Online Cheating Detection using YOLOv8

Model Performance Analysis

By

Hamza Kholti

Email: hamza.kholti@etu.uae.ac.ma

October 2024

0.1 Training and Validation Metrics

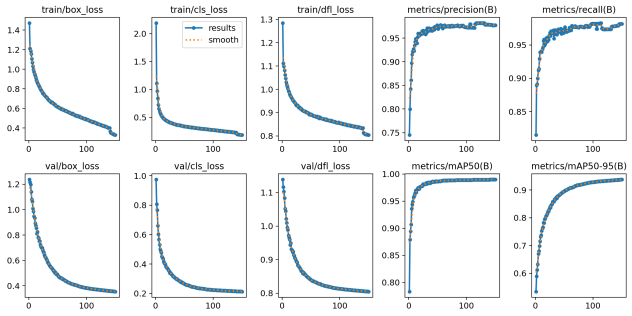

Figure 1: Confusion matrix illustrating the model's performance across different classes.

Model Performance: The model is learning effectively, as indicated by the consistent decrease in both training and validation losses without signs of overfitting.
High Precision and Recall: High precision and recall suggest a good balance between correct detections and the model's ability to detect all relevant instances.
Increasing mAP Scores: The mAP scores confirm that the model is improving in its object detection capabilities across different IoU thresholds.

0.2 Confusion Matrix

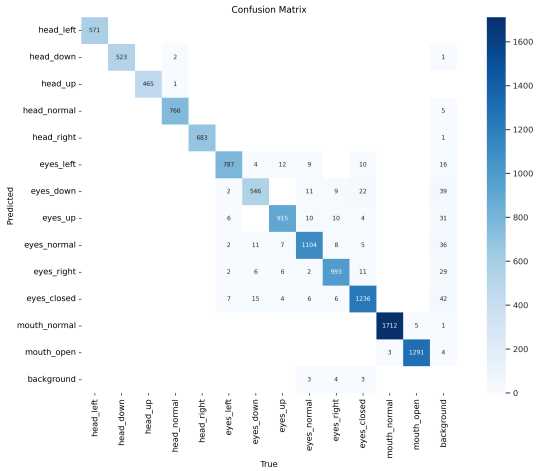

Figure 2: Confusion matrix illustrating the model's performance across different classes.

For most classes, there is a high number of correct predictions along the diagonal of the matrix. The matrix shows some misclassifications, where predictions fall outside the diagonal. For example, eye classes like *eyes_normal* and *eyes_up* have minor confusions with similar classes, such as *eyes_right* and *eyes_down*. This is likely due to subtle differences that are challenging for the model to distinguish.
The background class has a few misclassifications, but most background samples are correctly identified. The majority of classes have high true positives with relatively low misclassifications. This suggests that the model has good overall performance, with room for improvement in differentiating between similar classes (e.g., eye and head orientations).

0.3 Labels

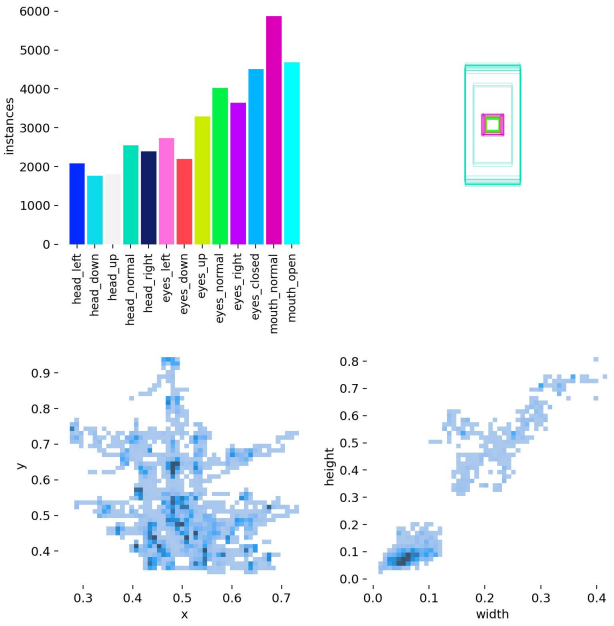

Figure 3: Various graphs illustrating the label distribution and bounding box characteristics in the dataset.

Top Left (Bar Chart): The bar chart shows the number of instances per label, giving an overview of the dataset's distribution across different classes. The labels *eyes_closed* and *mouth_open* have the highest counts, which may indicate a potential class imbalance that could affect the model's performance.
Bottom Left (Scatter Plot for x and y Coordinates): This scatter plot depicts the distribution of bounding box center coordinates along the x and y axes, possibly for the facial landmarks or regions. It shows a general distribution pattern of bounding boxes, with more density around the center, indicating that most features are centered in the image.
Bottom Right (Scatter Plot for Width and Height of Bounding Boxes): This scatter plot visualizes the distribution of bounding box width and height. It shows a concentration at certain size ranges, especially small width and height, which could indicate that the majority of bounding boxes are relatively small.

0.4 Precision-Confidence Curve:

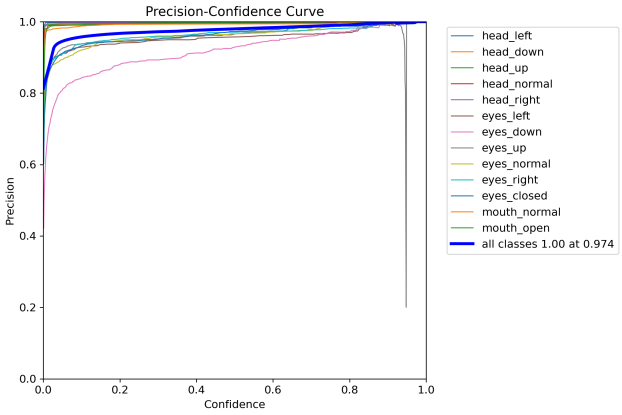

Figure 4: Precision-confidence curve illustrating the model's performance across different classes.

This graph shows a curve of precision versus confidence for different classes:
High Precision at High Confidence Levels: For all classes, precision rises quickly with confidence, reaching nearly 1 around a confidence of 0.3 to 0.4, and remains high as confidence nears 1.
Variability Across Classes: The precision of certain classes (such as *eyes_down* in pink) starts lower but joins the other classes as confidence increases. This may indicate that some classes are more challenging to detect reliably at lower confidence levels.

0.5 Recall-Confidence Curve:

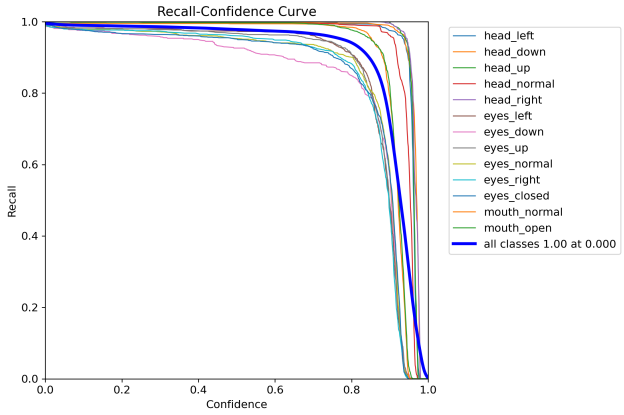

Figure 5: Recall-confidence curve illustrating the model's performance across different classes.

This graph shows a curve of recall versus confidence for different classes:
High Recall at Low Confidence Levels: When detection confidence is low (toward the left side of the graph), recall is almost 1 for all classes, meaning that nearly all positive instances are detected, though with little certainty.
Drop in Recall at High Confidence Levels: Toward the right side of the graph, as confidence increases, recall quickly declines. This means that at higher confidence levels, some detections are missed, which is typical because the algorithm becomes more selective.

0.6 Precision-Recall Curve:

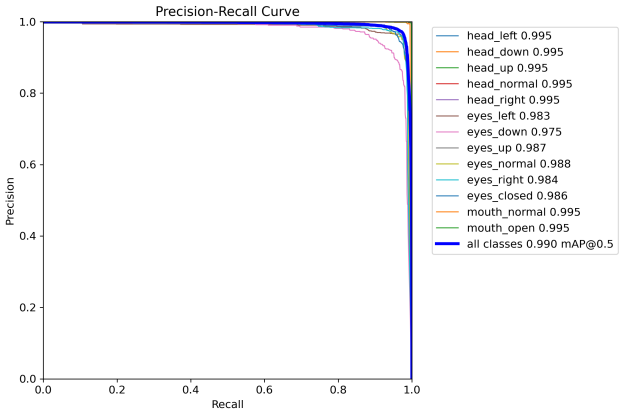

Figure 6: Precision-recall curve illustrating the model's performance across different classes.

This graph shows a curve of precision versus recall for different classes:
High Precision over a Wide Range of Recall: Precision is close to 1 across a large portion of recall, indicating a very low false detection rate. The classes remain precise even as recall increases, which is a sign of high detection quality.
Decline in Precision at High Recall Levels: Toward the right side of the graph (recall near 1), there is a slight decrease in precision for some classes. This suggests that when the model tries to capture all positive instances (high recall), it begins to include more false positives, reducing precision.

0.7 F1-Confidence Curve:

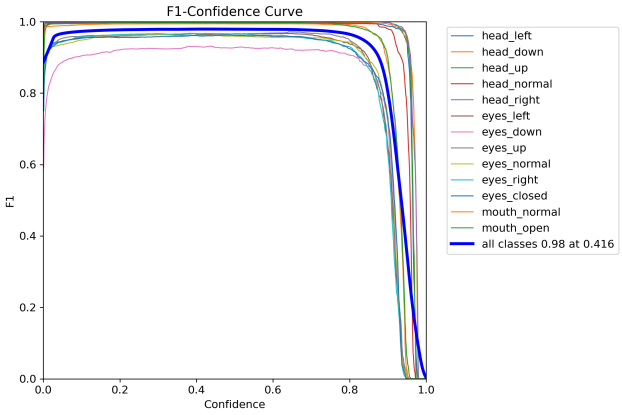

Figure 7: F1-confidence curve illustrating the model's performance across different classes.

This F1-confidence curve illustrates the relationship between confidence thresholds and F1 scores across different classes, indicating the model's performance as confidence levels change. Here's an analysis:

High F1 Scores Across Classes: Most classes achieve F1 scores close to 1, even at lower confidence thresholds (e.g., around 0.4). This indicates the model is generally accurate across a range of classes and performs well without requiring high confidence levels to achieve high F1 scores.

Variance Among Classes: Some classes (e.g., *eyes_down*, shown in light pink) have lower F1 scores, especially at lower confidence thresholds, but they still reach high F1 scores as confidence increases. This suggests that these classes may have more variability or are harder to classify accurately at lower confidence levels.

Overall Model Performance: The curve labeled "all classes" shows an overall F1 score of approximately 0.98 at a confidence threshold of 0.416, indicating that the model maintains high accuracy even with moderate confidence. This reflects a well-balanced performance across classes.

Impact of Confidence Thresholds: As confidence thresholds approach 1.0, all classes' F1 scores tend to converge near 1.0. This suggests that while the model can reach nearly perfect precision and recall at high confidence levels, setting the threshold too high may reduce the number of predictions made. A balanced threshold (around 0.4-0.5) appears to be optimal.