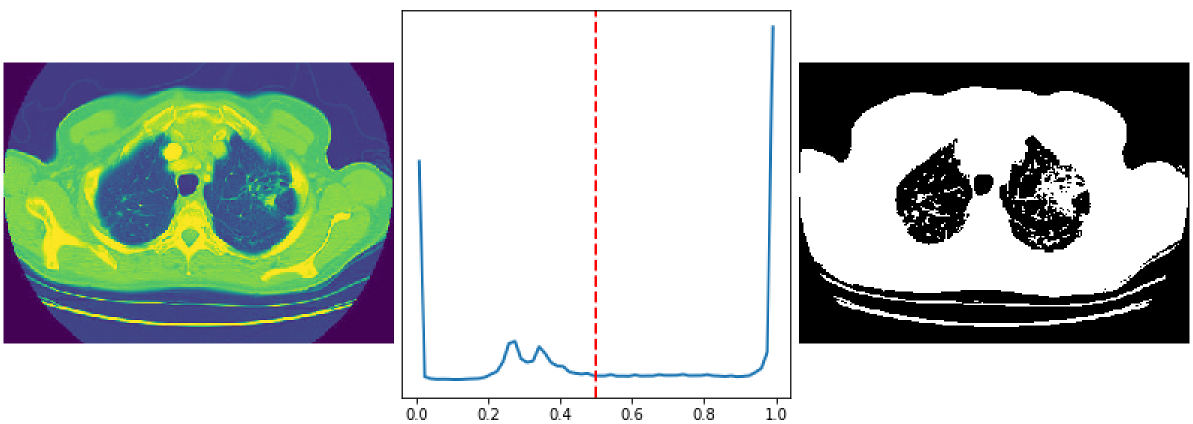
With regards to limitations, GMM was not able to process multiple images at one time, which would be excellent for classifying incoming CT scans. We had to run each image individually, one at a time. This was not only a time consuming process, but also provided the inability for the algorithm to prove it could correctly group negative scans against positive scans, which would certainly prove beneficial for supervised learning.

Below are samples of images generated from the GMM algorithm. The images are labeled as either a Covid-19 negative patient or Covid-19 positive patient. The leftmost image is the original CT scan, the graph represents the bin clustering, indicating how many clusters may exist in the data. The rightmost image is the clustered, grayscale image generated by GMM. We can draw the conclusion that, in general, the center between the lungs was bigger in Covid positive patients and smaller in Covid negative patients, which aligns with our findings from DBSCAN. Whether this is medically significant would be left up to the doctors, but it was consistant across many images.

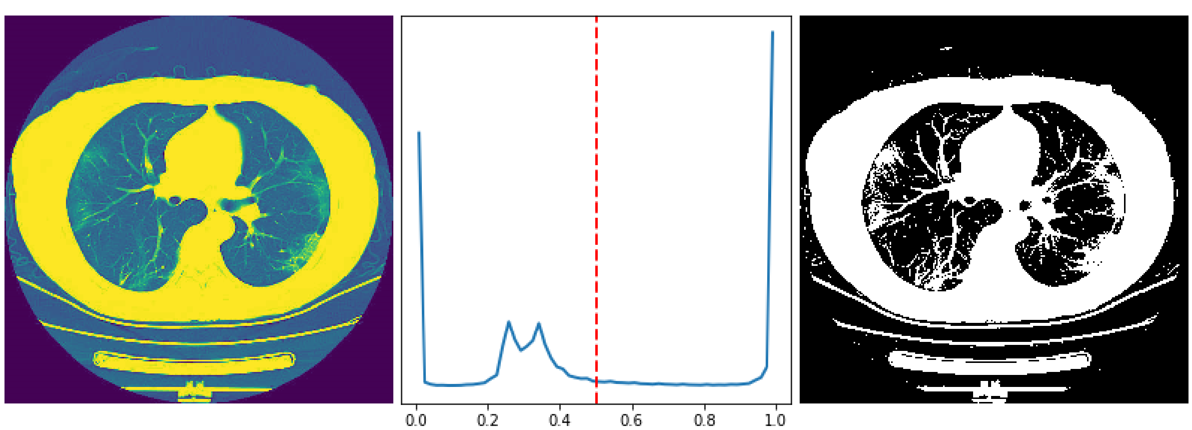
COVID Negative GMM Clustering (Image 1)



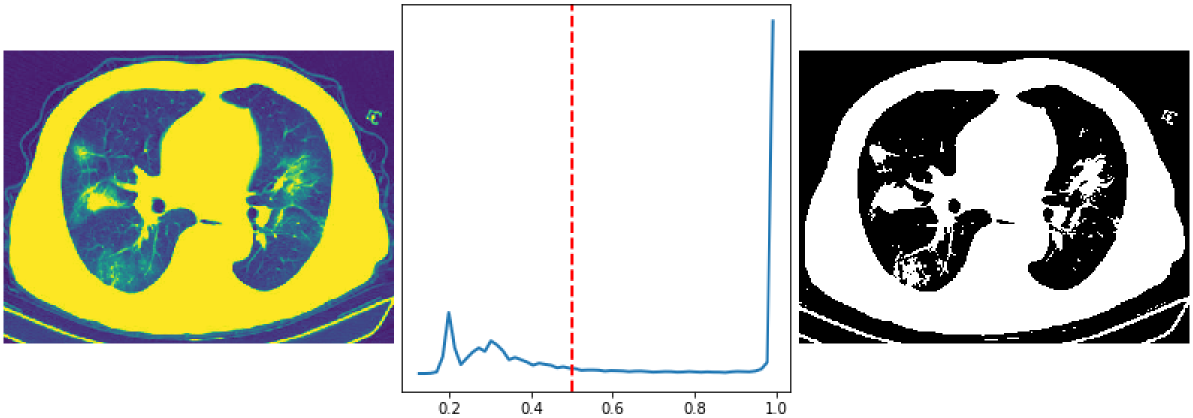
COVID Negative GMM Clustering (Image 2)



COVID Positive GMM Clustering (Image 1)



COVID Positive GMM Clustering (Image 2)



References

- [Description of pneumonia, and its impact on lungs]()
- [Types of damages to lungs]()
- [Extent of injury due to COVID-19 on the lungs]()
- [COVID-19 lung CT scan image dataset]()

[aazmi6]() maintains [Cs4641-group39.GitHub.io]()