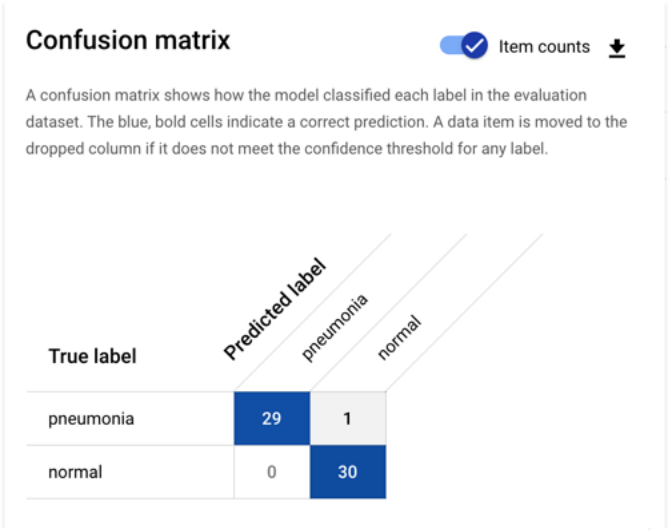
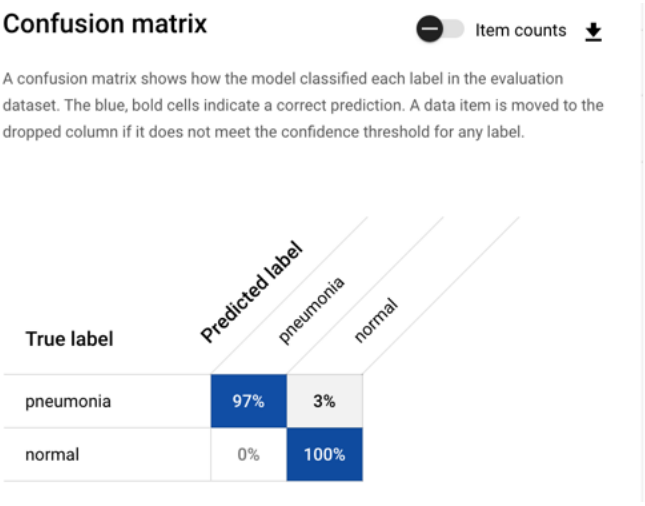
AutoML Modeling Report

Stephen D. Gardner

Binary Classifier with Clean/Balanced Data

Train/Test Split How much data was used for training? How much data was used for testing?	600 images total were used. 300 normal and 300 pneumonia. 480 images were used for training. 60 Images were used for testing. The model training took 2hrs 58mins to complete. Below is a screenshot: All labels Average precision 0.984 Precision 98.3% Recall 98.3% Created Sep 28, 2024, 4:56:25 PM Total images 600 Training images 480 Validation images 60 Test images 60
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?	Each cell in the confusion matrix represents the outcomes of predictions: True Positives (correctly predicted pneumonia), True Negatives (correctly predicted normal), False Positives (normal predicted as pneumonia), and False Negatives (pneumonia predicted as normal). From the provided confusion matrices, the values observed were 29 True Positives and 1 False Negative for pneumonia, and 30 True Negatives and 0 False Positives for normal. The true positive rate for the pneumonia class was 97% (TP/(TP + FN) = 29/30), while the false positive rate for the normal class was 0% (FP/(FP + TN) = 0/30). A screenshot of confusion matrix is below:

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




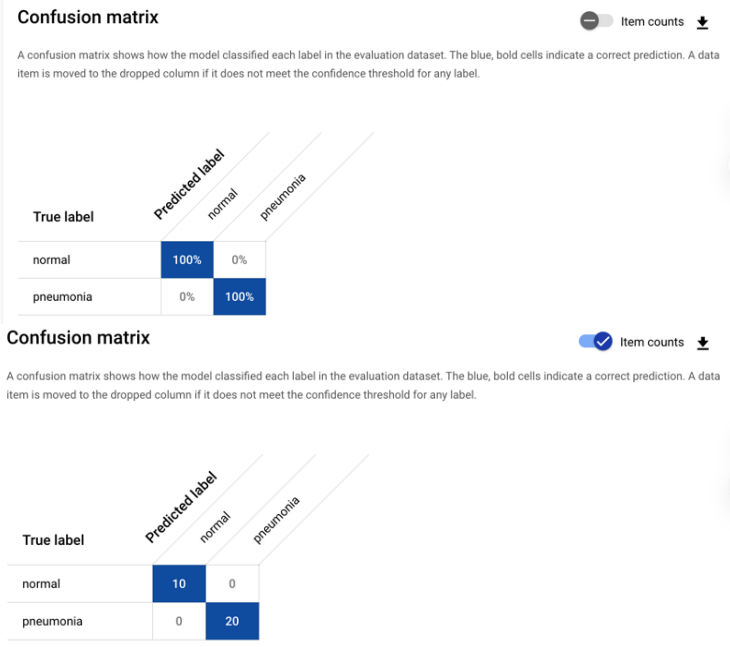
</td>
</tr>
<tr>
<td>Precision and Recall
What does precision measure? What does recall measure? What precision and recall did the model achieve (report the values for a score threshold of 0.5)?</td>
<td>Precision measures the proportion of true positive predictions out of all positive predictions made by the model. It reflects how many of the predicted positive instances are actually positive. Recall measures the proportion of true positives out of all actual positive cases. It tells you how well the model can identify all relevant positive cases.

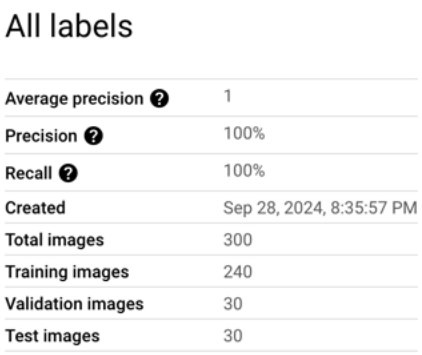
In reference to the Binary Classifier with Clean/Balanced Data model Precision was measured at: 98.3% and Recall was measured at 98.3%.

A screenshot is provided below:</td>
</tr>
</table>

	Evaluation details Confidence threshold 0.5 All labels Average precision 0.984 Precision 98.3% Recall 98.3% Created Sep 28, 2024, 4:56:25 PM Total images 600 Training images 480 Validation images 60 Test images 60
Score Threshold When you increase the threshold what happens to precision? What happens to recall? Why?	When the confidence threshold is increased in a classification model, precision improves because the model becomes more selective, predicting "pneumonia" only when it is more certain. This leads to fewer false positives, increasing precision. However, recall decreases because the model becomes stricter, missing some true "pneumonia" cases (false negatives), resulting in lower recall. For example, with a higher threshold, the model might correctly classify more "normal" cases as not pneumonia, but it might fail to catch all pneumonia cases, sacrificing recall. This trade-off between precision and recall is common in classification models. A screenshot is provided below: Evaluation details Confidence threshold 0.8 All labels Average precision 0.984 Precision 98.3% Recall 96.7% Created Sep 28, 2024, 4:56:25 PM Total images 600 Training images 480 Validation images 60 Test images 60

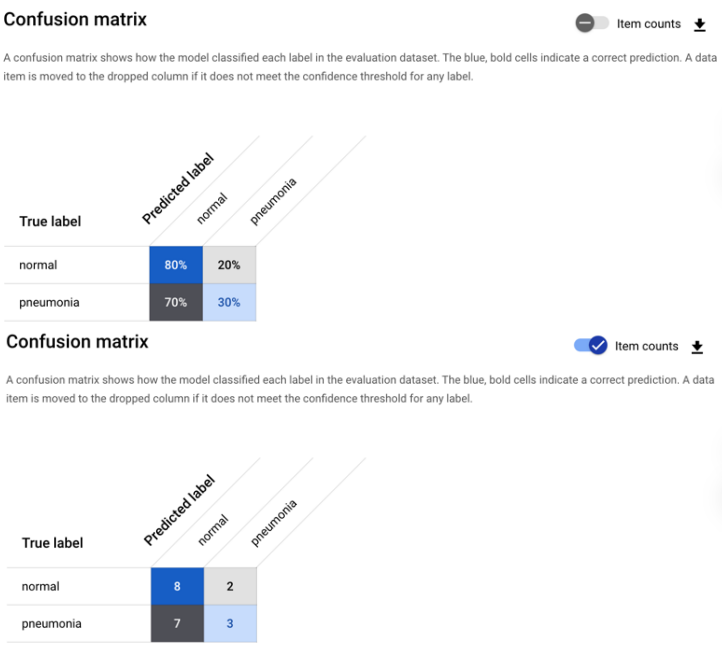
Binary Classifier with Clean/Unbalanced Data

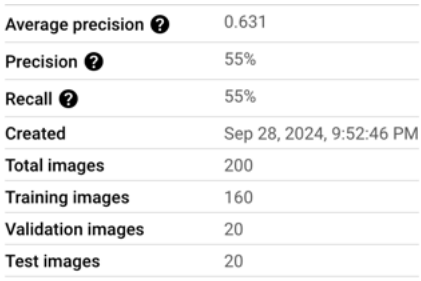
Train/Test Split How much data was used for training? How much data was used for testing?	300 images were used. 100 normal images and 200 pneumonia images. 30 images were used for testing. The entire model took 1hr 43mins to complete.
Confusion Matrix How has the confusion matrix been affected by the unbalanced data? Include a screenshot of the new confusion matrix.	The confusion matrix reflects a perfectly accurate model performance, where 100% of both "normal" and "pneumonia" images were classified correctly. The model identified all 10 normal images and 20 pneumonia images in the test set without any errors. However, the dataset is unbalanced, with a higher number of pneumonia images (200) compared to normal images (100). This imbalance might give the impression that the model is more robust than it actually is because it is tested on fewer normal images. While the confusion matrix looks ideal, this imbalance could cause the model to be overly biased towards classifying pneumonia cases correctly, potentially overlooking the minority class in future scenarios. Further testing on a more balanced dataset might be required to evaluate the model's performance comprehensively. A screenshot of the confusion matrix is below: 

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)?	In this confusion matrix, and with a score threshold of .5 both precision and recall are 100% because the model correctly classified all normal (10 out of 10) and pneumonia (20 out of 20) images with no errors. Precision is 100% as there were no false positives, indicating every prediction was accurate. Similarly, recall is 100% since the model did not miss any true positive cases, meaning there were no false negatives. A screenshot of the precision and recall is below:  All labels Average precision 1 Precision 100% Recall 100% Created Sep 28, 2024, 8:35:57 PM Total images 300 Training images 240 Validation images 30 Test images 30
Unbalanced Classes From what you have observed, how do unbalanced classed affect a machine learning model?	Given the unbalanced nature of the dataset, with a greater number of pneumonia images than normal ones, suggests a potential bias in the model. This bias could cause the model to perform better on pneumonia cases while not necessarily generalizing well across the entire dataset of 300 images. Thus, while this subset shows perfect performance, further evaluation is needed to assess the model's behavior across all data points.

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.	This model took 2hrs 33mins to complete with a total of 100 "normal" and 100 "pneumonia" images. In this confusion matrix, after switching the labels of 30 images from each class (normal and pneumonia), the model's performance reflected in the confusion matrix showed some misclassifications. For the normal class, out of 10 true normal images, 7 were correctly classified, while 3 were misclassified as pneumonia. Similarly, for the pneumonia class, 7 images were correctly identified, but

	3 were incorrectly labeled as normal. This change results from intentionally altering the labels of some images, leading to more errors in classification, as seen in the confusion matrix. The label-switching causes increased misclassification, impacting the precision and recall scores. These numbers illustrate how altering label distribution, even slightly, can affect the model's accuracy and reliability when evaluating a dataset. A screenshot of the confusion matrix is below: 
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?	In reference to the model, the precision and recall are both 55%, as the model correctly classified 8 normal images and 3 pneumonia images. Precision is calculated as the number of true positives (correct predictions) divided by the total number of predicted positives. In this case, for the normal class, 8 out of 10 predictions were correct, and for the pneumonia class, 3 out of 10 predictions were correct. Therefore, the precision is low due to the presence of false positives, where 7 pneumonia images were incorrectly classified as normal. Similarly, recall is 55% because the model failed to correctly identify 7 true positive pneumonia cases. This drop in performance indicates that the model struggles to differentiate between normal and pneumonia images, likely due to the distribution and quality of the training data.

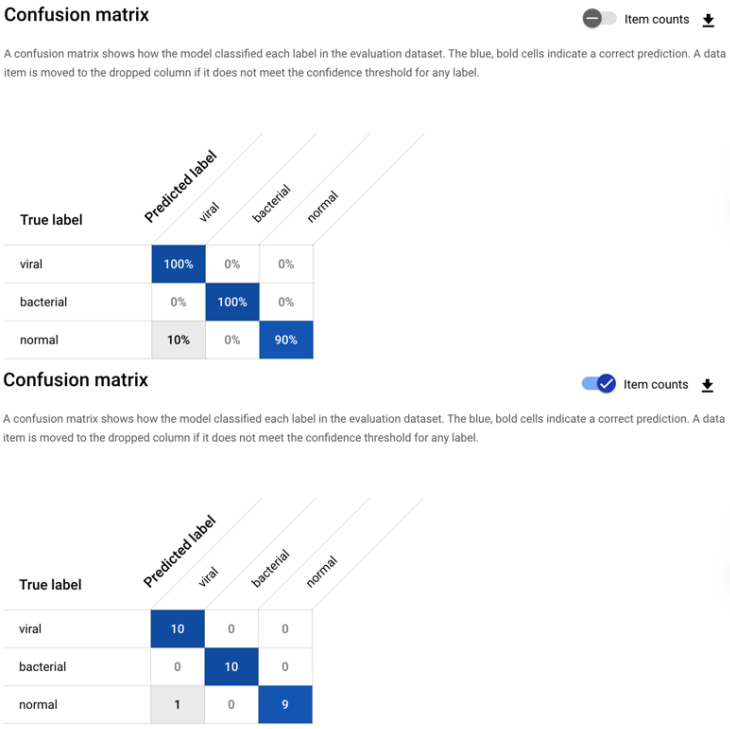
	A screenshot of the model's precision and recall is below:  Comparing Model #1, Model #2, and Model #3 reveals key differences in their precision and recall due to the data characteristics and how clean or unbalanced it was. **Model #1** was trained on a clean, balanced dataset with 300 normal images and 300 pneumonia images. It achieved excellent results with both precision and recall measured at 98.3%. The balanced dataset helped the model perform well without biases toward any class, as seen from its high precision and recall. **Model #2** was trained on a clean, unbalanced dataset with 100 normal images and 200 pneumonia images. Despite the imbalance, it performed perfectly, achieving 100% precision and recall. The confusion matrix showed no misclassifications, as all images in the test set were correctly identified. The unbalanced data did not impact this model's performance due to its clean training data. **Model #3** was trained on a balanced but dirty dataset where labels of 30 images were switched between normal and pneumonia. This label-switching caused the model to perform poorly, with both precision and recall dropping to 55%. The confusion matrix showed several misclassifications, reflecting how dirty data severely impacted the model's accuracy. Model #3 struggled to differentiate between normal and pneumonia images. In conclusion, **Model #2** performed the best with perfect scores, followed closely by **Model #1**. **Model #3** was the weakest, affected by the intentional label switching.
Dirty Data From what you have observed, how does dirty data affect a machine learning model?	Dirty data can significantly impact the performance of machine learning models by introducing errors, biases, and inaccuracies into the training process. When data contains issues such as incorrect labeling, missing values, duplicates, or inconsistencies, models trained on this data may learn incorrect patterns or fail to

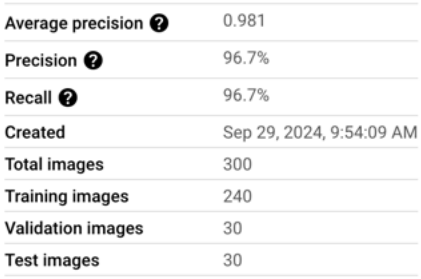
Average precision	0.631
Precision	55%
Recall	55%
Created	Sep 28, 2024, 9:52:46 PM
Total images	200
Training images	160
Validation images	20
Test images	20

	generalize well to new data. For example, if a dataset includes mislabeled or incomplete entries, the model might misclassify certain categories, leading to lower accuracy, precision, and recall scores. Moreover, the presence of dirty data can cause biased models, as it skews the underlying relationships that the model relies on for prediction. Cleaning and maintaining high quality data is essential to ensure that machine learning models have a solid foundation from which they can learn, thereby improving their overall reliability and performance .

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 three-class model took 5hrs and 13mins to complete. The even breakdown included 100 "normal" images, 100 "bacterial pneumonia" images, and 100 "viral pneumonia" images (for a total of 3 classes).The confusion matrix shows that the model has classified 10 images of the "viral" class, 10 images of the "bacterial" class, and 9 images of the "normal" class correctly. There is only one misclassification, where the model predicted one "bacterial" image as "normal." This suggests that the model is most likely to confuse the "bacterial" and "normal" classes, though the confusion is minimal. The "viral" class seems to be the easiest for the model to classify correctly, as there are no errors related to this class. To remedy the model's confusion between "bacterial" and "normal" classes, additional data could be collected to further train the model, particularly with more diverse images from both classes. Adjusting the model's hyperparameters or applying techniques like data augmentation could also help to improve the model's ability to distinguish between these two classes. A screenshot of the confusion matrix is below:

	
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)?	Utilizing a score threshold of .5, the model achieved a precision of 96.7% and a recall of 96.7%. Precision indicates the percentage of positive predictions that were correct, meaning 96.7% of the model's positive predictions were accurate. Recall represents the percentage of actual positive cases that the model correctly identified, which in this case was also 96.7%. These values were calculated using the confusion matrix. **True Positives (TP):** The number of instances correctly predicted for each class. **False Positives (FP):** The number of instances incorrectly predicted for each class. **False Negatives (FN):** The number of instances missed by the model for each class. **Precision** is the ratio of true positives to the sum of true positives and false positives. Example: Precision = TP / (TP + FP) **Recall** is the ratio of true positives to the sum of true positives and false negatives. Example:

	Recall = TP / (TP + FN) With these values, the model performs well, making it reliable in identifying positive cases while minimizing false positives and false negatives. A screenshot of precision and recall is below:  The individual values for precision and recall across the four models were: **Model 1**: Precision = 98.3%, Recall = 98.3%; **Model 2**: Precision = 100%, Recall = 100%; **Model 3**: Precision = 55%, Recall = 55%; **Model 4**: Precision = 96.7%, Recall = 96.7% The calculation for the average precision and recall for all four models was 87.5%, using the following calculations AP = (98.3+100+55+96.7)/4 = 87.5% AR = (98.3+100+55+96.7)/4 = 87.5% I calculated this by averaging the precision and recall values for each model. This demonstrates a high level of performance overall, even though model 3 had significantly lower precision and recall due to the label switching. The consistently high scores in models 1, 2, and 4 contributed to a strong overall performance, indicating that the classifiers worked well in most cases, except for the intentionally disrupted data in model 3.
F1 Score What is this model's F1 score?	This three-class model has a strong balance between precision (its ability to avoid false positives) and recall (its ability to detect true positives), making it highly reliable in this case. The F1 score was .967. This was determined by using

Average precision	0.981
Precision	96.7%
Recall	96.7%
Created	Sep 29, 2024, 9:54:09 AM
Total images	300
Training images	240
Validation images	30
Test images	30

	the following formula: $$F1 = 2 \times ((\text{Precision x Recall})/(\text{Precision} + \text{Recall}))$$ In reference to the three-class model, the F1 was evaluated as: $$F1 = 2 \times ((0.967 \text{ x } 0.967)/(0.967 + 0.967)) = 0.967$$ According to Google AutoML documentation, the best results meaning a precision and recall higher than 85% can be established by using clean and properly labeled data, and by having approximately 1,000 images were class. In reference to three-class we have needed 1,000 of each normal; bacterial; and viral images across a well-balanced 3,000 total image dataset to help properly classify all three classes in the long-term. Additionally, given the time and cost of training such a model, properly stake holder approvals would be needed to keep the project within budget.