Tree Analysis

Even with a tiny parent and child node requirement of 2 members and 1 respectively, it required only 5 layers of nodes to reach 99.5% accuracy on the model. Even with only 4 nodes, it would have been over 95%.

Component analysis:

The SVD's Principle component 1 for covid and normal lungs were too similar to each other, leading to Principle component 2 being far more important as they were different enough to bring better splitting of the nodes.

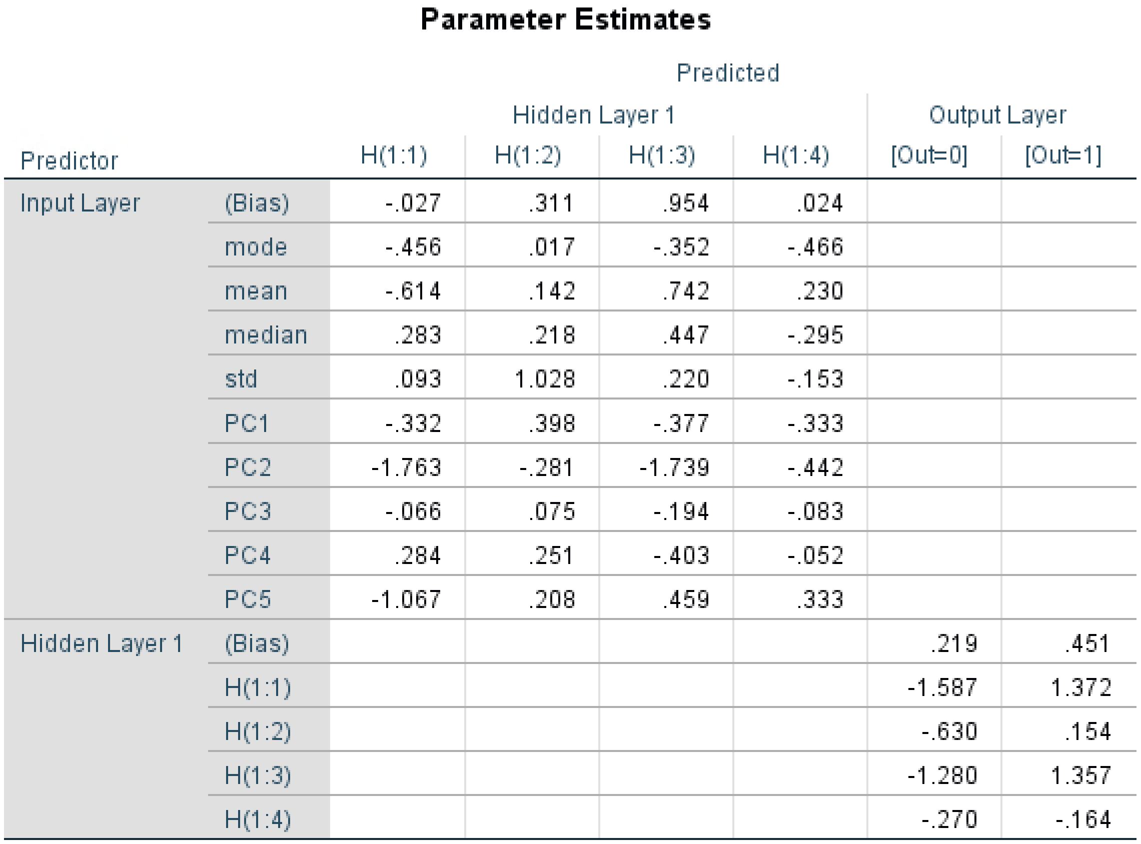
Then, the median and mean of the greyscale were both more important than PC1.

The model will likely fall in accuracy as more outlier images are formed, as only 1 of the 200 images had a x-ray that had the patient meaningfully off center. It's likely overfit to this data set for that reason.

Still, it shows that only 4 or 5 variables are necessary to diagnose covid vs normal lungs if no other diseases are present. The dataset would need to be much larger for other diseases.

Neural Network analysis

Also had a 95% accuracy, but with far less nodal layers than a tree analysis.



Parameter Estimates

Predictor		Predicted					
		Hidden Layer 1				Output Layer	
		H(1:1)	H(1:2)	H(1:3)	H(1:4)	[Out=0]	[Out=1]
Input Layer	(Bias)	-.027	.311	.954	.024		
	mode	-.456	.017	-.352	-.466		
	mean	-.614	.142	.742	.230		
	median	.283	.218	.447	-.295		
	std	.093	1.028	.220	-.153		
	PC1	-.332	.398	-.377	-.333		
	PC2	-1.763	-.281	-1.739	-.442		
	PC3	-.066	.075	-.194	-.083		
	PC4	.284	.251	-.403	-.052		
	PC5	-1.067	.208	.459	.333		
Hidden Layer 1	(Bias)					.219	.451
	H(1:1)					-1.587	1.372
	H(1:2)					-.630	.154
	H(1:3)					-1.280	1.357
	H(1:4)					-.270	-.164

A bias term needed to fix the hidden layer node 3 was very large, showing that node in particular needs more data to have the variables themselves show the proper diagnosis rather than a constant.

In the training, only 1 error came from missed diagnosis, while 0 from the testing section. In terms of a pandemic, false positives are much better than missed cases and having 9 false positives per missed case is still fine with a 95% accuracy.

Area under the curve for the specificity vs sensitivity is really high, showing it doesn’t take a lot to get high accuracy in true positives and negatives. That is good, then increasing the dataset appears to not be super huge due to the narrowness of the problem.