Building 4 Types of Classification Models: A Quick Research into predicting the Xray outcomes regarding Pneumonia

Image Classifcation using Google's Auto ML:A Study done by Udacity Research Institute

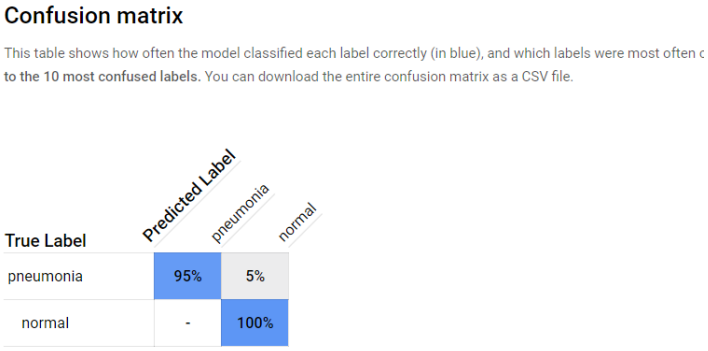
Study done by: Frederick Zoreta
** Other AutoML platforms being used were: Azure's Custom Vision and SentiSight.AI

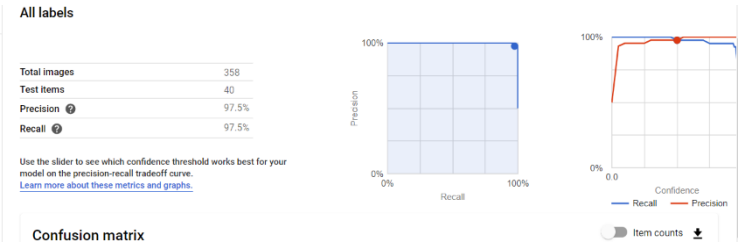
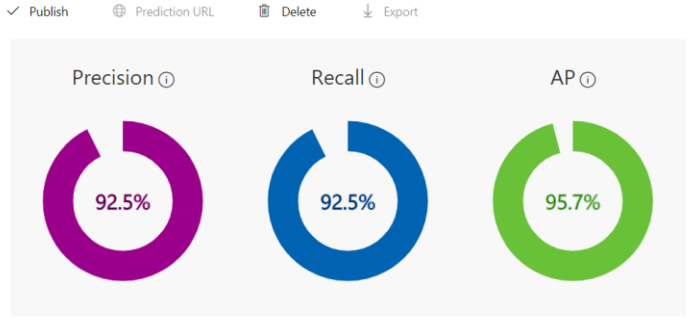
AutoML Modeling Report

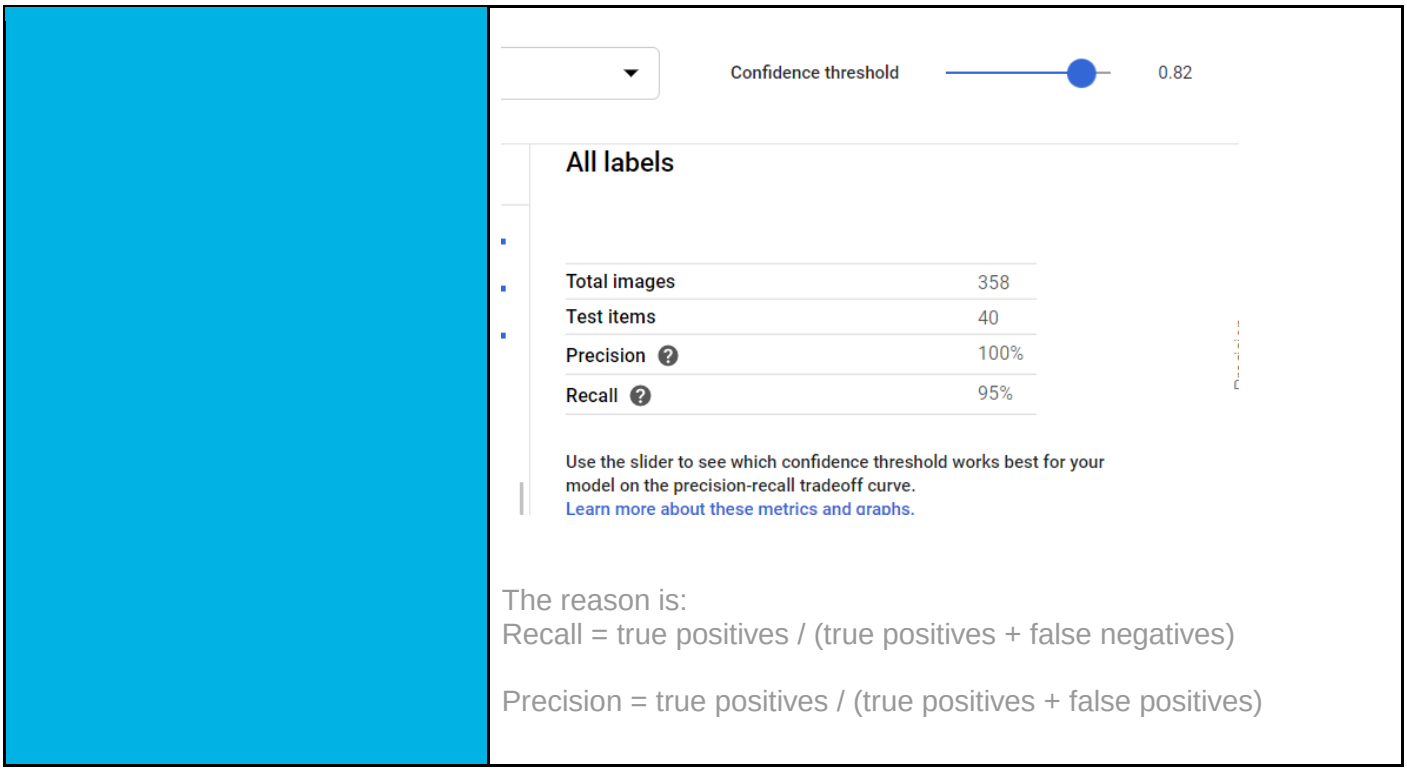
Researcher / Analyst : FREDERICK ZORETA

Binary Classifier with Clean/Balanced Data

**** I used both 200 images for normal and pneumonia in this class. Since it was given in the instructions, I could use between 100 to 300. As also discussed on the discussion boards.**

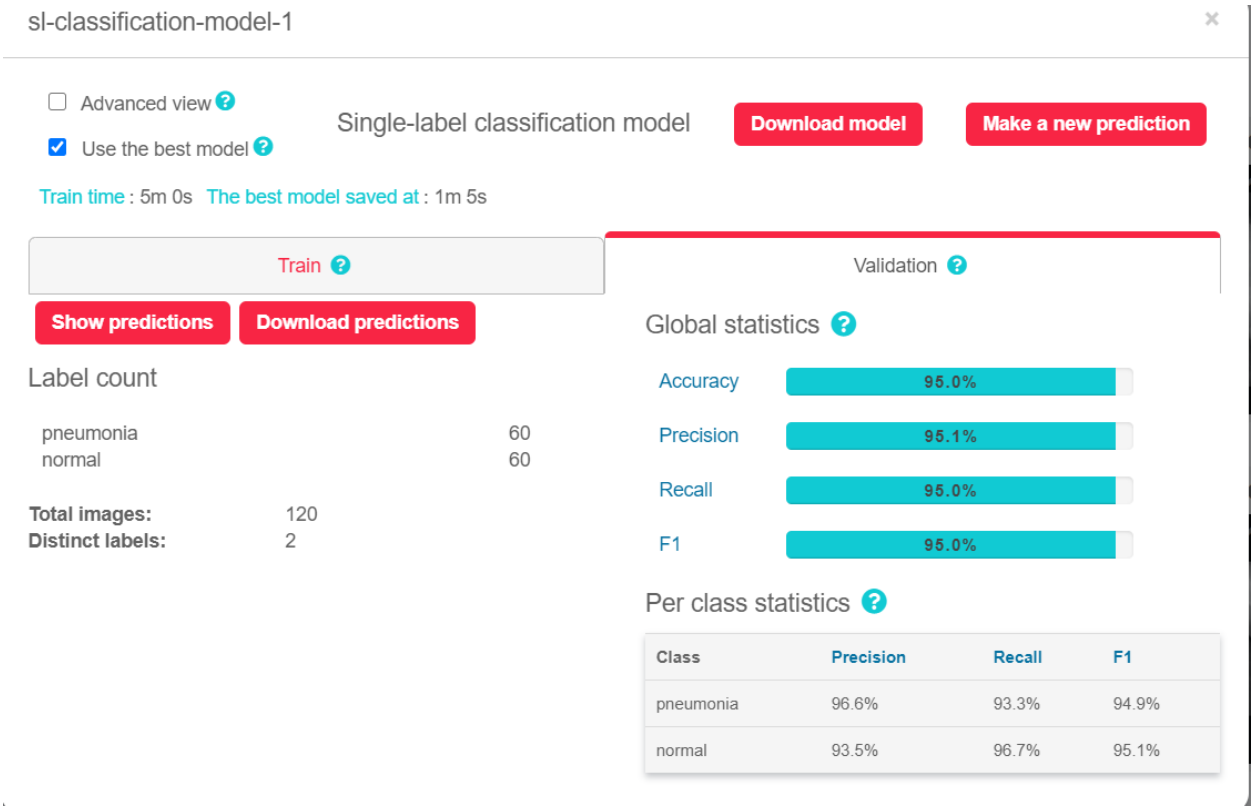
Train/Test Split How much data was used for training? How much data was used for testing?	**<Normal>** **For training: 160** **For testing: 20** **<Pneumonia>** **For training: 158** **For testing: 20** Data used: 200
Confusion Matrix What do each of the cells in the confusion matrix describe? What values did you observe (include a screenshot)? What is the true positive rate for the “pneumonia” class? What is the false positive rate for the “normal” class?	Predicted: Pneumonia: 95% labeled as ‘pneumonia’ 5% labeled as ‘normal’ Normal: 0% labaled as ‘pneumonia’ 100% labeled as ‘normal’  The true positives for Pneumonia is 95% and 0% false positive for normal
Precision and Recall	Using Google’s AutoML, both Precision and Recall was at 97.5%

What does precision measure? What does recall measure? What precision and recall did the model achieve (report the values for a score threshold of 0.5)?	with a threshold of 0.5 or 5%.  Below is the Precision & Recall using Azure's Custom Vision AI , also with a threshold of 0.5 or 50%  Performance Per Tag 
Score Threshold When you increase the threshold what happens to precision? What happens to recall? Why?	Increasing the threshold makes precision at 100% while recall is at 95% It means that Precision increased while recall decreased by a small margin



Additional Insights Regarding Clean Balanced Data

** Image below shows the analysis using Neurotechnology's SentiSight.AI
Set to 0.5 threshold



The precision and recall show a 95% accuracy on the validation set.

Below are 2 images that shows my model predicted the photos with a VERY HIGH %. It clearly shows that <CLEAN BALANCED> datasets results in higher % of accurate predictions.

Image 1 shows Pneumonia:

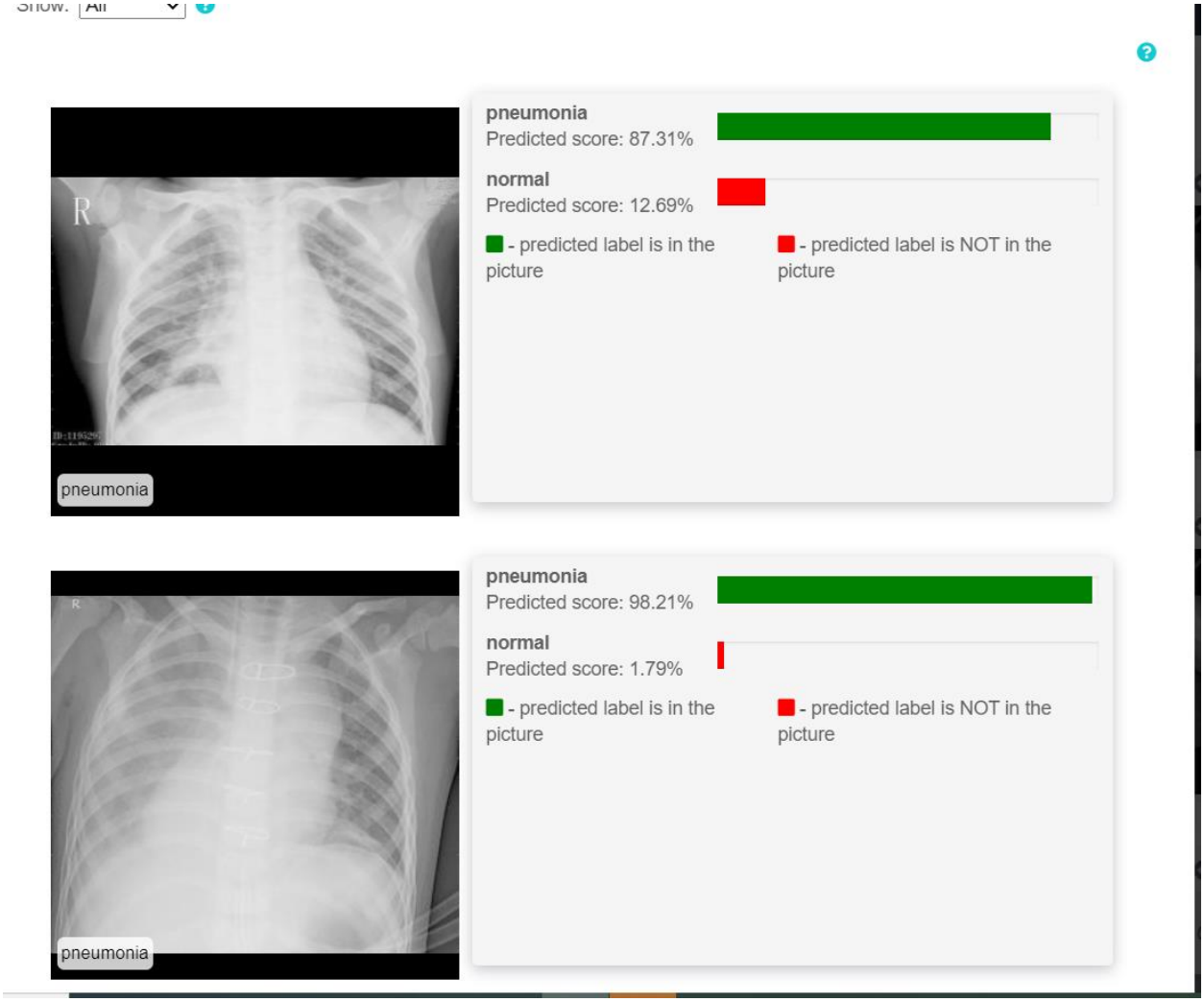
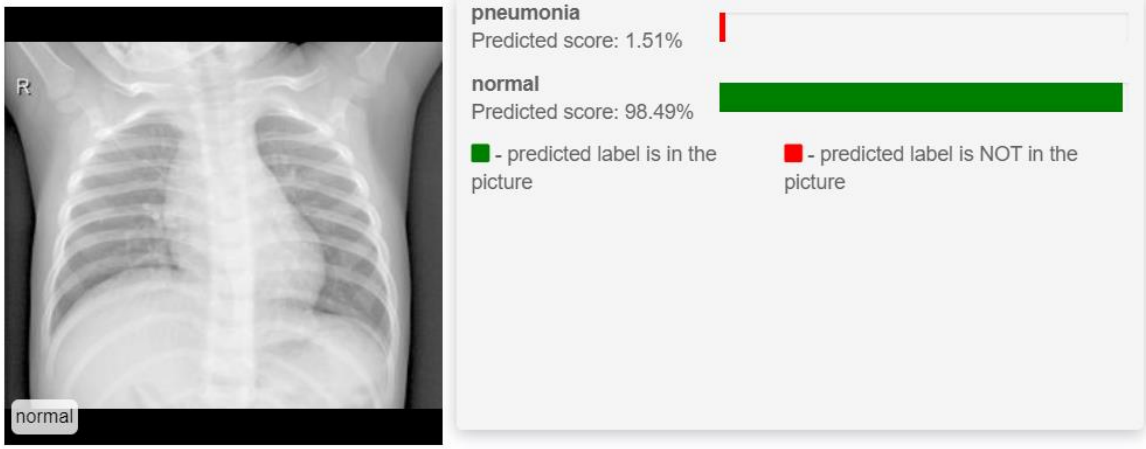
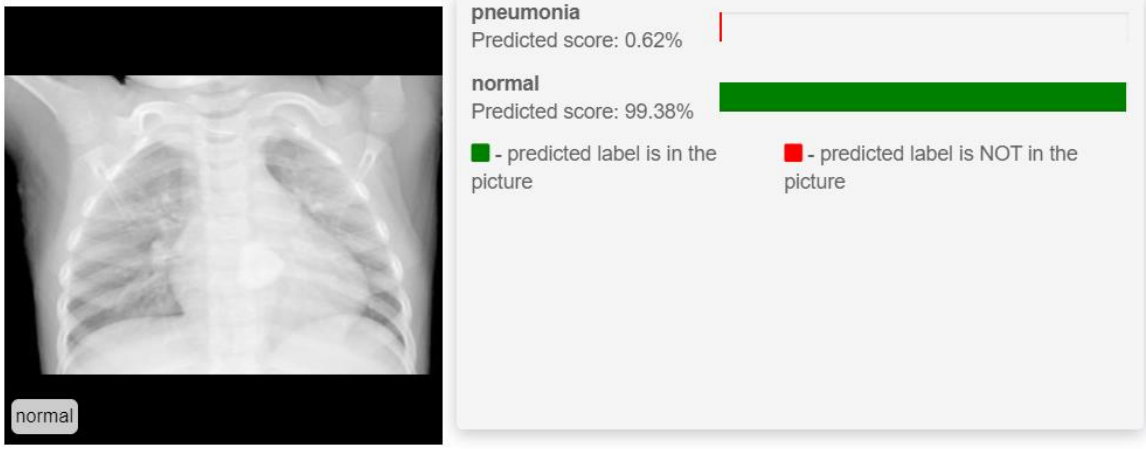


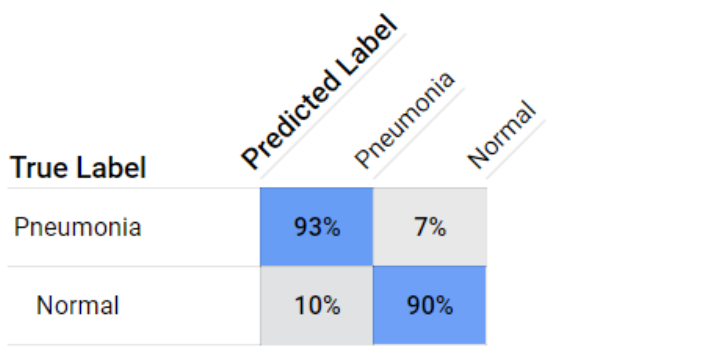
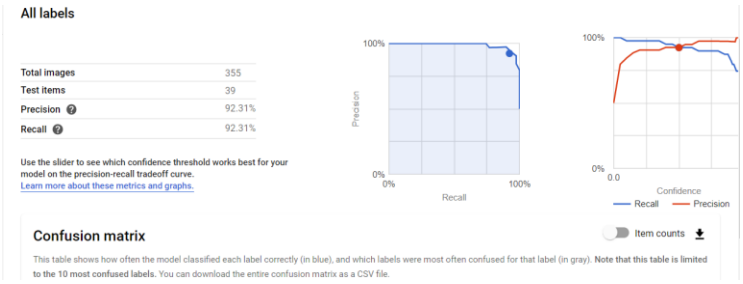
Image 2 shows a normal Xray

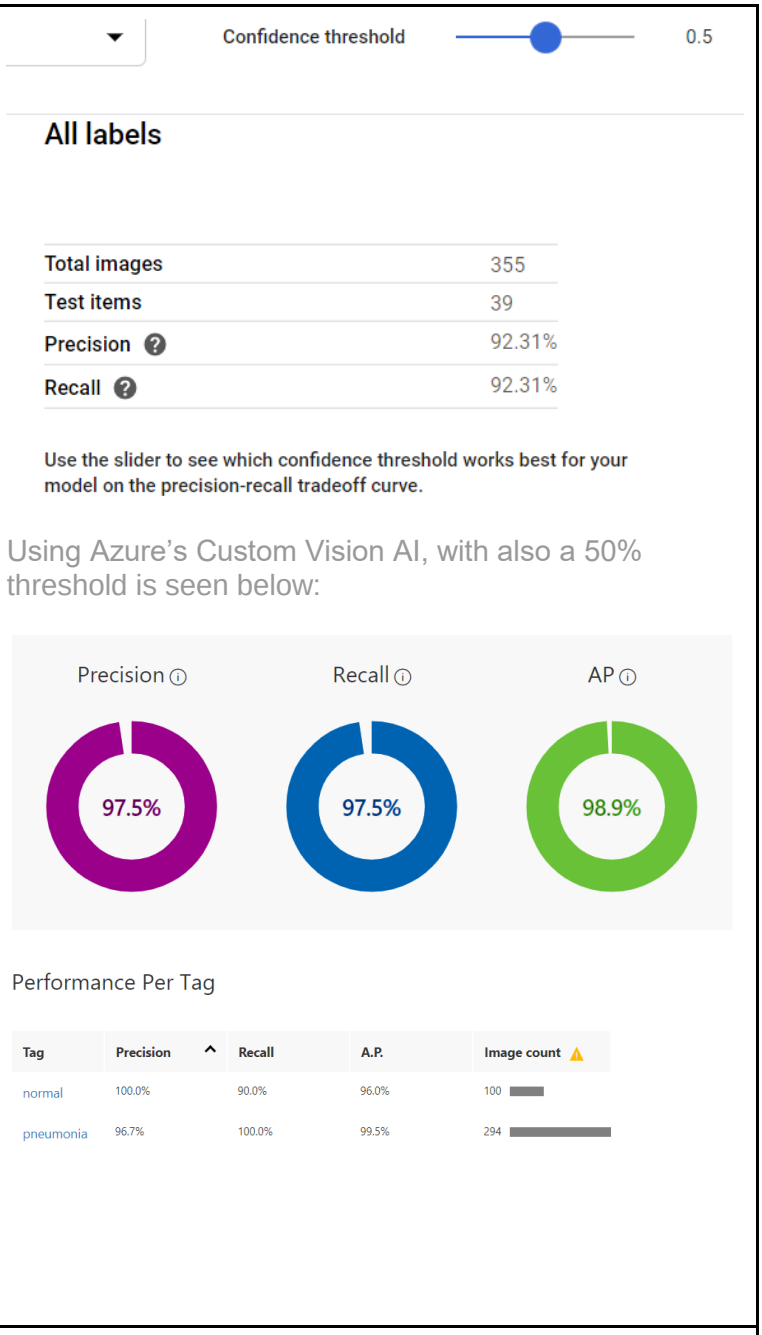




Binary Classifier with Clean/Unbalanced Data

Train/Test Split How much data was used for training? How much data was used for testing?	<Normal> Training: 80 Testing: 10 Validation: 10 <Pneumonia> Training: 235 Testing: 29 Validation: 30

	Data used: 400
Confusion Matrix How has the confusion matrix been affected by the unbalanced data? Include a screenshot of the new confusion matrix.	Based on the table below, it appears that there is very minimal errors. Pneumonia has 93% true positives and 90% true positives for Normal 
Precision and Recall How have the model's precision and recall been affected by the unbalanced data (report the values for a score threshold of 0.5)?	 The above image shows the 0.5 threshold of precision and recall using Google's AutoML A summary is below:



Unbalanced Classes
From what you have observed, how do unbalanced classed affect a machine learning model?

Unbalanced classed definitely 'throws off' the percentage predictions, although not by a huge amount. This could all be proven by looking at the graphs.

It is very useful to note that realistically, there will be a lot of unbalanced classes in real life.

Binary Classifier with Dirty/Balanced Data

Confusion Matrix
How has the confusion matrix been affected by the dirty data? Include a screenshot of the new confusion matrix.

Confusion matrix

This table shows how often the model classified each label correctly (i
to the 10 most confused labels. You can download the entire confusio

True Label \ Predicted Label	normal	pneumonia
normal	86%	14%
pneumonia	57%	43%

There is a huge increase in false positives, especially for pneumonia. It is astonishing to note that the predicted outcome for pneumonia is 57% FALSE POSITIVE. This is more than 50% error rate.

Clearly the 30% 'wrong photos', which we did on purpose have further confused the machine's algorithm.

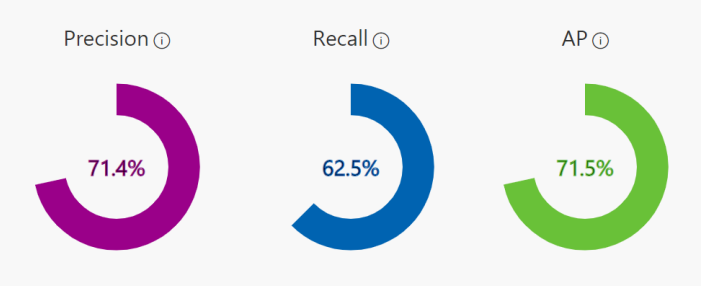
Precision and Recall
How have the model's precision and recall been affected by the dirty data (report the values for a score threshold of 0.5)? Of the binary classifiers, which has the highest precision? Which has the highest recall?

Confidence threshold 0.5

Total images	124
Test items	14
Precision	64.29%
Recall	64.29%

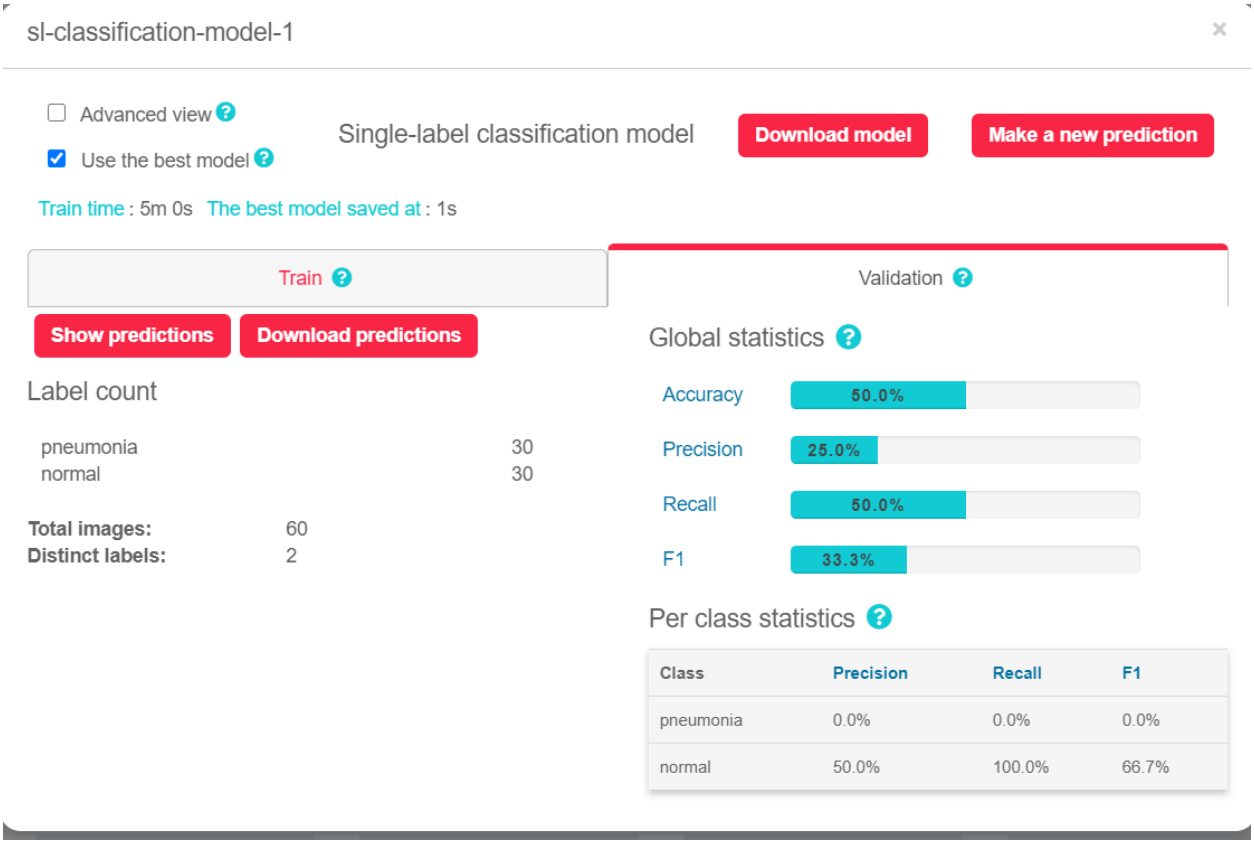
Use the slider to see which confidence threshold works best for your model on the precision-recall tradeoff curve.
Learn more about these metrics and graphs.

The highest precision went to the normal images. Despite both normal and pneumonia having each 30 wrong images, the 86% of correct prediction for normal

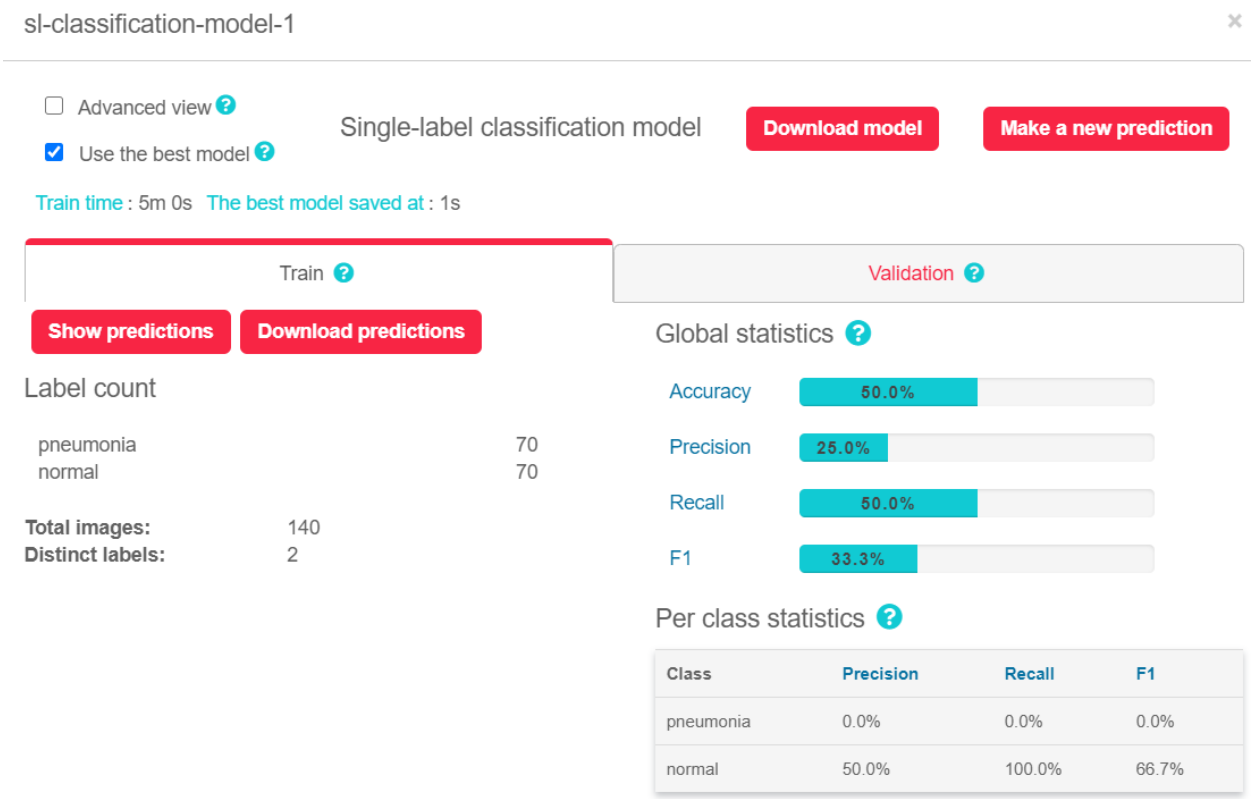
	is still a good score. On the other hand, the 43% correct prediction for pneumonia is way below the half mark. **Image below was derived using Azure's CustomVision.AI The precision % and the recall % is very near to that of Google's AutoML  Performance Per Tag Tag | Precision | Recall | A.P. | Image count pneumonia | 90.0% | 45.0% | 79.3% | 100 normal | 64.0% | 80.0% | 67.0% | 100
Dirty Data From what you have observed, how does dirty data affect a machine learning model?	Based on the findings, it definitely has a huge negative impact. The confusion matrix derived from Google's AutoML. The prediction % especially that of pneumonia is a huge factor. The lesson learned here is, data cleansing and data preparation is VERY VITAL.

Additional Insights regarding 'DIRTY DATA'.

**The image below was derived by using SentiSight.AI It shows the training was split into 30 pneumonia and 30 normal.

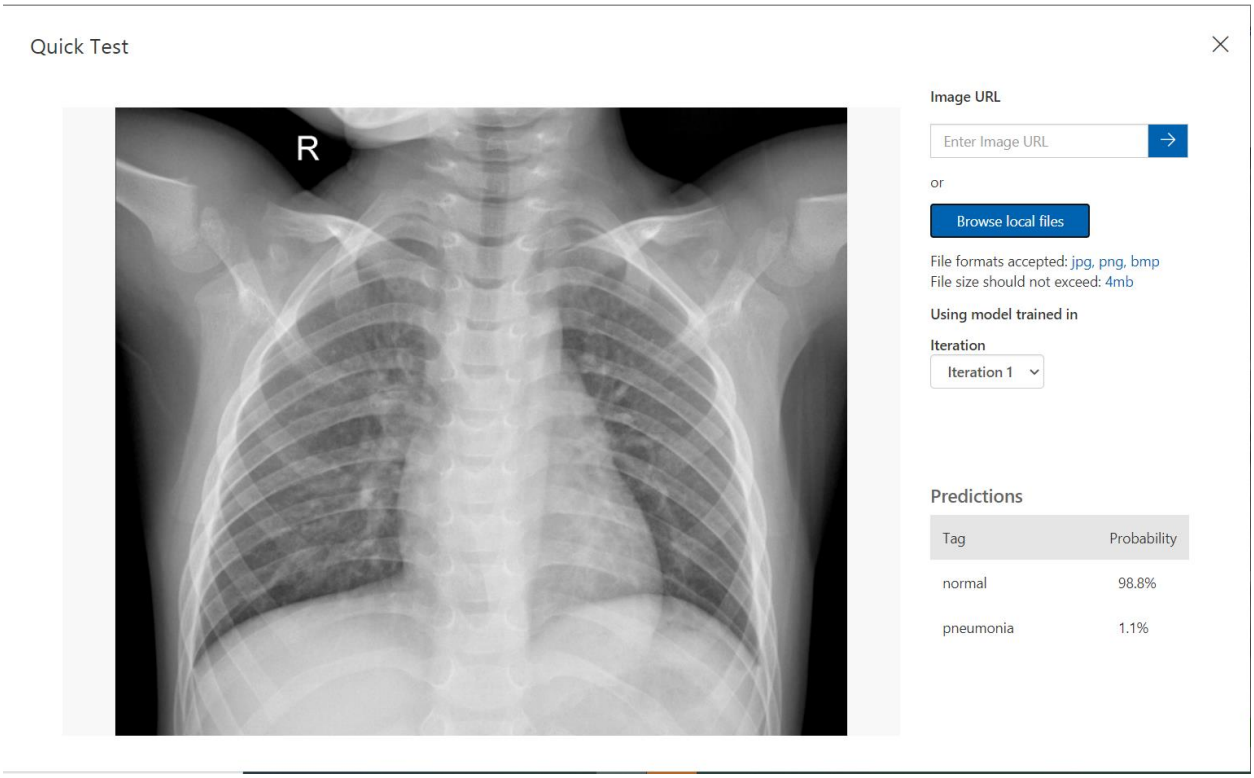


** The image below shows the validated 70 correct pneumonia and 70 correct normal images. This was produced using SentiSight.AI



**Another insight about dirty data, is that despite having a low accuracy, it still can predict correct results. As shown below using Azure's Custom Vision AI. The first 2 images were correct. The 3rd was taken from the dirty file, the folder was labeled 'pneumonia' but the actual photo was a normal Xray.

Correctly predicted Pneumonia -> Image above

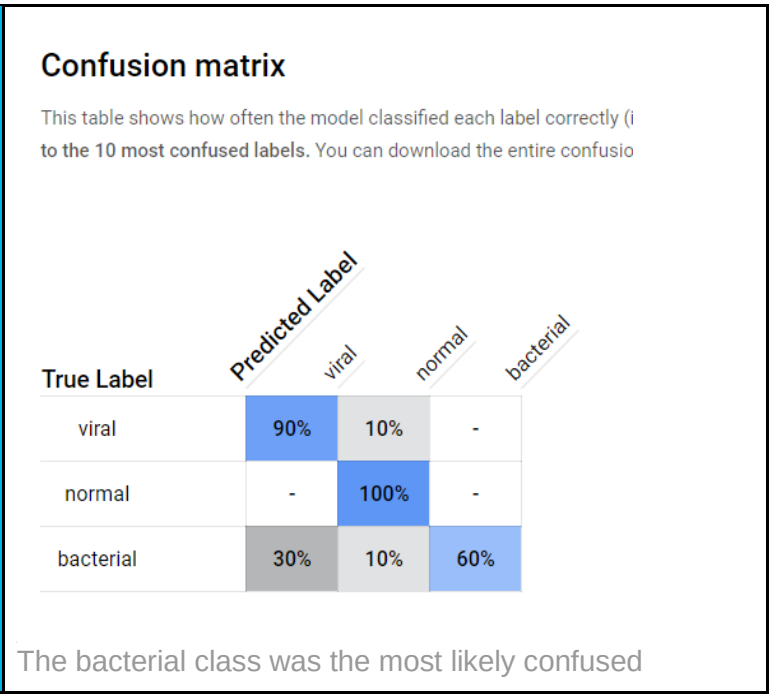


Correctly predicted Normal -> Image above

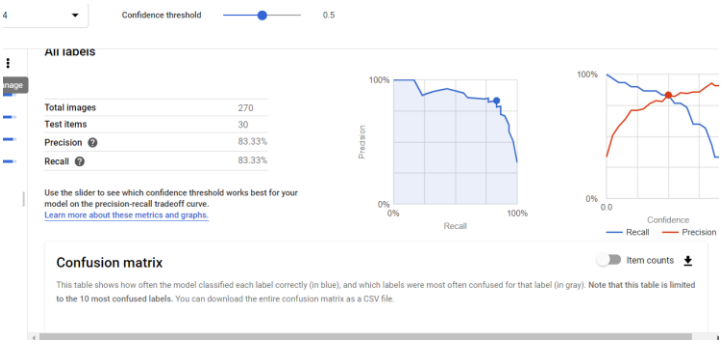
Image above is a normal xray taken from one of the dirty data folders. This is an example of *'dirty data messing up the prediction'*** scenario.

3-Class Model

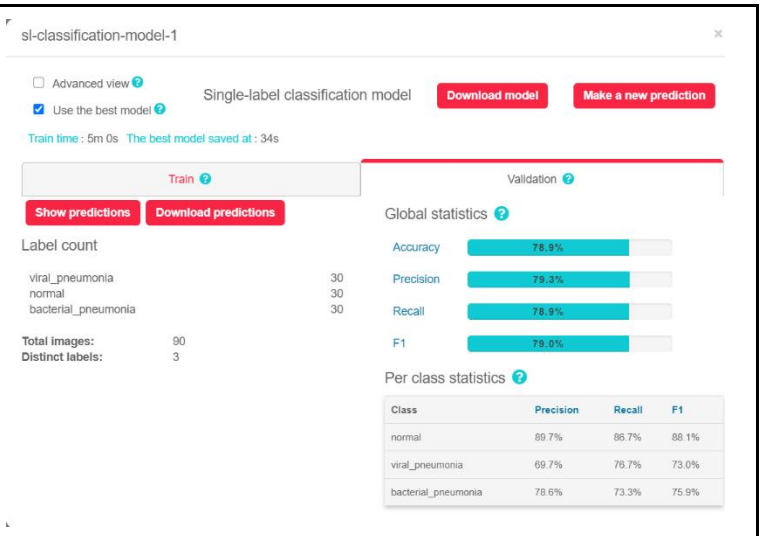
Confusion Matrix
Summarize the 3-class confusion matrix. Which classes is the model most likely to confuse? Which class(es) is the model most likely to get right? Why might you do to try to remedy the model's "confusion"? Include a screenshot of the new confusion matrix.



The bacterial class was the most likely confused

	prediction. Only 60% of it was predicted accurately. The normal was predicted with 100% accuracy. As far as the above data is concerned, Viral and Normal are actually good. As far as this model is concerned, I would look into the image quality of the bacterial Xrays. Since the % for both normal and viral is very high, I see no reason to further tweak the model.
Precision and Recall What are the model's precision and recall? How are these values calculated (report the values for a score threshold of 0.5)?	 Caluclated using: Precision = True Positives / True Positives + False Positives Precision = (100 + 60 + 90) -> TP --- (100 + 60 + 90) -> TP + (30 + 10 + 10) -> FP Therefore, Precision = .83 or 83% Recall = = (100 + 60 + 90) -> TP --- (100 + 60 + 90) -> TP + (30 + 10) -> FN Recall = .85 or 85%

F1 Score
What is this model's F1 score?



The over-all F1 Score is 79% using Sentisight.AI

Below is a computation of the F1 Score using the Google AutoML

** The higher the F Score, the better the predictive power.

$$F = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}$$

$$= 2 \times \frac{(6{,}944)}{167}$$

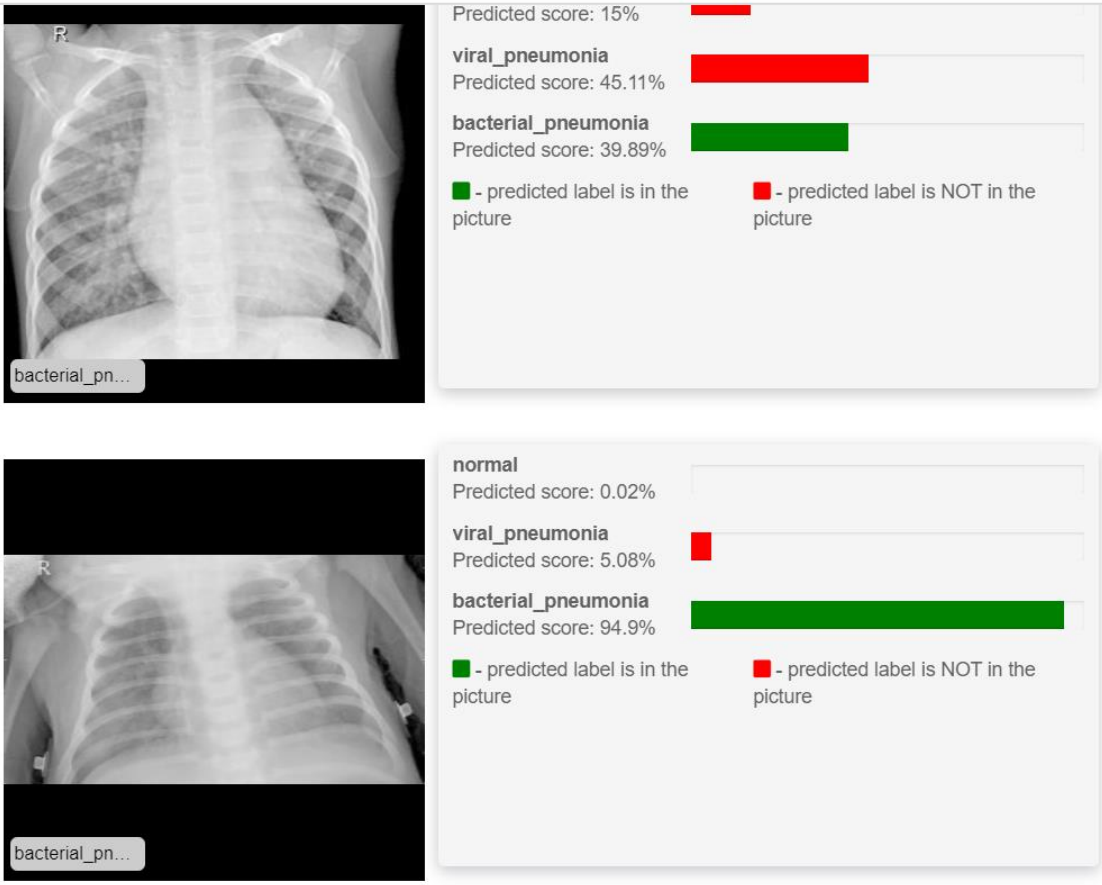
$$= 41.58 \times 2,$$

Therefore F Score = 83.16

Additional Insights Regarding 3 Class

Below are 3 images from each of the 2 of the 3 platforms used that showed correct predictions via testing different images from my desktop
*NOTE: Google AutoML now charges money for testing models

SentiSight.AI



Azure CustomVision.AI

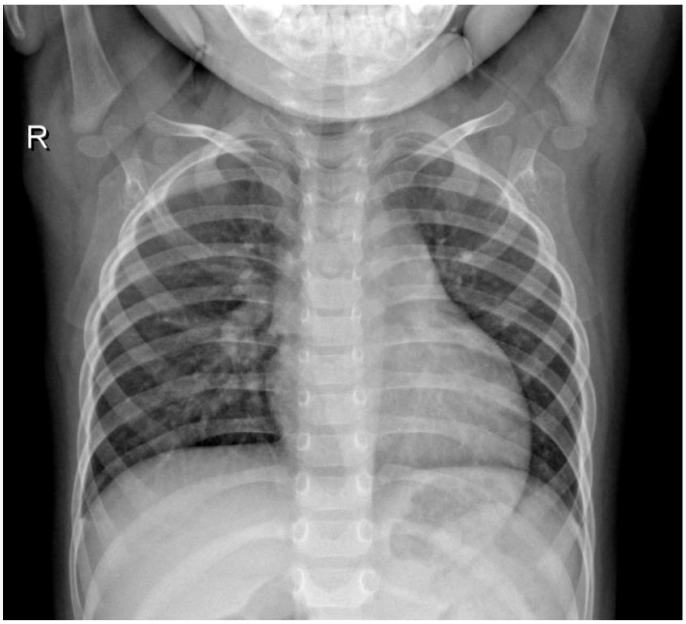


Image URL

Enter Image URL

or

Browse local files

File formats accepted: jpg, png, bmp
File size should not exceed: 4mb

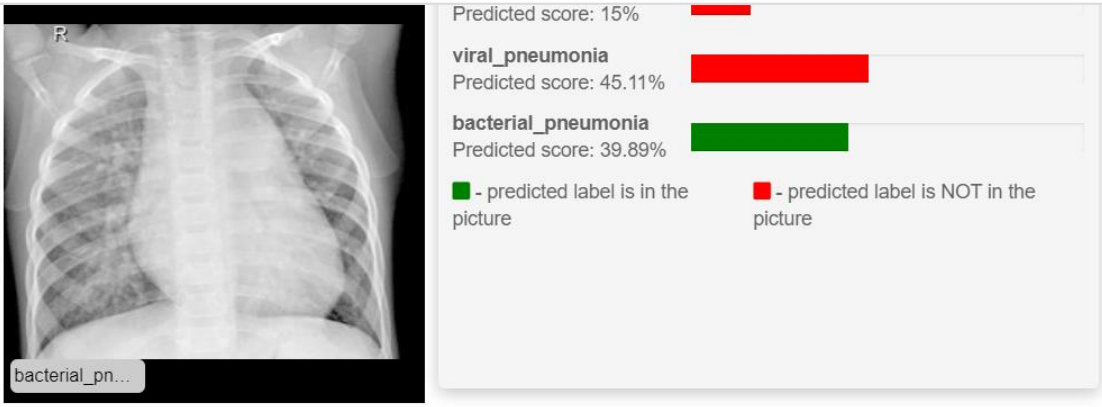
Using model trained in

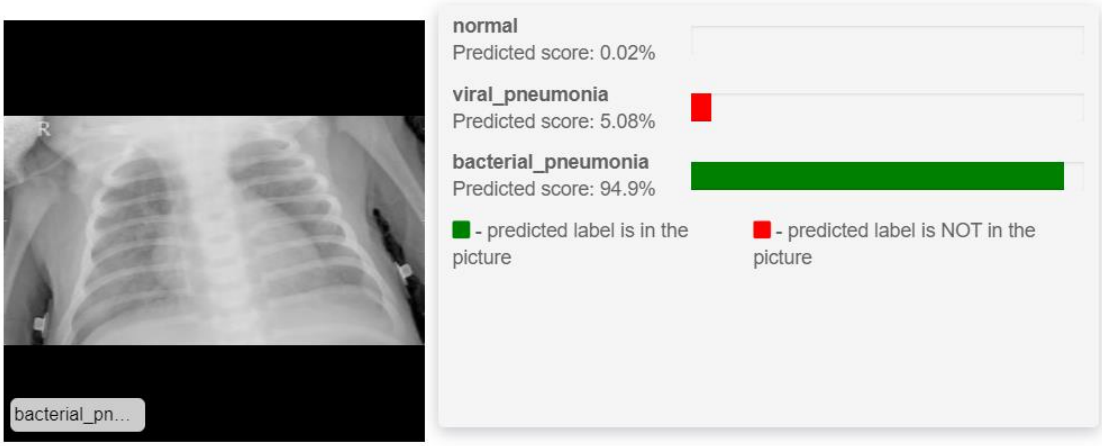
Iteration

Iteration 1

Predictions

Tag	Probability
normal	99.9%
virus	0%
bacteria	0%





Over all summary:

One of the reasons I have used 3 platforms is to have a comparison. Although the results are almost identical, I have noticed that SentiSight.AI was the fastest in processing the models. It was also the only platform that never missed some photos to be uploaded.

I had to repeat the process of simply uploading the photos using Google's AutoML and Azure's Custom Vision AI.

Another important factor is to always label the classes as accurately and as appropriately as possible. For a simple research like this, it would be easy to detect what the errors are. In an ultra large-scale project, with thousands of images and even videos involved, it would be extremely difficult.

Finally, related to the 3 class model, certain xray images may have a very high striking similarity. On higher level research that would involve thousands of images or videos, scenarios that pertain to 'bacterial' and 'viral' may indeed cause some errors.

Appendix :

** Platforms being used:
a.Google Cloud’s AutoML Vision
b. Azure’s Custom Vision AI
c. Neurotechnology’s SentiSight.AI

Towards Data Science: Beyond Accuracy & Precision Recall :

https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c#:~:text=As%20the%20threshold%20decreases%2C%20the,we%20increase%20the%20false%20positives

Analytics Vihdya – Confusion Matrix

https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/

Accuracy, Precision, Recall & F1 Score
https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/

Towards Data Science: Understanding Confusion Matrix
https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62