A computational framework based on convolutional neural network for classifying interstitial lung disease in computed tomography scans

Christian Havranek, Skyler Penna, Dr. Gurman Gill

Department of Computer Science, Sonoma State University, Rohnert Park

Introduction

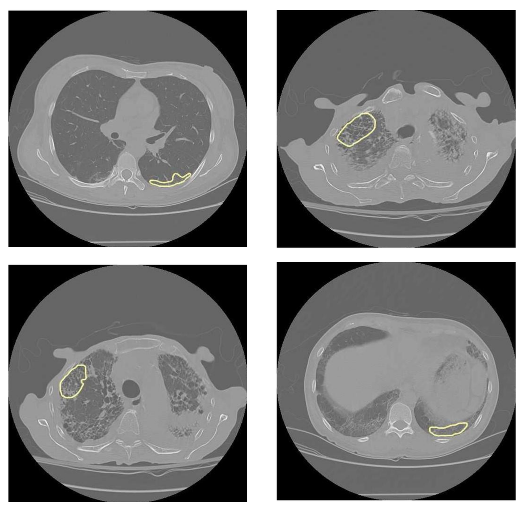
Chest Computed Tomography (CT) Scans are a form of X-ray that provide detailed images of organs so that they can be diagnosed as either healthy or diseased. CT scans are commonly used in the analysis of a lung to classify whether an Interstitial Lung Disease (ILD) is present or not. Computer Aided Diagnostic (CAD) systems are often used in conjunction with CT scans during the diagnosing process in an attempt to get quicker results. Convolutional Neural Networks (CNN) have proven to be a reliable method of choice in the use of CAD systems thanks to their diverse ability to learn and categorize different images based on what the network is being trained on. CNNs normally require large amounts of data in order to be accurate but the use of tools like transfer learning and our image patch extraction framework have allowed us to work with a smaller set of data.

Datasets

ImageNet

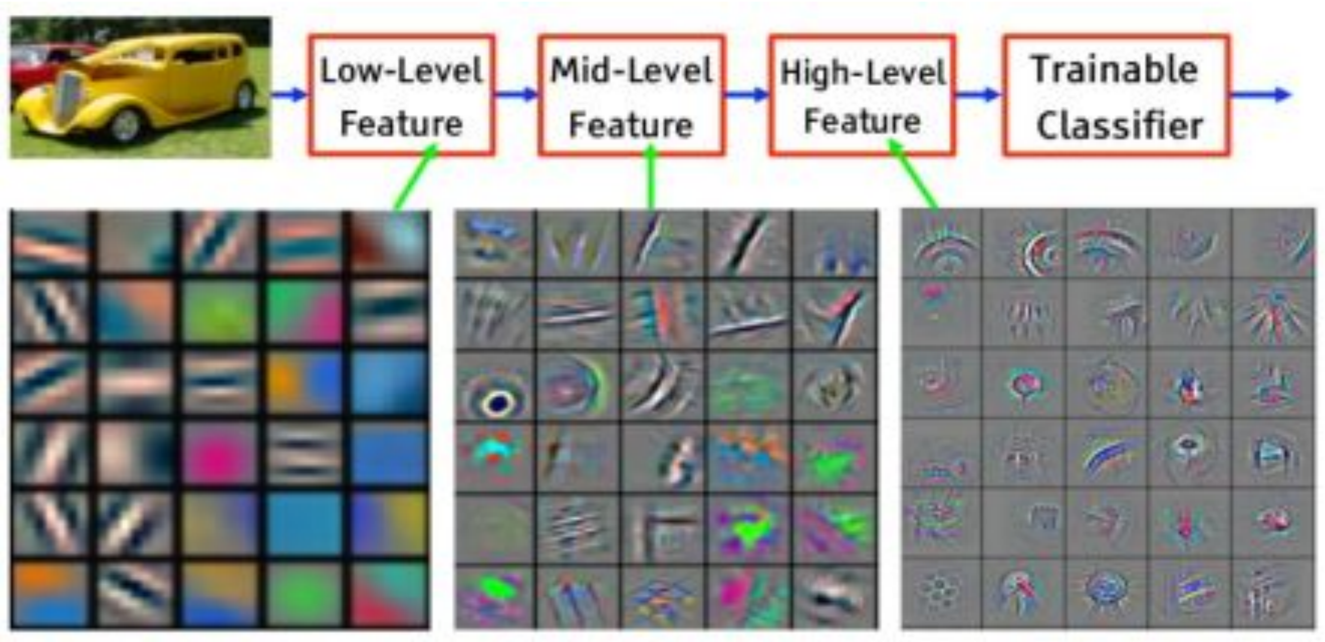
- 14,197,122 total images
- 21,841 Categories of images which include animals, plants, crops, people, cars, and many others which creates a diverse dataset to be used for different purposes.
- The large amount of images makes it a very appealing option to perform transfer learning when training various different Convolutional Neural Networks.

MedGIFT

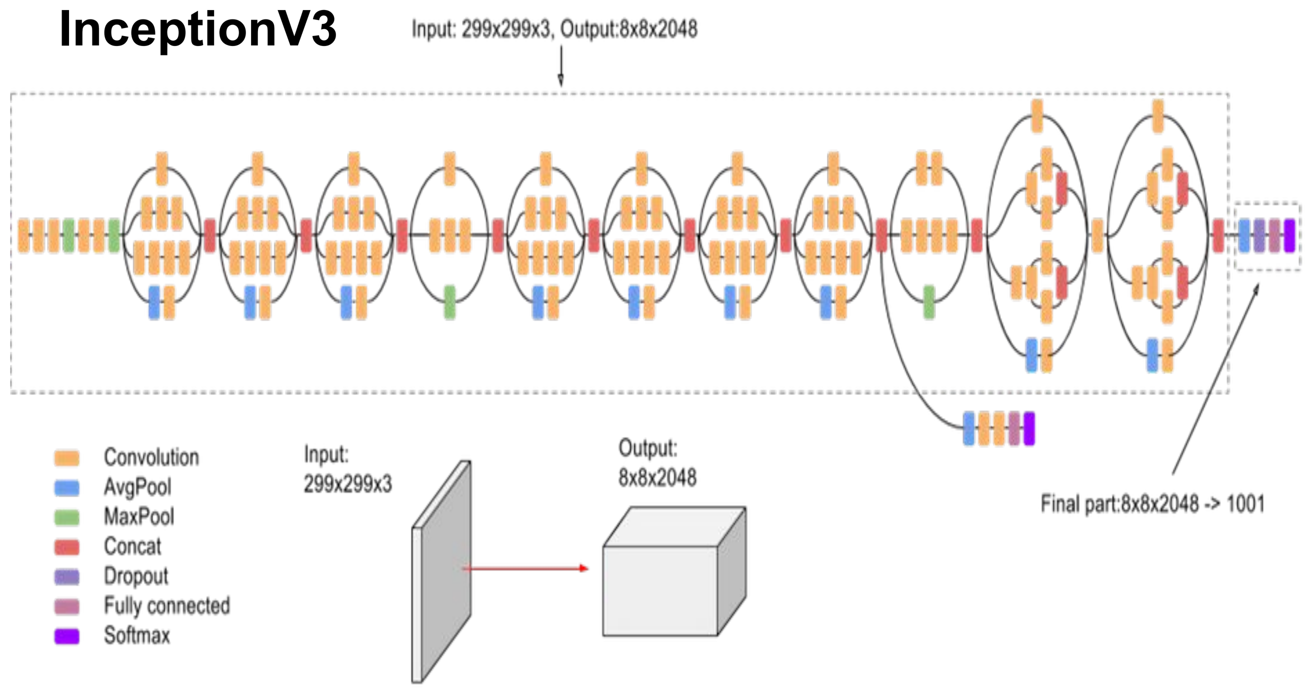


- 128 patients
- 1,946 Region of interests (ROI)
- 12 different potential disease labels for CT images (including healthy)
- Disease Labels used for testing: Emphysema, Fibrosis, Ground Glass, Healthy, Micronodules.
- Allows us to try working with different types of Interstitial Lung Diseases and is closer to the type of data that a neural network would be interacting with in a real word application.

Convolutional Neural Networks

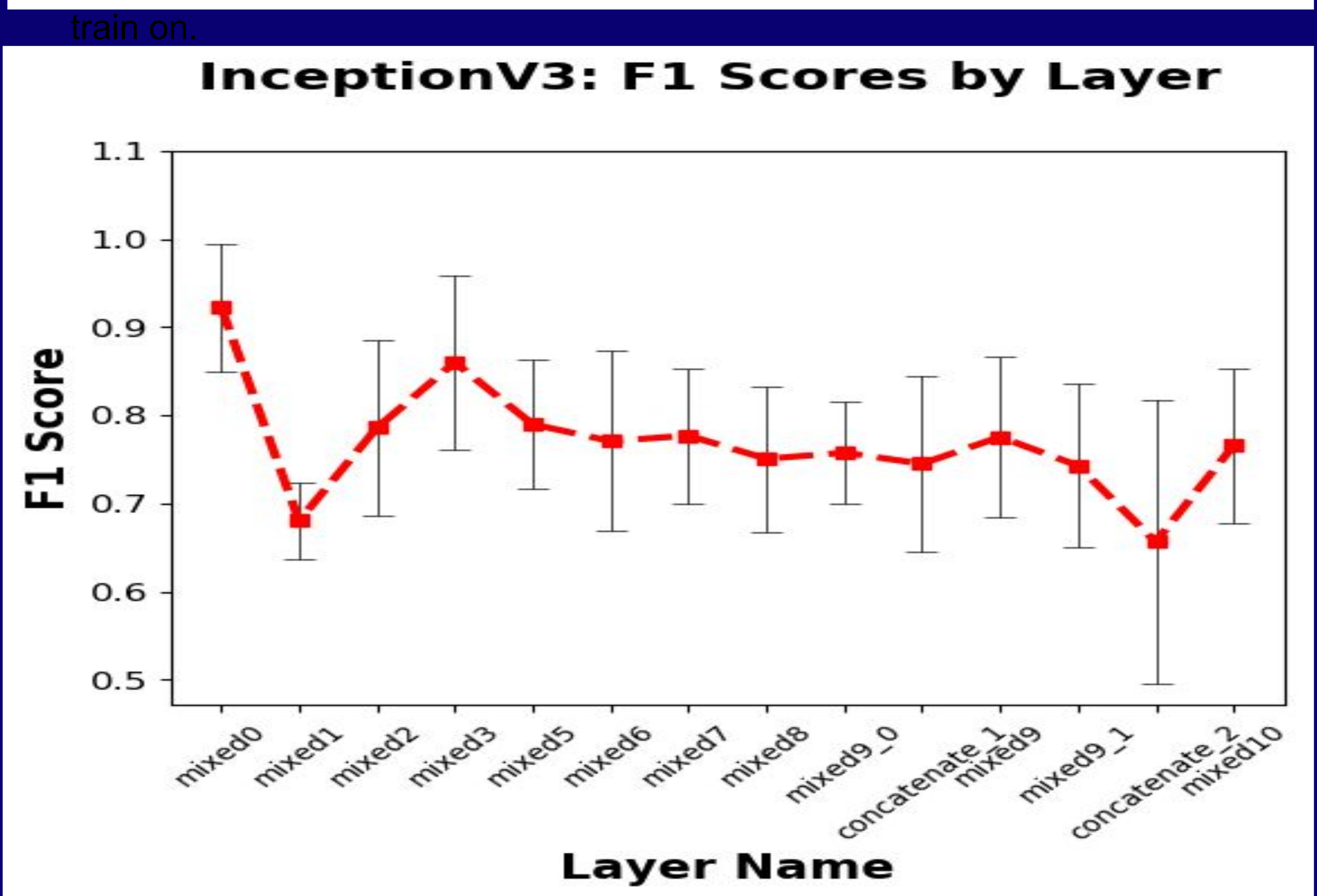


InceptionV3



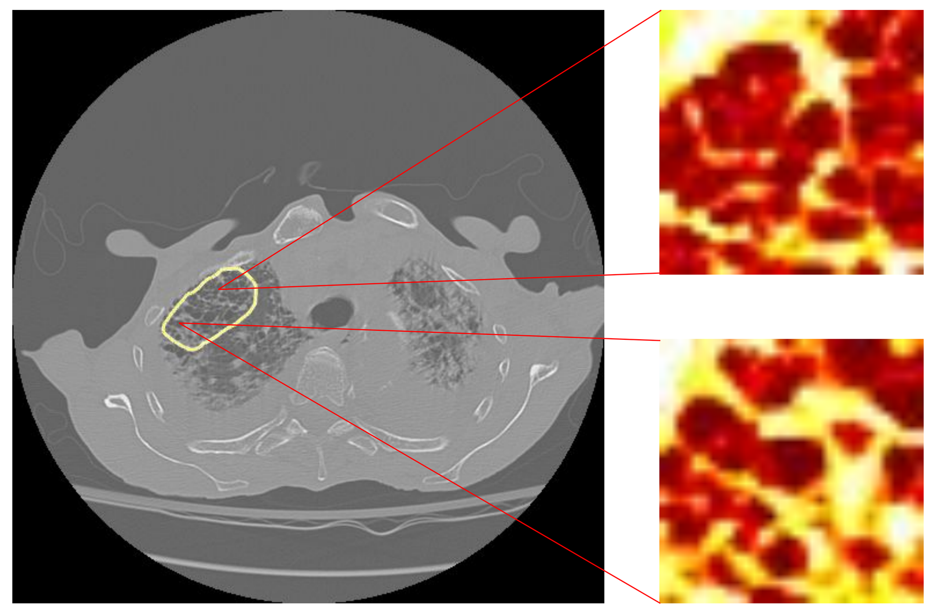
Transfer Learning:

- Using a model that has been trained with an alternative dataset. (ImageNet)
- Features learned from that previous training can be used to train and test on a new data set.
- Additional layers can be added to a network to train new data while utilizing the features from the previously trained model.
- Speeds up the process since the whole network may not need to be trained.
- Leads to greater accuracy if there is only a small amount of original data to train on



MedGIFT Dataset: Image Patch Framework

Fibrosis



Configurable Values:

- Patch Size: Desired size of the patch to be used while searching and extracting from the CT scan.
- Shift Value: How much the patch window should move while looking through the CT scan.
- Patch Qualification Value: How much of the ROI needs to be present in current patch to qualify for extraction.
- Specific Disease Labels: Can be any amount of the 12 diseases from the database.

Converting Hounsfield Units to RGB:

a. R: -1400 to -950, G: -1400 to -200, B: -160 to 240
b. R: -1400 to -853, G: -852 to -306, B: -305 to 240
c. R: -1400 to -601, G: -600 to -601, B: -200 to -200

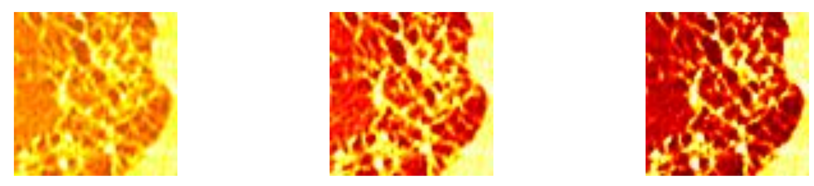


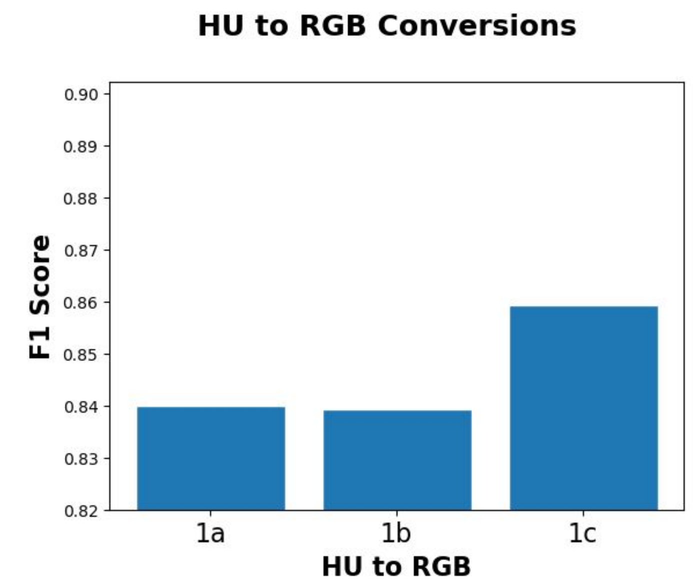
Figure 1: Comparing different HU to RGB methods.

Purpose of the Framework:

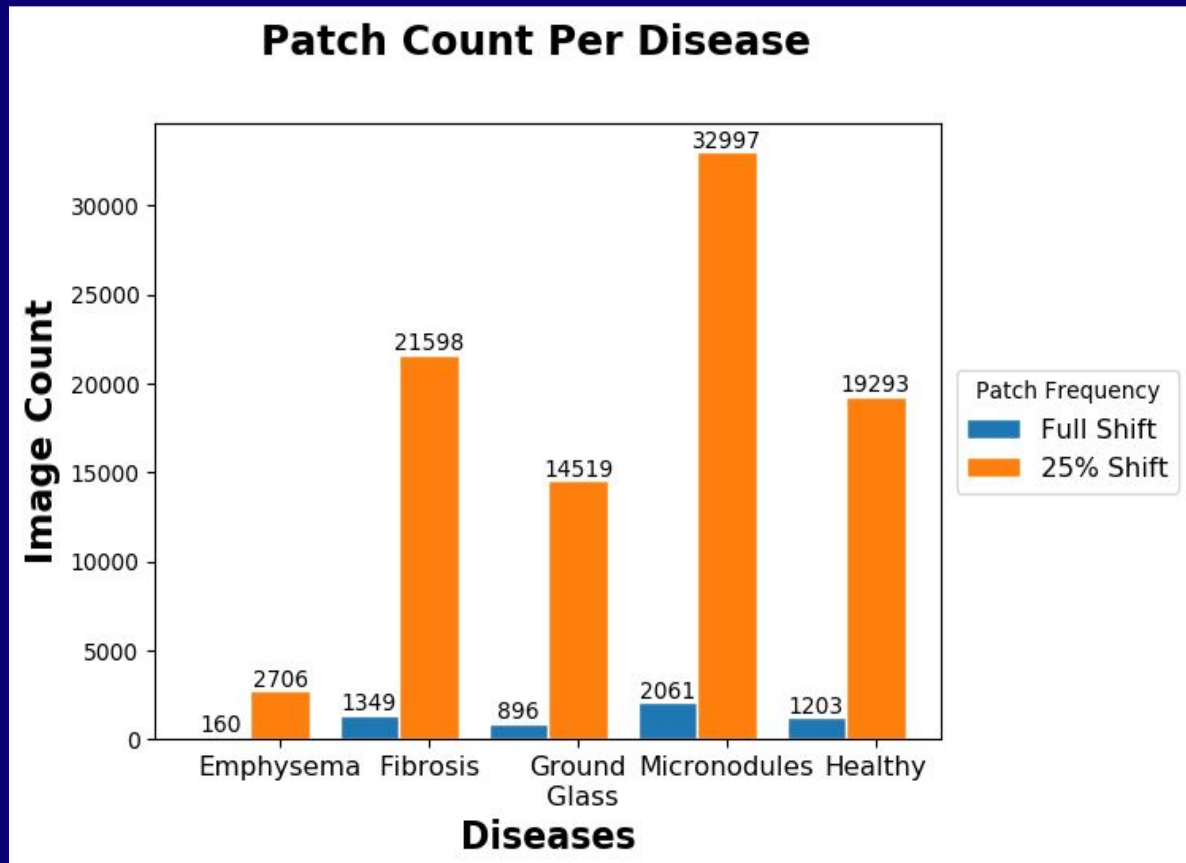
- Extract patches from CT scan that are from the ROI to avoid passing unnecessary CT info to the CNN.
- Convert the patches that are in single channel Hounsfield Unit to 3 channel red, green, blue png images.
- Use good naming conventions so that the patch names contain important information like disease label, CT series number, slice number, patch number (relative to the label), and patient number.
- Provide a framework with configurable values so that the user can get patches of a specific size, frequency, ROI concentration, and disease label.

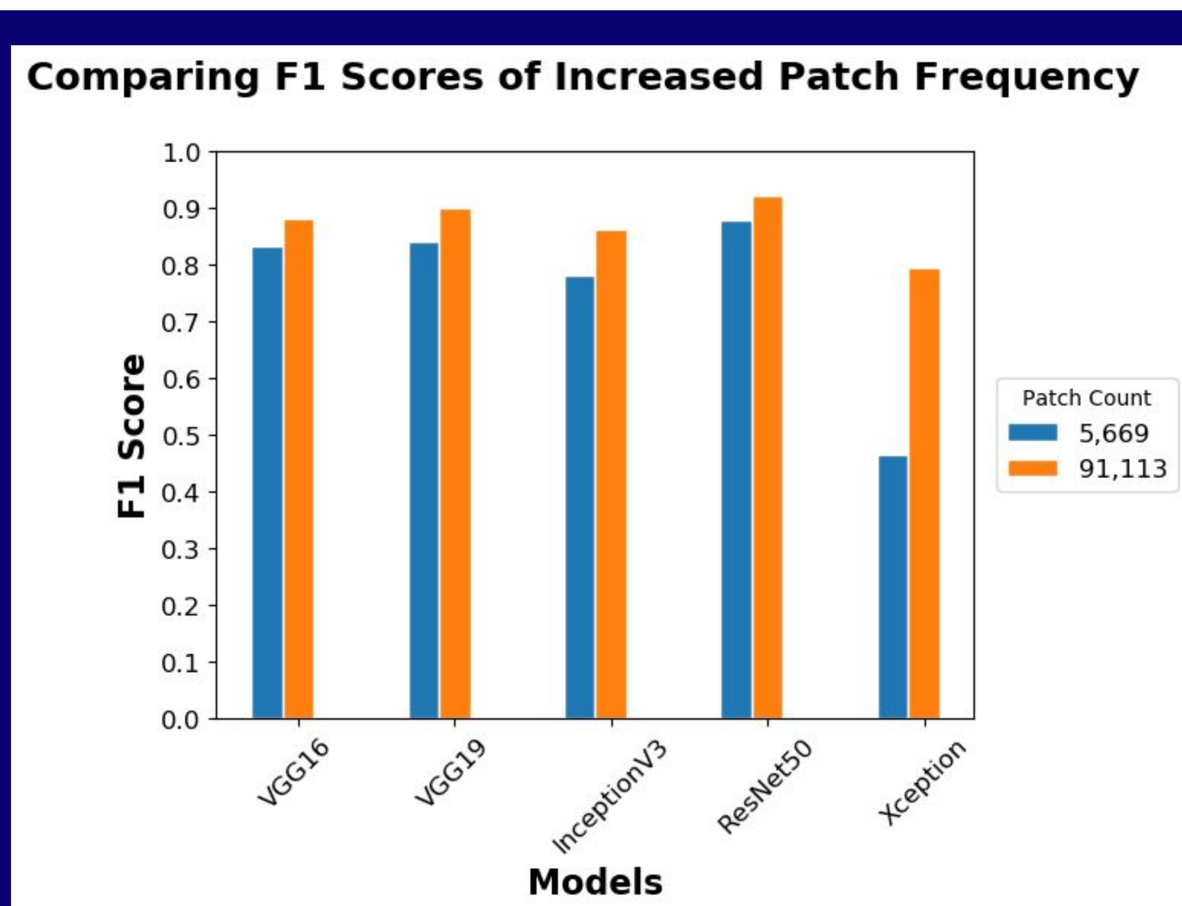
Why Use MedGIFT?

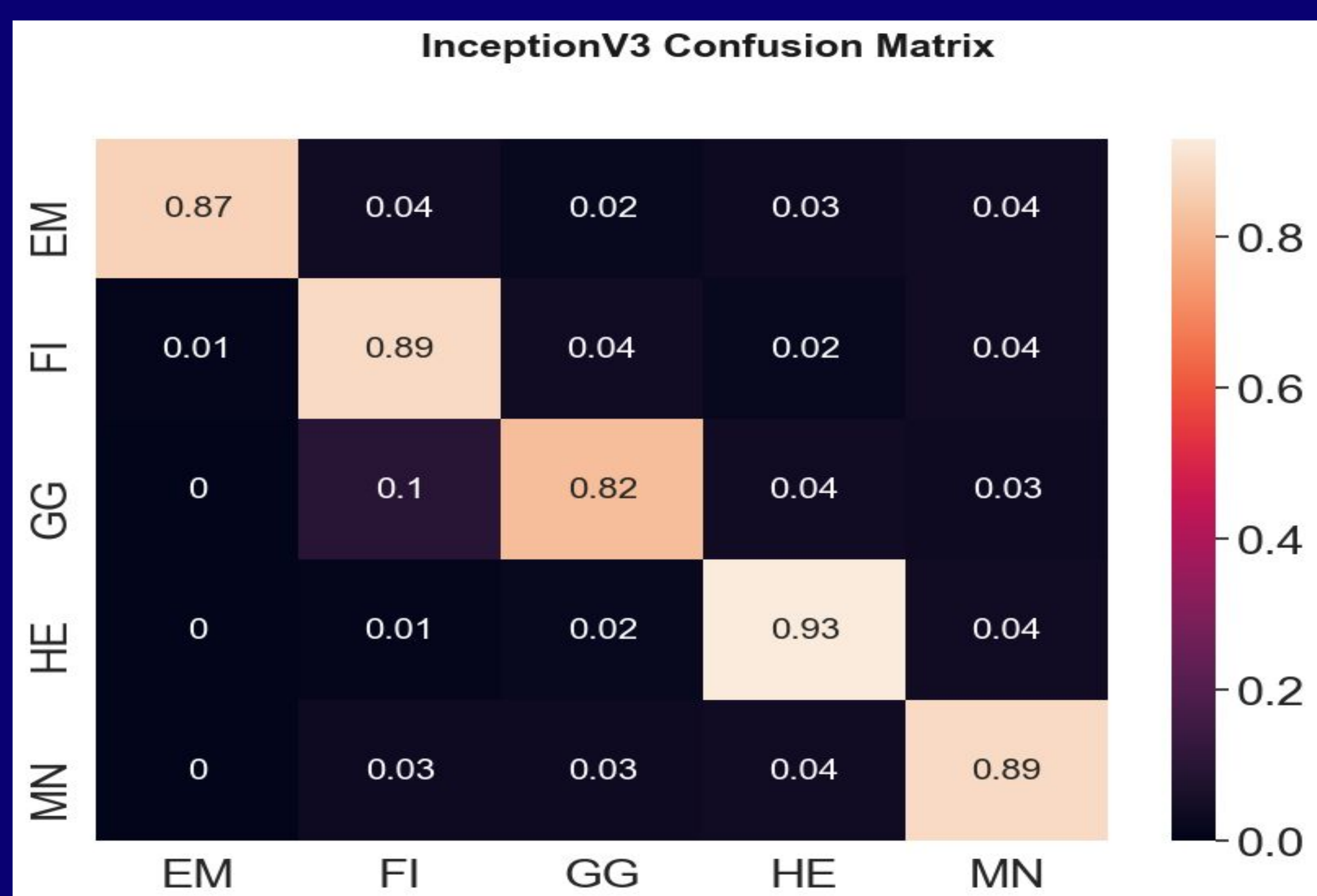
- The amount of diseases available means testing can be done to create a CNN framework that is more dynamic.
- Interacting with full CT scans is closer to a real world application.
- Using a full set of CT scans from a single patient opens up the door for the possibility of using multiple CT scans at once to gain some intuition of how a disease presents itself. This could lead to more accurate classifications of diseases from CT scans.



Patch Frequency: The patch extracting framework has premade settings for extracting patches at a given frequency. Patches can be extracted at 100% (full), 50%, and 25% increments. 25% increments yield the most images but this means patches will have a 75% overlap from the previous image.







Confusion Matrix Results

- Despite having substantially fewer images to pool from Emphysema was still more accurately predicted than Ground Glass.
- Ground Glass seems to be the hardest disease to predict for the inception network. This is probably because fibrosis appears to have similar patterns to fibrosis since that was choses the most in its place and vice versa.
- Healthy is the easiest to predict disease probably due to having the fewest similarities with the other diseases.

Future Work

- Take advantage of data augmentation techniques to increase the amount of data even further.
- Implement techniques such as ensembles, fine tuning, and multi-stage classification to further increase accuracy.
- Evaluate patches alongside other patches from a similar ROI in neighboring CT scans.
- Evaluate patches from regions of the CT scan at slices that are one level deeper and one level more shallow to achieve higher dimensionality.
- Create an original Convolutional Neural Network to exclusively train and test with MedGIFT data.
- Train and test with all of the diseases in the MedGIFT data set.
- Since Ground Glass and Fibrosis were less accurate despite having larger datasets some exploration on different HU to RGB conversions to differentiate the two classes could possibly be beneficial.

Conclusions

- Patches extracted using an even distribution of HU to RGB yielded the best results even above methods which were meant to highlight lung tissue.
- Larger datasets using 25% shifts proved to yield much better F1 scores when compared to a full shift data set.
- Class accuracy seems largely dependant on how different the class is compared to other classes being tested.
- While the larger dataset is more accurate it comes at a cost; running tests take substantially longer.