

AutoML Modeling Report



Classifying Pneumonia in Children Chest X-ray Images Using Google AutoML Vision.

Building four different types of ML models

<Bakary Badjie>

Binary Classifier with Clean/Balanced Data

Train/Test Split

How much data was used for training? How much data was used for testing?

160 image samples were used for training while 40 image samples were used for testing.

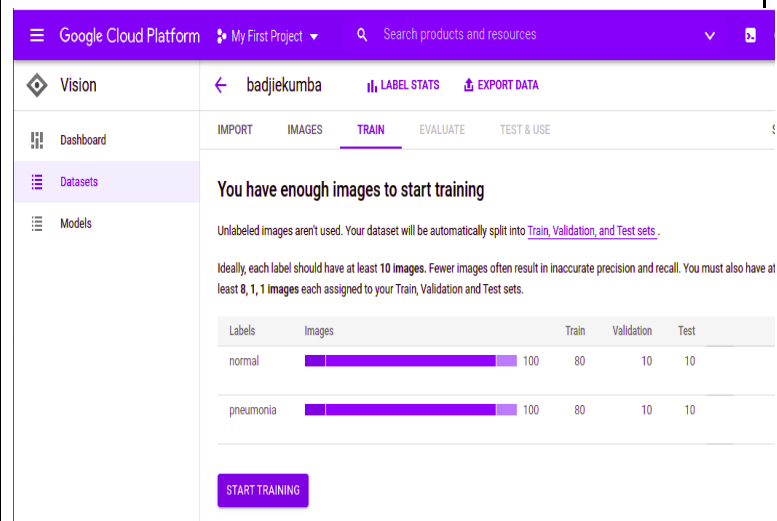


Figure 1.

Figure1 indicates that, 80 image samples were used for training for both normal and pneumonia sets. 10 image samples were used for validation for both normal and pneumonia, while 10 image samples were also used for tests for both normal and pneumonia.

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

The confusion matrix displays how many times our model has categorized each label (normal and pneumonia). It also demonstrates the level of accuracy our desired classes were recognized by the model. The graph indicates that, in the normal category, 100% of our data was correctly classified

positive rate for the “pneumonia” class? What is the false positive rate for the “normal” class?

as normal while 0% were classified as pneumonia. However, in the pneumonia category, 95% were correctly classified as ‘pneumonia’ while 5% were misclassified as normal. The true positive rate for “pneumonia” is 95 percent while the false positive rate for “normal” is 0 percent according to our confusion matrix

Confusion matrix

This table shows how often the model classified each label correctly (in blue), and which labels were most often confused. You can download the entire confusion matrix as a CSV file.

True Label	Predicted Label	
	pneumonia	normal
pneumonia	95%	5%
normal	0%	100%

Figure 2.

Precision and Recall

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)?

Precision and recall are two typical metrics used to assess the model, and this helps us to comprehend the level of our model's performance for a single class but also in several classes. They also show us how effectively our model has truly classified our data and how much our model has wrongly classified our data.

Precision enable us determining the number of test samples that are assigned to both normal and pneumonia sets. It also helps us understand the correctness of our model's predictions.

The recall gave us the percentage of how likely or how correctly the prediction was made by the model for both the normal and pneumonia classes.

Talking about the score threshold of 0.5, our model had 85 percent for both precision and recall. (See figure3). This shows that, our data is well organized and well generalized.

Score Threshold

When you increase the threshold what happens to precision? What happens to recall? Why?

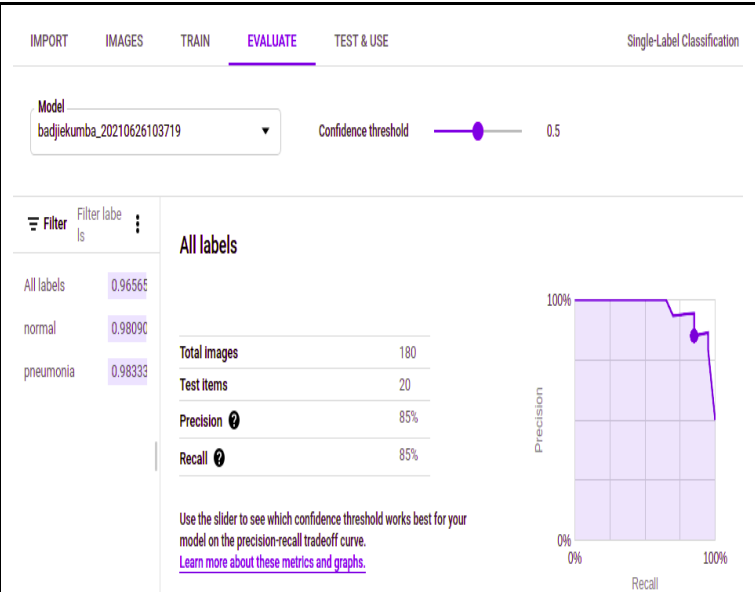


Figure 3.

At a confidence threshold of 0.5, we have 85% for both precision and recall.

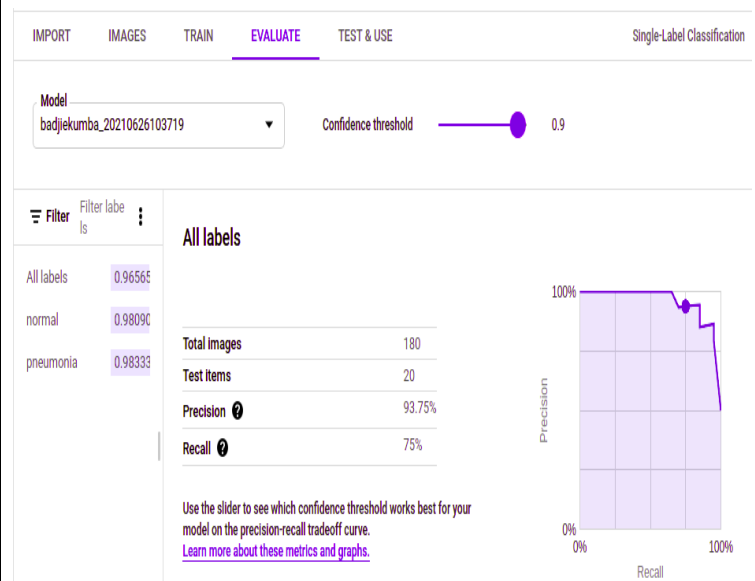


Figure 4.

When we increased the confidence threshold to 0.9, we noticed that, the precision had increased from 85% to 93.75%, whereas the recall dropped from 85% to 75%. (See figure4.)

This occurs due to the fact that, as the degree of confidence-threshold in assigning a category rises, our model categorized lesser data while having the chance of reducing the number of misclassification.

Binary Classifier with Clean/Unbalanced Data

Train/Test Split

How much data was used for training? How much data was used for testing?

319 image samples were used for training while 40 image samples were used for testing as indicated in the figure below.

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	 299	239	30	30

Figure 5.

Confusion Matrix

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

The confusion matrix demonstrates that, 100 percent of our data labeled as normal class had been perfectly classified in the normal category while 100% of our data labeled as pneumonia class was perfectly classified as pneumonia. This indicated that, 0% of our data labeled as normal was misclassified and 0% of our data labeled as pneumonia had been misclassified as well. This shows that, the model produced no false positives and false negatives as compared to the previous model with a balanced dataset. In other words, the model has made no misclassifications on both the images labeled as normal or pneumonia. (See figure6 below)

Confusion matrix

Item counts

This table shows how often the model classified each label correctly (in blue), and which labels were most often confused that label (in gray). You can download the entire confusion matrix as a CSV file.

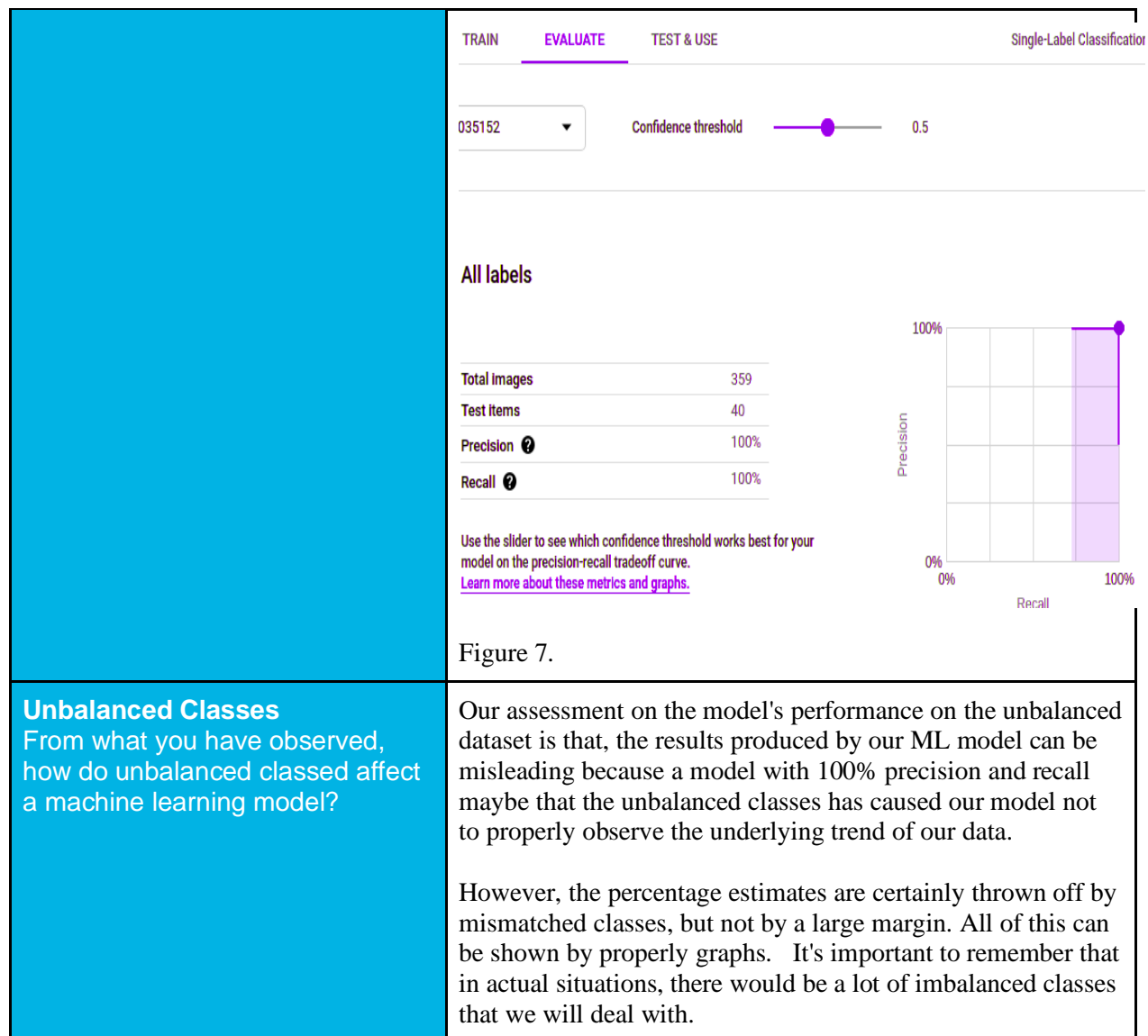
True Label	Predicted Label	
	normal	pneumonia
normal	100%	-
pneumonia	-	100%

Figure 6.

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)?

Looking at the confidence threshold of 0.5, our model has attained 100% recall and precision for every image labeled as pneumonia and normal respectively, which is as a result of the impact of the unbalanced dataset used to train the model (See figure7). However, this prediction could be acceptable, but it could also be a generalization problem for our model. This means that the data has not been too well generalized. Because it is unusual or out of the ordinary for a model to be 100% precision and recall at the same time.



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

☐ Item counts

This table shows how often the model classified each label correctly (in blue), and which labels were most often confused that label (in gray). You can download the entire confusion matrix as a CSV file.

True Label	Predicted Label	
	pneumonia	normal
pneumonia	70%	30%
normal	10%	90%

Figure 8.

Looking at the model's performance on our unclean datasets, 30% of our data in the pneumonia category is misclassified. This is not a surprise because we deliberately put 30% of the normal images into the pneumonia category. However, there is also a 10% misclassification performed by the model in the normal category, which the model was confused about. Therefore, comparing the model's performance on the clean-balanced dataset and the dirty-balanced dataset, there is a huge increase in the error rate, especially in the pneumonia category. Therefore, the dirty-balance data has increased the error rate of our model in both categories.

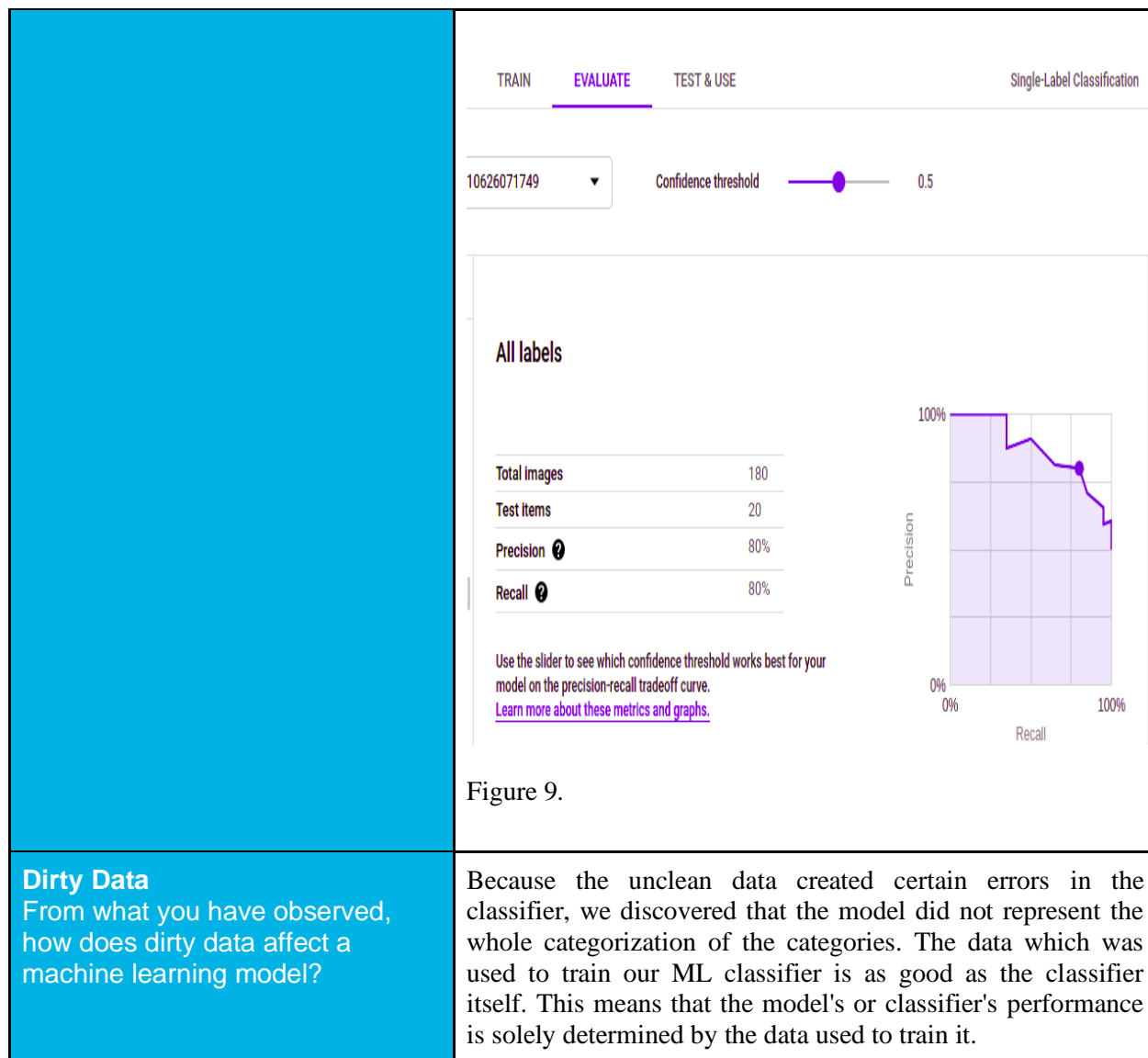
Precision and Recall

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?

The precision and recall of our model are of the same values as shown in the image below. We have Precision of 80% and recall of 80% as well.

Since we deliberately mislabeled 30% of our data for each category, we observed that the mislabeled data shared the same visual characteristic with our correctly labeled data and the model is bias in its predictions.

At a confidence threshold of 0.5, our binary classifier with clean-balance data has a higher precision (85%) and recall (85%) than this classifier trained with unclean-balanced dataset which has a precision and recall of 80% each.



Dirty Data
From what you have observed, how does dirty data affect a machine learning model?

Because the unclean data created certain errors in the classifier, we discovered that the model did not represent the whole categorization of the categories. The data which was used to train our ML classifier is as good as the classifier itself. This means that the model's or classifier's performance is solely determined by the data used to train it.

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 confusion matrix indicates that our model has accurately classified the entire image label as normal in the normal category at 100%, while the model has classified 90% accuracy for all the images labeled as bacterial pneumonia and 90% for viral pneumonia in their respective classes. Therefore, the model was most likely confused about its predictions for the bacterial pneumonia and viral pneumonia classes, with 10% of confusion for each class. While the model most likely get the normal class right with 0% confusion.

However, one of the ways to remedy the confusion of the model with its predictions is to add more data to the classes that the model was confused with.

Confusion matrix

 Item counts 

This table shows how often the model classified each label correctly (in blue), and which labels were most often confused for that label (in gray). You can download the entire confusion matrix as a CSV file.

True Label	Predicted Label		
	normal	bacterial_pneumonia	viral_pneumonia
normal	100%	-	-
bacterial_pneumonia	-	90%	10%
viral_pneumonia	-	10%	90%

Figure 10.

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)?

We have a precision of 93.33% and a recall of 93.33%. They are calculated as follows.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ positives + False\ Negatives}$$

However, the average precision is the precision averaged over all instances of recall between 0 and 1, and it is computed as a single number that represents the entire performance of our model, is a common metric that captures the precision and recall of the entire model. 'Average precision evaluates how well a model performs across all score thresholds, by computing the areas under the accuracy-recall trade off curve.

$$Average\ Precision = \frac{clean/balance + clean/unbalanced + dirty/balance + 3class}{4}$$

$$= \frac{0.85 + 1.0 + 0.80 + 0.9333}{4} = 0.896$$

At a confidence threshold of 0.5, the model has precision and recall as shown in the figure below.

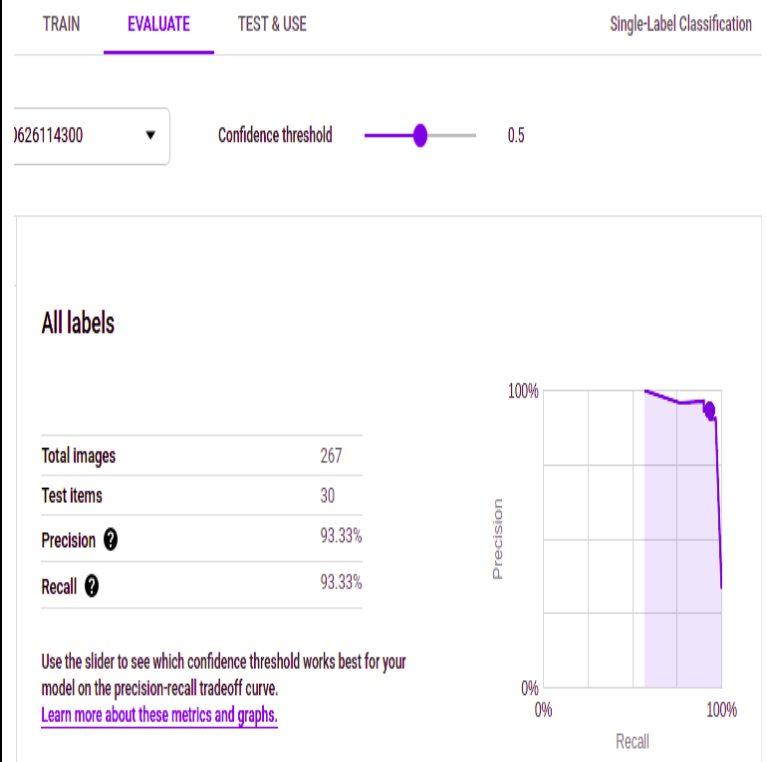


Figure 11.

F1 Score

What is this model's F1 score?

The F1 score analyses or measures the overall performance of the model by incorporating precision and recall together.

$$\begin{aligned} F1\ Score &= 2 \times \frac{Precision \times Recall}{Precision + Recall} \\ &= 2 \times \frac{0.9333 \times 0.9333}{0.9333 + 0.9333} = 0.9333 \cong 1.0 \end{aligned}$$