AutoML Modeling Report

Sukanya Rammohan

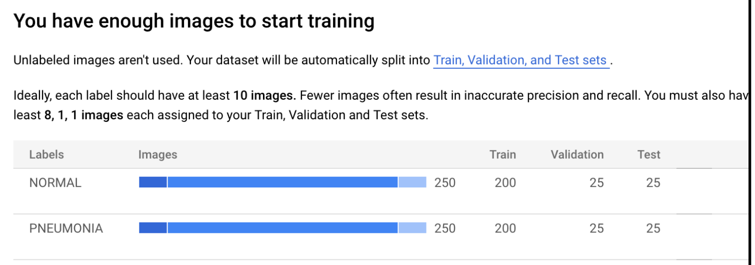
Binary Classifier with Clean/Balanced Data

Train/Test Split
How much data was used for training? How much data was used for testing?

Total data used: 500

Pneumonia: 250
Normal: 250

Following image gives Train validation and Test values.



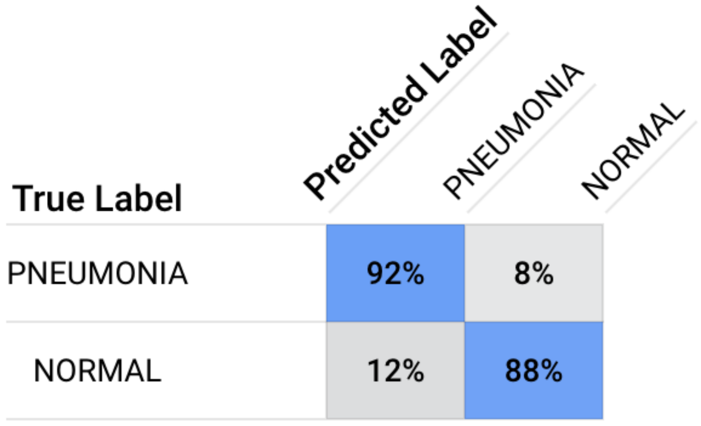
You have enough images to start training

Unlabeled images aren't used. Your dataset will be automatically split into Train, Validation, and Test sets .

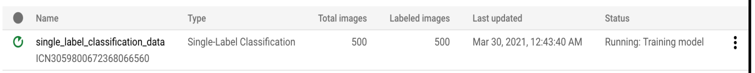
Ideally, each label should have at least **10 images.** Fewer images often result in inaccurate precision and recall. You must also hav least **8, 1, 1 images** each assigned to your Train, Validation and Test sets.

Labels	Images	Train	Validation	Test
NORMAL	250	200	25	25
PNEUMONIA	250	200	25	25

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?



True Label	Predicted Label PNEUMONIA	NORMAL
PNEUMONIA	92%	8%
NORMAL	12%	88%

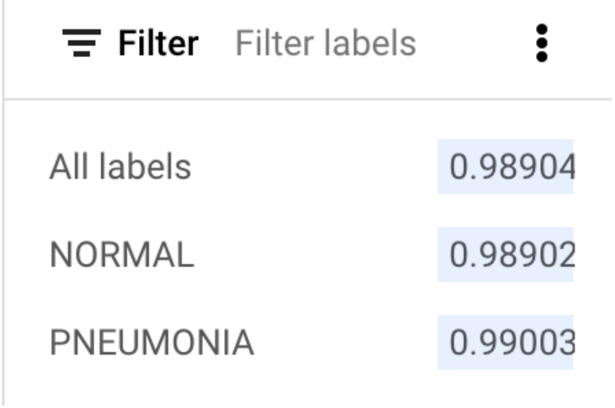
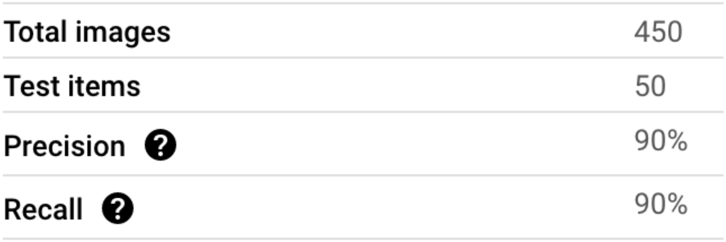


●	Name	Type	Total images	Labeled images	Last updated	Status
	single_label_classification_data ICN3059800672368066560	Single-Label Classification	500	500	Mar 30, 2021, 12:43:40 AM	Running: Training model

The confusion matrix is shown in the above image.

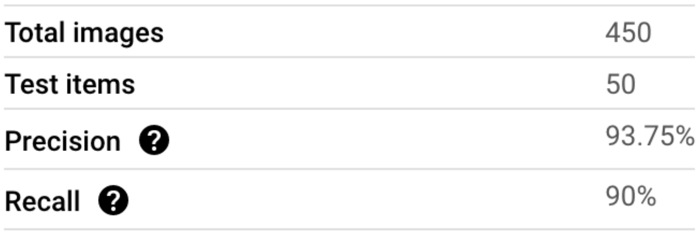
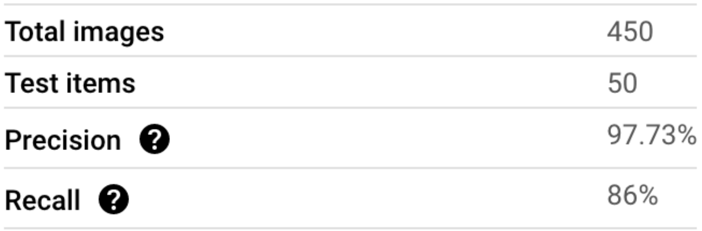
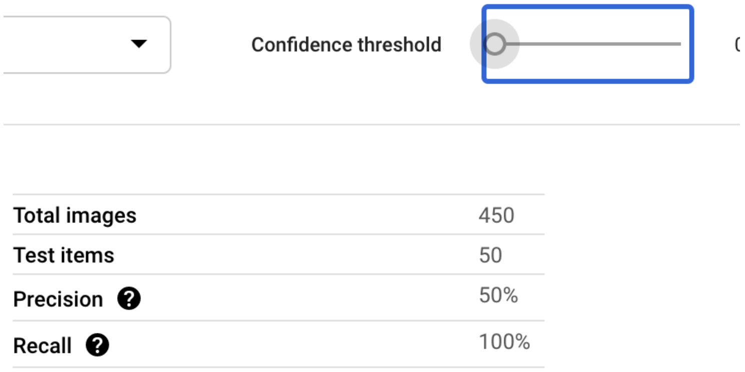
<table>
<tr>
<td></td>
<td>The values shown depict accuracy of the predictions that are done. It shows the degree to which the model is flawed.

The true positives from the matrix is 92%, which indicates that the model classified the true pneumonia affected cases as affected. The true positives constitute 92% for pneumonia affected cases.

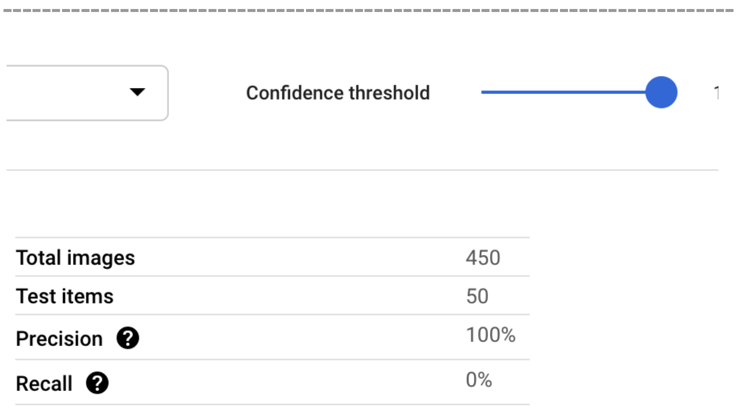
It also wrongly classified 8% of the affected cases as normal. It also miscategorized 12% of the normal cases as affected, which is the false positive percentage of the normal class. Presence of a high percentage of true positives indicates the better performance of the mode.</td>
</tr>
<tr>
<td><b>Precision and Recall</b>
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)?</td>
<td>The image below shows the Precision and Recall for score threshold of 0.5

Filter Filter labels
All labels 0.98904
NORMAL 0.98902
PNEUMONIA 0.99003

Total images 450
Test items 50
Precision 90%
Recall 90%

Precision is the measure of the model’s ability to predict accurately without flaw. Recall measures the model’s ability to predict without regards to accuracy. In recall, the accuracy is not taken into consideration.

The image above indicates that the model has 90%</td>
</tr>
</table>

	precision and recall, which indicates that the model makes more predictions that are also accurate.
Score Threshold When you increase the threshold what happens to precision? What happens to recall? Why? 1. **0.60---------------->** 2. **0.75----------------->**	 Total images 450 Test items 50 Precision ? 93.75% Recall ? 90%  Total images 450 Test items 50 Precision ? 97.73% Recall ? 86% When the confidence threshold is increased, the value of precision increased and the value of recall declined considerably. When the confidence threshold was at 0.60, the precision was set at 93.75% and it increased to 97.73% when the confidence threshold was increased to 0.75 On the contrary, when the confidence threshold was increased from 0.60 to 0.75, recall declined from 90% to 86%. This is a clear indication that the precision is proportional to the threshold while the recall and the confidence threshold were negatively correlated.
Threshold **Confidence threshold : 0**	 Total images 450 Test items 50 Precision ? 50% Recall ? 100%

Confidence threshold : 1



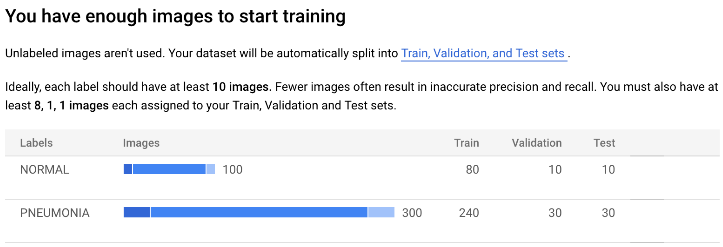
Total images	450
Test items	50
Precision ?	100%
Recall ?	0%

Increasing the threshold from 0 to 1 shows the increase of precision while the recall quickly declines. At the confidence threshold of 0, precision was 50% while recall was 100%.

On the other hand, when the confidence threshold was at 1, the precision was at its peak at 100% while recall was 0%

Binary Classifier with Clean/Unbalanced Data

Train/Test Split
How much data was used for training? How much data was used for testing?



You have enough images to start training

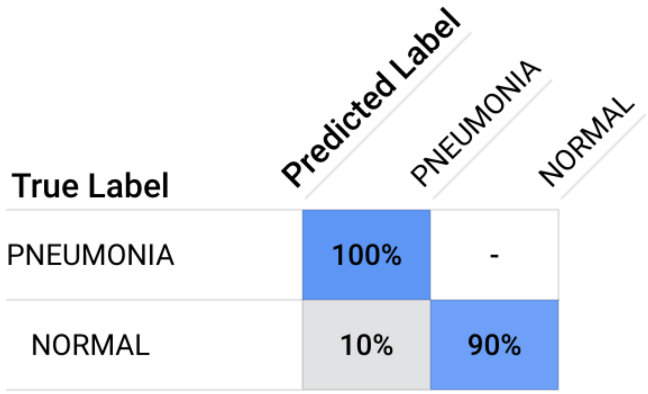
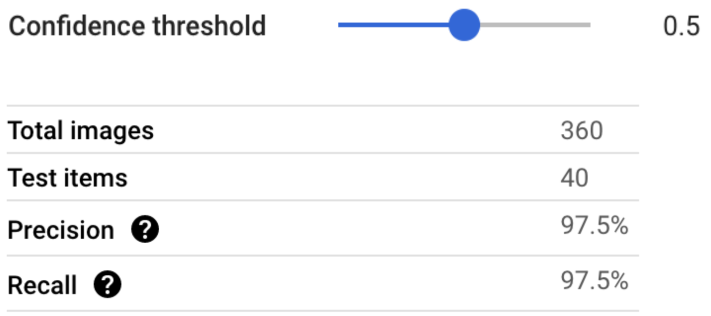
Unlabeled images aren't used. Your dataset will be automatically split into Train, Validation, and Test sets .

Ideally, each label should have at least **10 images.** Fewer images often result in inaccurate precision and recall. You must also have at least **8, 1, 1 images** each assigned to your Train, Validation and Test sets.

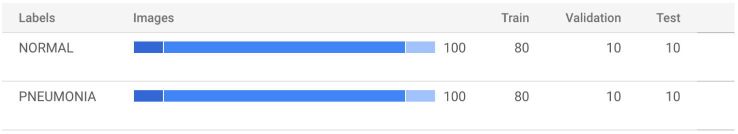
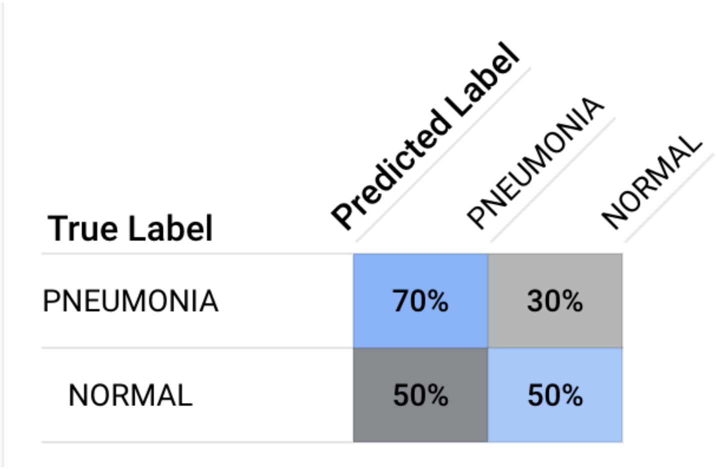
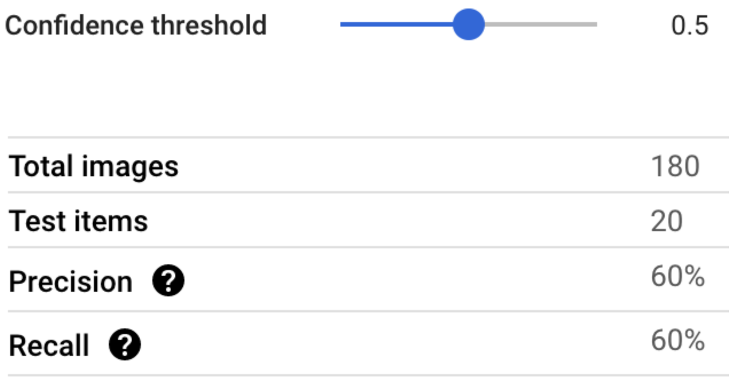
Labels	Images	Train	Validation	Test
NORMAL	100	80	10	10
PNEUMONIA	300	240	30	30

Data set includes a total of 400 images including an unbalanced set of 100 data points for normal ases and 300 cases of pneumonia indicating X-ray images.

For testing 10 data points were chosen from normal X-ray images and 30 from pneumonia affected dataset.

<table>
<tr>
<td><b>Confusion Matrix</b>
How has the confusion matrix been affected by the unbalanced data? Include a screenshot of the new confusion matrix.</td>
<td>
<table>
<tr><th>True Label</th><th>Predicted Label
PNEUMONIA</th><th>NORMAL</th></tr>
<tr><td>PNEUMONIA</td><td>100%</td><td>-</td></tr>
<tr><td>NORMAL</td><td>10%</td><td>90%</td></tr>
</table>
Due to the unbalanced dataset, having more pneumonia data set, the precision of pneumonia indicating data has reached a perfect score of 100%. The unbalanced dataset has become efficient in predicting pneumonia than the normal cases.</td>
</tr>
<tr>
<td><b>Precision and Recall</b>
How have the model’s precision and recall been affected by the unbalanced data (report the values for a score threshold of 0.5)?</td>
<td>
Confidence threshold 0.5
<table>
<tr><td>Total images</td><td>360</td></tr>
<tr><td>Test items</td><td>40</td></tr>
<tr><td>Precision ?</td><td>97.5%</td></tr>
<tr><td>Recall ?</td><td>97.5%</td></tr>
</table>
The improvement in precision at a confidence threshold of 0.5 seems to perform better. Unfortunately, this improvement is unreliable due to the unbalanced data sets.</td>
</tr>
<tr>
<td><b>Unbalanced Classes</b>
From what you have observed, how do unbalanced classed affect a machine learning model?</td>
<td>Unbalanced classes affect the machine learning model in a disguise of improving efficiency of the model. But in reality, the model’s precision is adversely affected making it unreliable and the efficiency is at stake.</td>
</tr>
</table>

Binary Classifier with Dirty/Balanced Data

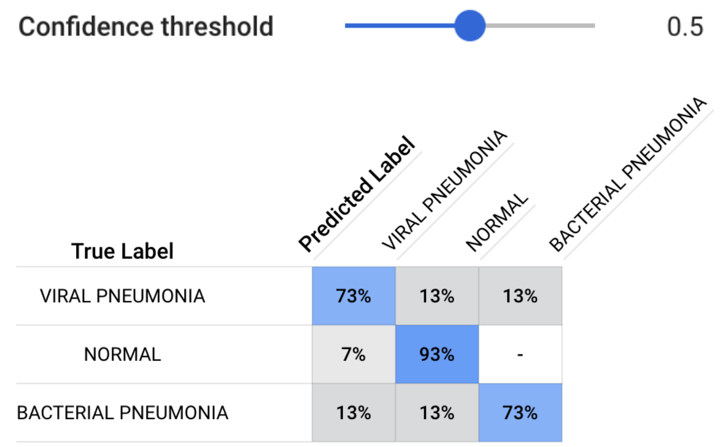
Confusion Matrix How has the confusion matrix been affected by the dirty data? Include a screenshot of the new confusion matrix.	 Labels · Images · Train · Validation · Test NORMAL · 100 · 80 · 10 · 10 PNEUMONIA · 100 · 80 · 10 · 10  True Label · Predicted Label · PNEUMONIA · NORMAL PNEUMONIA · 70% · 30% NORMAL · 50% · 50% Using dirty data which contains 30% of mixed data labels, the score of true positives has reduced by about 20% compared to the previous clean data models. This has also increased the mis-categorization to about 50%.
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 180 Test items 20 Precision 60% Recall 60% From observation, it is found that the precision and recall declined considerably given the dirty data. Both precision and recall have decreased from 97.5% to 60%. A decrease in precision and recall indicates the decline in accuracy and results.
Dirty Data From what you have observed, how does dirty data affect a	Dirty data, which is characterized by data points with some incorrect labeling has the ability to negatively impact precision and recall.

machine learning model?

From the confusion matrix, it is evident that there is significant increase in false positives and significant decrease in true positives. Hence, the performance of the machine learning model is unreliable.

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.



Confidence threshold 0.5

True Label	Predicted Label VIRAL PNEUMONIA	NORMAL	BACTERIAL PNEUMONIA
VIRAL PNEUMONIA	73%	13%	13%
NORMAL	7%	93%	-
BACTERIAL PNEUMONIA	13%	13%	73%

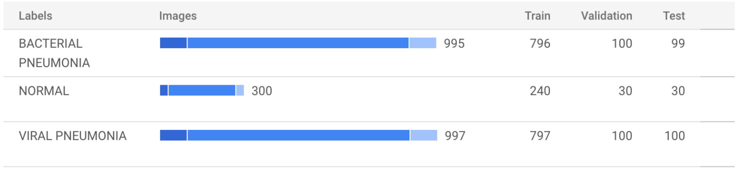
The model with 150 data points of each category viz bacterial pneumonia, viral pneumonia and normal resulted in almost equal levels of true positives.

New Improved Data:

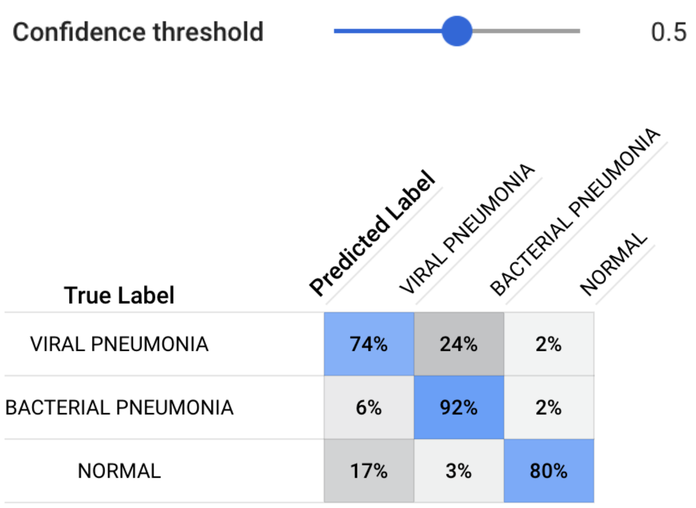
To reduce the models confusion, adding more data points to the pneumonia label resulted in better performance. 1000 data points for bacterial pneumonia, 1000 data points of viral pneumonia had some issues where some of the data produced error while loading.

Hence the final data point was as follows:

- Bacterial pneumonia count = 796
- Viral pneumonia count = 997
- Normal count = 300

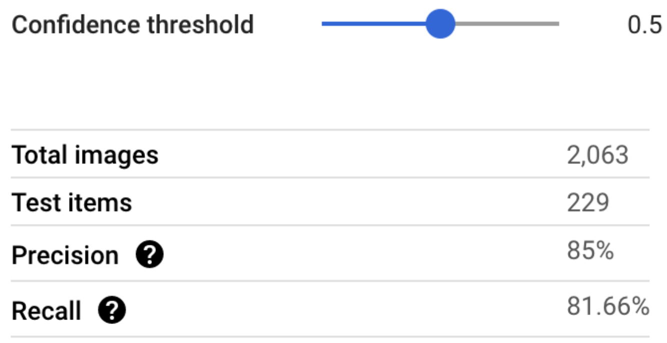


Labels	Images	Train	Validation	Test
BACTERIAL PNEUMONIA	995	796	100	99
NORMAL	300	240	30	30
VIRAL PNEUMONIA	997	797	100	100

<table>
<tr>
<td></td>
<td>


Confidence threshold 0.5

<table>
<tr><th>True Label</th><th>Predicted Label
VIRAL PNEUMONIA</th><th>BACTERIAL PNEUMONIA</th><th>NORMAL</th></tr>
<tr><td>VIRAL PNEUMONIA</td><td>74%</td><td>24%</td><td>2%</td></tr>
<tr><td>BACTERIAL PNEUMONIA</td><td>6%</td><td>92%</td><td>2%</td></tr>
<tr><td>NORMAL</td><td>17%</td><td>3%</td><td>80%</td></tr>
</table>

The resulting confusion matrix for the new data set with increased data points resulted in less false positives. The model was better at predicting bacterial pneumonia with true positives at 92%
</td>
</tr>
<tr>
<td>
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)?
</td>
<td>


Confidence threshold 0.5

<table>
<tr><td>Total images</td><td>2,063</td></tr>
<tr><td>Test items</td><td>229</td></tr>
<tr><td>Precision ?</td><td>85%</td></tr>
<tr><td>Recall ?</td><td>81.66%</td></tr>
</table>

The new data set used, with more data points for both viral and bacterial dataset compared to the normal points, improved the precision and recall. A precision of 85% and a recall of 81.66% is noted for the confidence threshold for 0.5 from the above image.
</td>
</tr>
<tr>
<td>
F1 Score
What is this model’s F1 score?
</td>
<td>
F1 score = [(2 * precision * recall) / (precision + recall)]
=[(2 * 0.85 * 0.8166) / (0.85 + 0.8166)]
= 0.8329
</td>
</tr>
</table>