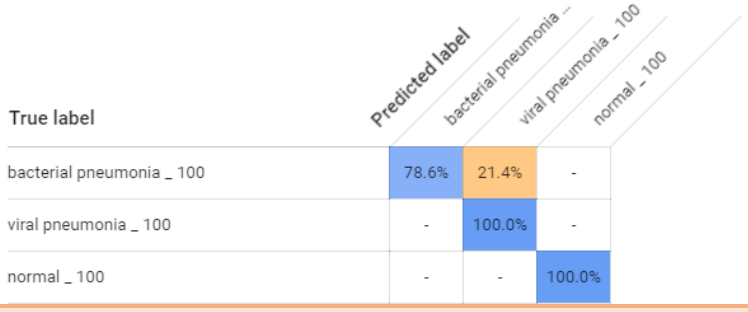
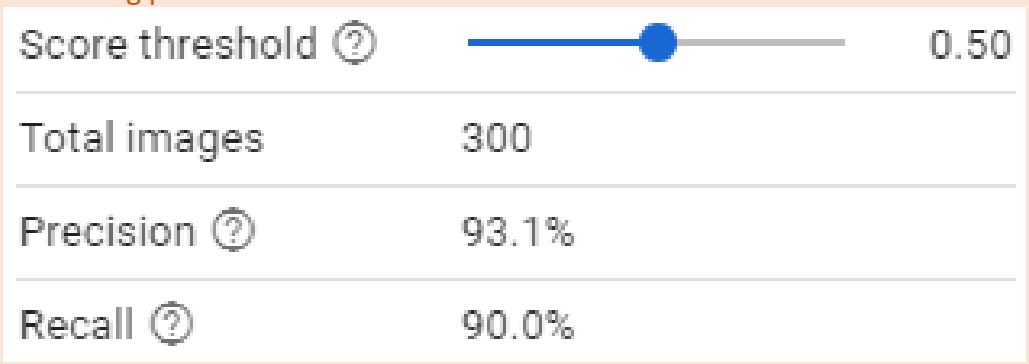
AutoML Modeling Report

<*B.Akintade*>

3-Class Model

Confusion Matrix	For a balanced dataset of 100 in each class (normal, viral pneumonia and bacterial pneumonia) the confusion matrix shows that the model highly classified 'normal' and 'viral pneumonia' labels 100% of the time. 'Bacterial pneumonia' label was predicted 78.6% and most often confused for 'viral pneumonia' labels 21.4%. High confusion implies incorrect predictions which may indicate inconsistent labels – a way to remedy this confusion is to provide more additional training data for the model. **Confusion matrix** This table shows how often the model classified each label correctly (in blue), and which labels were most often confused for that label (in orange). 
Precision & Recall	Precision = 93.1%; Recall = 90%; Avg Precision = 0.959 = ~1.0 Precision = True Positives ÷ (true positives + false positives) = pneumonia correctly identified ÷ (pneumonia correctly identified + 'normal' incorrectly labelled as 'pneumonia') There's 3.1% higher rate in precision where the model is predicting only true relevant instances/positive cases of pneumonia. With high precision, every children identified as testing positive does actually have pneumonia, but since the model is getting false negatives, it may also miss some children who actually have pneumonia but labelled as not having pneumonia. 

	Precision and recall of the whole model is a common metric known as the average precision which is precision averaged across all values of recall between 0 and 1, and calculated as a single number that characterizes the total performance of the classifier. Also, Google AutoML Vision states that, 'average precision measures how well a model performs across all score thresholds, by calculating the areas under the precision-recall trade off curve. 1.0 is the maximum score' Average Precision = [Clean/Balanced Data + Clean/Unbalanced Data + Dirty/Balanced Data + 3-Class Model] ÷ 4 Avg. Precision = [1.0 + 0.989 + 0.775 + 0.959] ÷ 4 = 3.723 ÷ 4 = 0.93075 = ~1.0
F1 Score	The F1 score combines the precision and recall to produce an overall performance measure model. F1 = 2*(precision*recall) ÷ (precision + recall) = 2*(93.1% * 90.0%) ÷ (93.1% + 90.0%) = 2*(0.8379 ÷ 1.831) = 91.5% or 0.915 = ~1 The F1-score ranges from zero to one, with one being perfect precision in recall

* Google Cloud, Evaluating models - https://cloud.google.com/vision/automl/docs/evaluate

EVALUATE

Model

AutoMLDataset_4b_v20190927145928

Created	Analyzed	Avg precision	Precision	Recall
Sep 27, 2019	300 images	0.959	93.103%	90.0%
1 compute hour	3 labels, 30 test images			

Precision and recall are based on a score threshold of 0.5

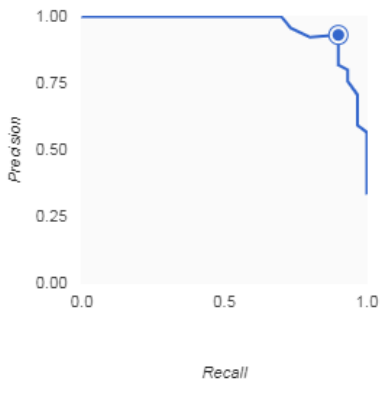
Type to filter labels...

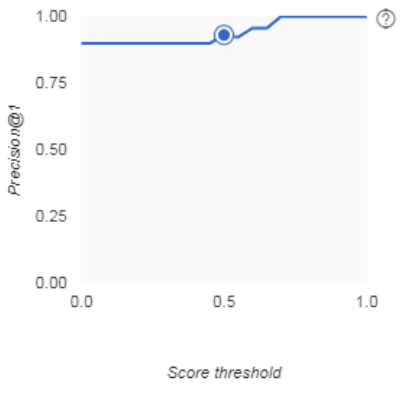
- All labels
- bacterial pneumonia _ 100
- normal _ 100
- viral pneumonia _ 100

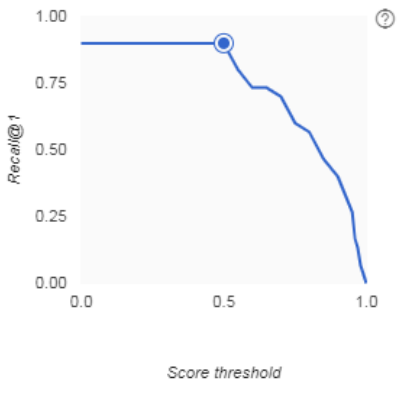
All labels

Score threshold	0.50
Total images	300
Precision	93.1%
Recall	90.0%

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







Confusion matrix

This table shows how often the model classified each label correctly (in blue), and which labels were most often confused for that label (in orange).

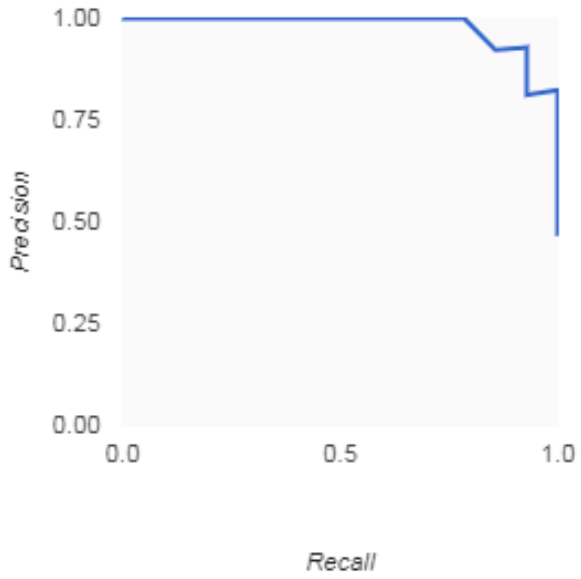
True label \ Predicted label	bacterial pneumonia ...	viral pneumonia _ 100	normal _ 100
bacterial pneumonia _ 100	78.6%	21.4%	-
viral pneumonia _ 100	-	100.0%	-
normal _ 100	-	-	100.0%

PREDICT

bacterial pneumonia _ 100

Score threshold	0.50
Total images	30
Precision	100.0%
Recall	78.6%

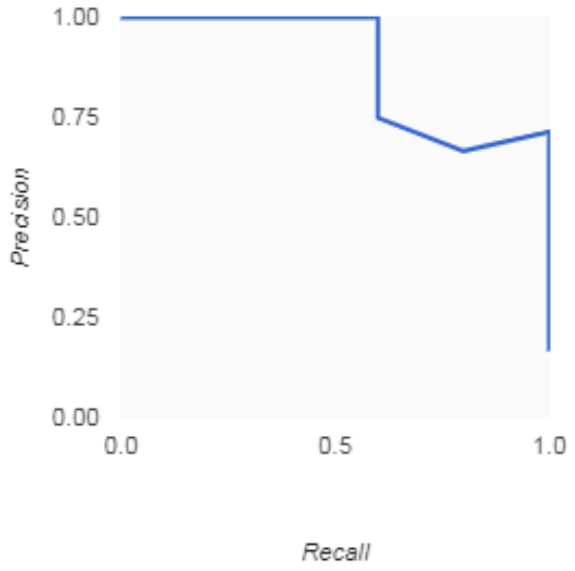
Use the slider to see which score threshold works best for your model on the precision-recall tradeoff curve. Learn more about these metrics and graphs



viral pneumonia _ 100

Score threshold	0.50
Total images	30
Precision	71.4%
Recall	100.0%

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



normal _ 100

Score threshold	0.50
Total images	30
Precision	100.0%
Recall	100.0%

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