COMP3850 Group 23 Testing Document

Revision History

Revision Number	Date	Person(s)	Changes
1.0	29/04/2024	Group 23	Initial version
1.1	16/05/2024	Michael Yee	Added Pulling Chain Performance Evaluation Results section for models trained only on ants in pulling chains

Test Specification

Model Evaluation

To evaluate the performance of our project's output and determine which model has the best performance, we will be analysing each model using Object Keypoint Similarity (OKS) metrics, Percentage of Correct Keypoints (PCK) metrics and producing a Confusion Matrix to compare the models' relative strengths and weaknesses.

The OKS score is measured based on the similarity between the predicted and ground truth keypoints and is calculated using the Euclidean distance (the length of the line segment between two points) and a scale factor related to the weaver ant part size. The resulting value ranges from 0 to 1, with 1 indicating a perfect match and 0 indicating no overlap.

PCK is used as an accuracy metric that measures whether the predicted keypoint and the true joint are within a certain distance threshold. The PCK is usually set with respect to the scale of the subject, which is enclosed within the bounding box. For example, the threshold could be:

- PCKh@0.5 is when the threshold = 50% of the head bone link
- PCK@0.2 = Distance between predicted and true joint < 0.2 * torso diameter
- This alleviates the shorter limb problem since shorter limbs have smaller torsos and head bone links.

Confusion matrix data will provide information about where each model has difficulties by mapping the model predictions to the original classes to which the data belong - giving an accuracy metric of true and false positives and negatives and identifying where each model generates less accurate predictions.

For our model evaluation process, we are using a combination of SLEAP application software and our own Project Evaluation Code, written in Python and executed using Jupyter Notebooks. In our project, we will be evaluating models based on varying the number of filters and the maximum stride using metrics produced by training labelled input, as well as testing different anchor parts and how they impact the accuracy metrics.

The SLEAP application can be used to label a high volume of weaver ant videos as input using a shared skeleton definition, and then train models using these labelled videos, producing metrics packages. As part of the training process for each model, SLEAP calculates metrics relating to instance and part location accuracy and part visibility accuracy. SLEAP metrics also include the average Euclidean distances between predicted and actual part positions, the confusion matrix values for the part visibility predictions, and the Object Keypoint Similarity and Percent of Correct Keypoint scores.

Our Project Evaluation Code uses the SLEAP generated metrics packages as input and produces results which we will be using to demonstrate model comparisons, as a foundation for reasoning as to the best performing model produced as part of this project. The Project Evaluation Code leverages SLEAP's Python library in order to visualise OKS precision comparisons and to define "match and recall" thresholds. This allows us to factor in difficulties such as isolating and locating weaver ant instances that are directly related to pulling-chain events, and also the size of the weaver ant which impacts the constrained accuracy ratios.

We are also employing mean Average Precision (mAP) and mean Average Recall (mAR) metrics for the evaluation of keypoint visibility prediction accuracy.

Mean Average Precision (mAP) is a metric often used to evaluate object detection models. The mean of average precision is calculated over recall results (see below for recall calculation details) from 0 to 1. We are considering mAP metrics as they encapsulate the tradeoff between precision and recall and maximise the effect of both metrics.

Precision is the ratio of True Positive and the total number of predicted positives, which can be calculated using the following formula:

```
Precision = True Positives / (True Positives + False Positives)
```

Recall measures how well the model can find true positives out of all predictions and is calculated by the following formula:

```
Recall = True Positives / (True Positives + False Negatives)
```

The mAP formula is therefore based on the following sub-metrics:

- Confusion Matrix
- Intersection over Union (IoU) - a higher IoU value indicates a better alignment between the predicted and actual regions, reflecting a more accurate model. It measures the overlap between predicted and ground truth regions and helps in quantifying alignment between predictions and reality.
- Recall
- Precision

The calculation of the Mean Average Recall (mAR) is similar to the mAP calculation. Instead of analysing precision versus recall, we analyse the recall behaviour using different IoU thresholds. Average Recall is the recall averaged over all IoU between 0.5 and 1.0 and can be calculated as two times the area under the recall-IoU curve:

$$AR = 2 \int_{0.5}^{1} recall(o) do$$

All of these metrics are calculated by SLEAP's Python library which is publicly available.

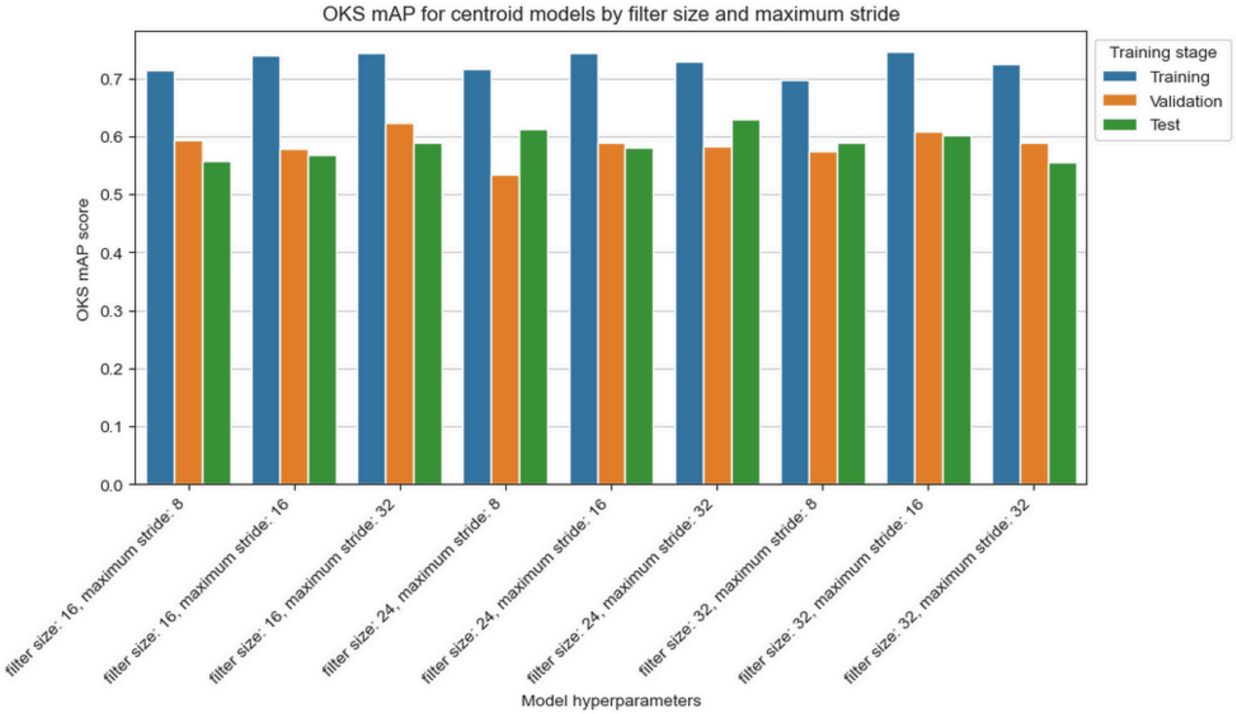
General Performance Evaluation Results

A SLEAP dataset consisting of 81 labelled frames which contained a total of 1,581 labelled ant instances was split into training, validation and test data in 3:1:1 split ratio respectively, which was used for training each model in the following analysis. All models were trained using the same training, validation and test data splits to remove this as a source of variance. The labelled ant instances consisted of ants engaging in different behaviours, including moving around and pulling on the leaf tip in the experiment environment. All hyperparameter values aside from the variables of interest were set at their default values and all models were trained until the validation loss consistently failed to improve, at which point the best performing versions of each model were saved. Future testing uses cross-validation to develop a more robust view of model performance.

Different models were trained with varying values for the filter size and maximum stride size as these were considered to be two of the most relevant hyperparameters for the UNet backbone used in this project and the long training time required for each model precluded testing most other hyperparameters in the available time. Separately, several models were trained using different choices of anchor parts, including no anchor part at all. The anchor part refers to the ant body part used as the reference point for the positions of all other body parts, which can help improve model accuracy and avoid the issue of body parts being predicted at physiologically unlikely locations.

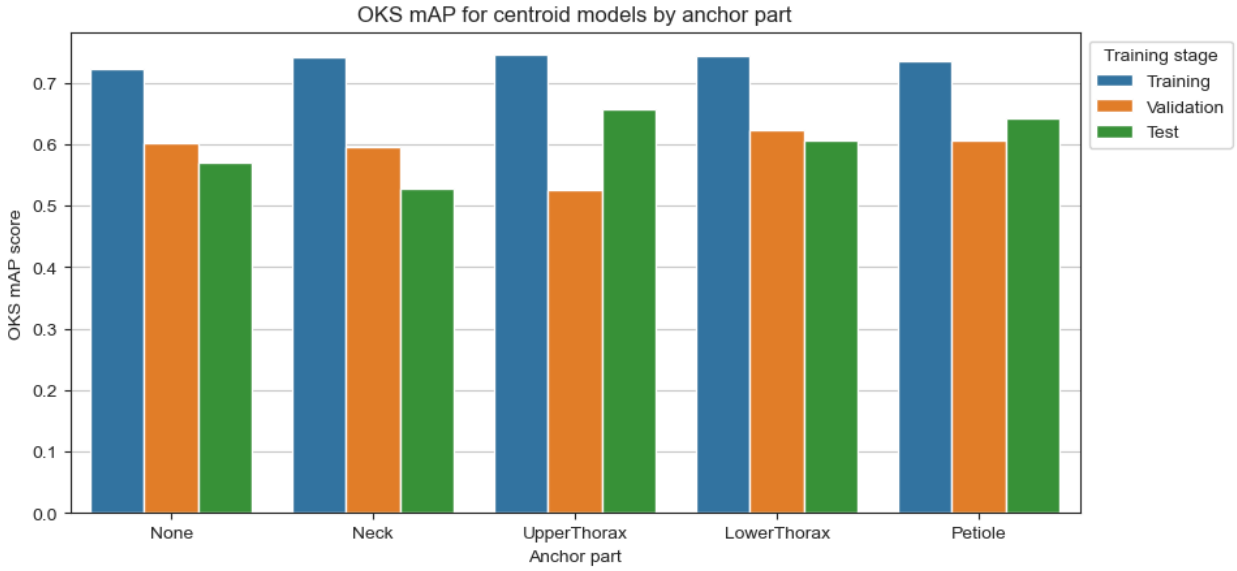
Centroid Location Accuracy

By filter size and maximum stride



From this graph, we can observe that the highest OKS mAP score for the Validation stage was obtained for the model using a filter size of 16 and a maximum stride of 32, and the highest score for the Test stage was obtained with the model with a filter size of 24 and a maximum stride of 32. In all of the combinations that were run, the performance during the Training stage was higher or similar to the Validation and Test stages, which is typical as models are tuned on the training data, but this is also indicative of overfitting, which likely means more training data are required. The model with filter size 32 and maximum stride 16 performs consistently well across all stages. This indicates it has a good balance between model complexity and performance. Overall, if the objective is to improve the centroid location accuracy in unseen data we should focus on models with a filter size of 32 and a stride of 16 based on this data.

By anchor part

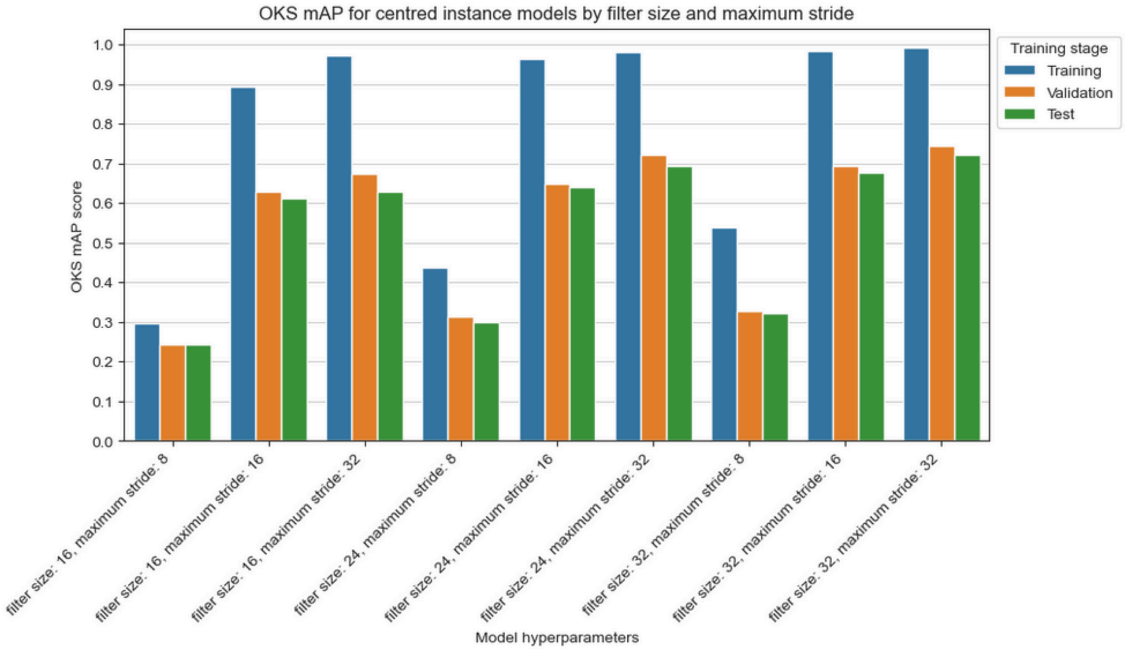


From this graph, it can be seen that the model that does not use any anchor part ("None") has a lower OKS mAP score in the Test stage compared to models that use a centrally located body part as the anchor part. This would indicate that using a centrally positioned anchor part improves the accuracy of centroid location estimation. Notably, when the Neck is used as the anchor part, the model performs worse in most cases except for the Validation stage of the model using the Upper Thorax as the anchor part. This reinforces that the anchor part chosen should be roughly towards the centre of the ant's body. The Upper Thorax anchor part has a lower performance than the Neck in the Validation stage but outperforms the Neck in the Test stage. This could mean that using Upper Thorax may be better for new data, but the inconsistency between the Validation and Test OKS mAP scores may cause issues with inference accuracy in a Production setting.

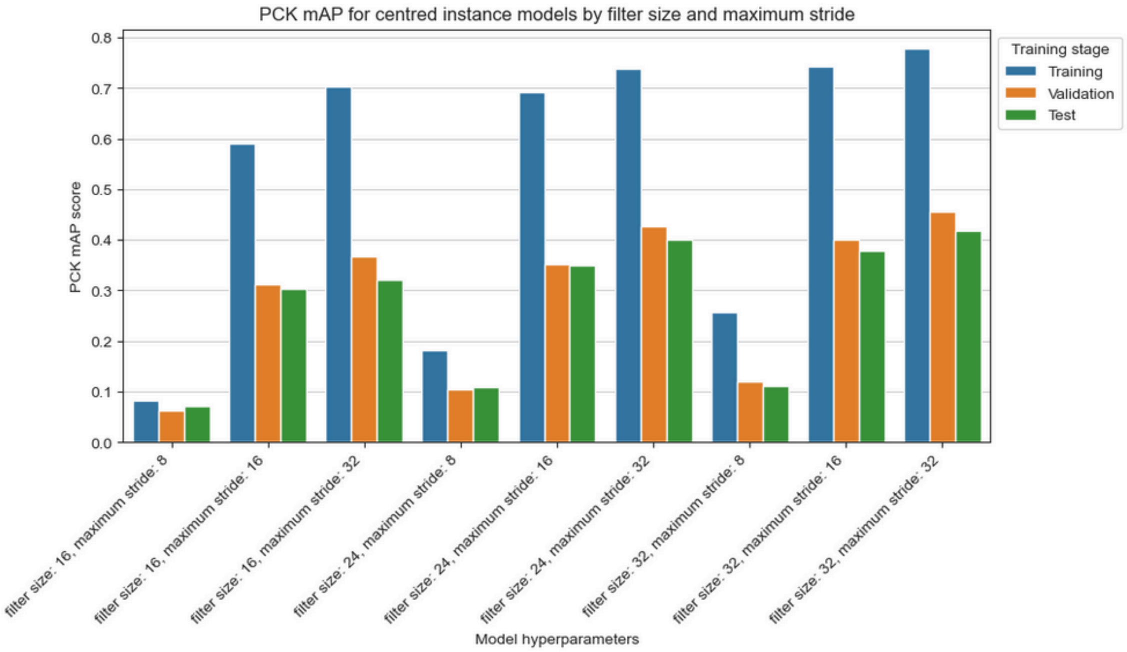
The models using the Lower Thorax and Petiole anchor parts perform nearly identically across all stages and are the second-best and best respectively in terms of overall performance. These results show that the choice of anchor part can significantly impact how well the model runs from training to unseen data and using the Petiole or the Lower Thorax as the anchor part would produce the most accurate results by OKS mAP score.

Centred Instance Pose Accuracy

By filter size and maximum stride

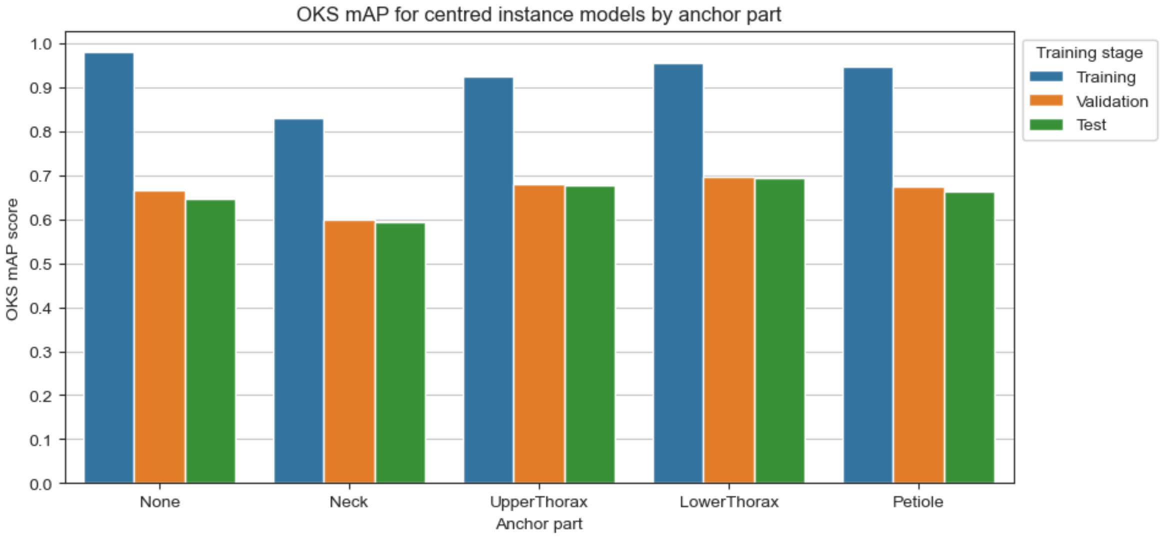


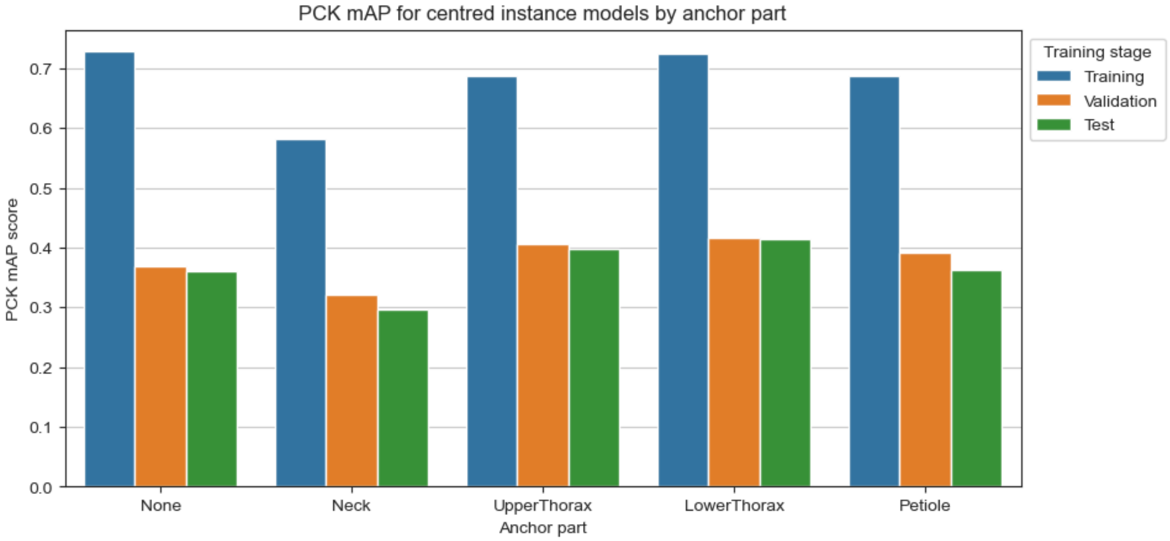
From this graph, it can be seen that models with a filter size of 32 and a maximum stride of 32 have the highest OKS mAP score in the training stage suggesting that they learn the training data very well. As before however, there is a significant drop in performance in the Validation and Test stages. This could be due to overfitting to the training data and hence more training data are required. The models with a maximum stride of 8 have the lowest OKS mAP scores across all stages, with an increase in the number of filters only slightly mitigating this decrease in accuracy. The graph suggests that for centred instance pose accuracy the choice of filter size and stride is crucial but the best-performing hyperparameters on training data may not always lead to the best generalisation on test data.



The combination of filter size 32 and stride 32 shows the highest PCK mAP score during the training stage. This indicates that this particular model configuration learns the training data very well. However, once again the performance of this configuration drops in the validation and test stages which may suggest that the model is overfitting to the training data. The configurations with a maximum stride size of 8 have a significantly lower PCK mAP score, especially with a filter size of 16. This could mean that the receptive field of the filter is too small to capture the necessary information for accurate pose estimation. Models with filter size 32 and stride 32 have the highest scores in the Test stage meaning they have a better ability to comprehend new data. Therefore, the best performing model for this particular dataset should use a filter size of 32 and a maximum stride size of 32 for the centred instance model.

By anchor part

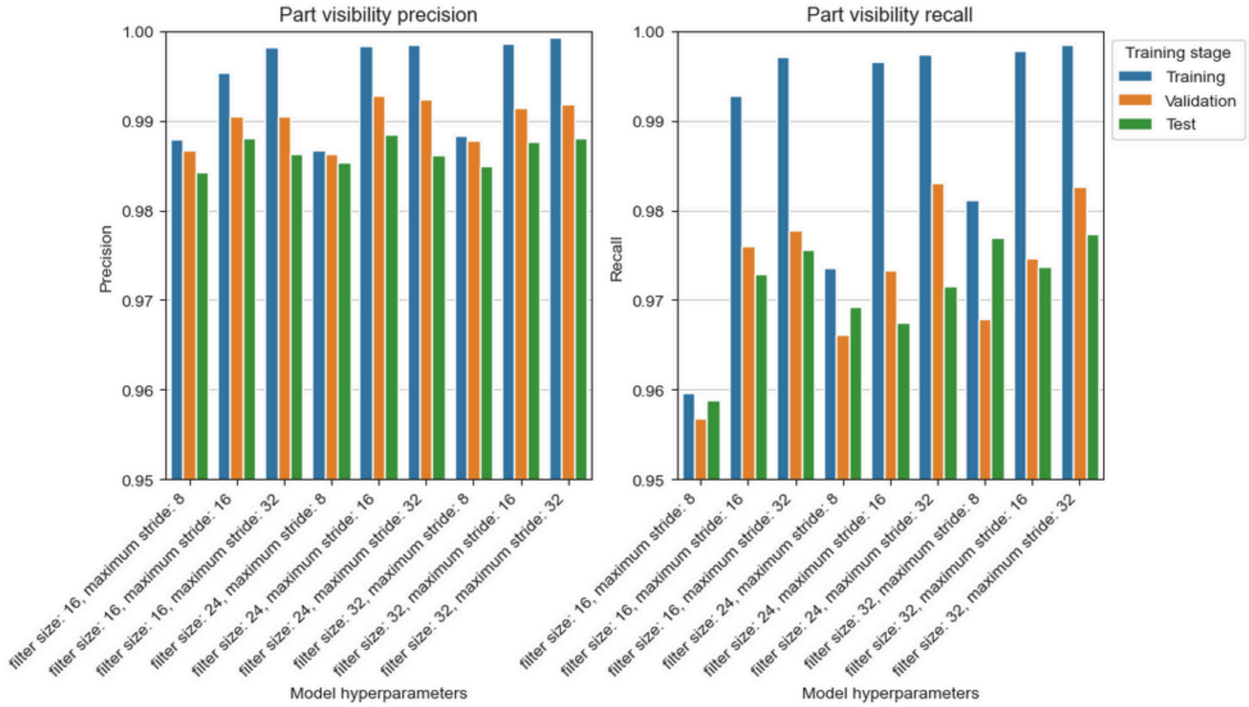




From the upper graph, it can be seen that all models perform exceptionally well in the Training stage but see a drop in performance in the Validation and Test stages further suggesting overfitting is occurring. Upper Thorax and Lower Thorax anchor part models display better performance on test data when measured by OKS mAP and PCK mAP scores. This means they perform better on new data when compared to using other anchor parts or no anchor part at all. Models using the Neck as the anchor part have the lowest performance, therefore we can deduce the importance of centrally located anchor parts for model accuracy. Lower Thorax appears to perform consistently across the Validation and Test stages but does not reach the high performance of the Training stage. In the lower graph, the Lower Thorax has the highest score in the Test stage. This indicates it is the most precise anchor for predicting key points in novel inputs according to the PCK mAP metric. Both None and Neck anchor part models have lower PCK mAP scores across all stages, while Upper Thorax and Lower Thorax anchor part models show higher PCK mAP scores in the validation and test stages compared to the previous two. Overall, we can see how different anchor parts significantly impact model performance for both OKS mAP and PCK mAP scores with the model using the Lower Thorax as the anchor point having the best overall performance across both metrics.

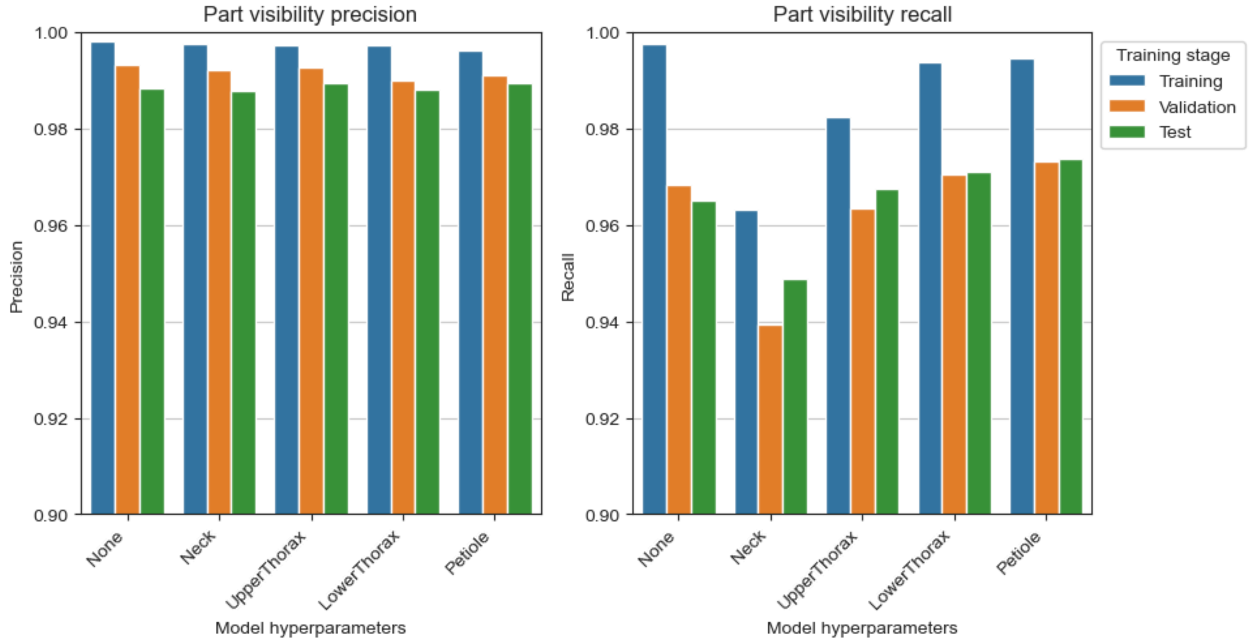
Part Visibility Accuracy

By filter size and maximum stride



The left graph shows that precision is quite high (>0.98) across all model parameters. The precision seems to decrease slightly for models with a lower stride size which may indicate that a smaller stride size leads to more false positives. The models with a filter size of 24 and stride 16 achieve the highest precision across all stages. However, the model with a filter size of 32 and stride 16 achieves the highest recall across the Training and Validation but drops in the Test stage. As with precision, recall generally increases with an increase in stride size. High precision with lower recall indicates that the model is very accurate when it predicts a part that is visible but it misses several actual visible parts. High recall with lower precision would mean the model identifies most of the visible parts but it also misidentifies some non-visible parts as visible. The models tend to perform better in the training stage than in the test stage. There is a trade-off present between precision and recall and it appears that filter size 32 and stride 32 offer a good balance for both metrics based on these graphs.

By anchor part



All of the anchor parts seem to provide high precision with scores close to 1 suggesting that when the model predicts a part is visible it is correct most of the time. Additionally, the difference in precision among the models using different anchor parts is minimal. This could mean that the choice of anchor part does not drastically affect the precision of visibility prediction. However, for the recall measures, the Neck category has the lowest recall showing the benefit of using a centrally located anchor part for visibility detection. All models have a noticeable drop in recall in the validation and test stages compared to the training stage due to potential overfitting. Overall, the use of central anchor parts significantly improves both the precision and recall of part visibility predictions compared to not using any anchor part. The high precision across all models suggests that when a part is predicted as visible it is most likely to be correct, regardless of the anchor part chosen. However, the model's performance in terms of recall varies more significantly, indicating differences in the model's confidence in identifying the visibility of the ant body parts.

Pulling Chain Performance Evaluation Results

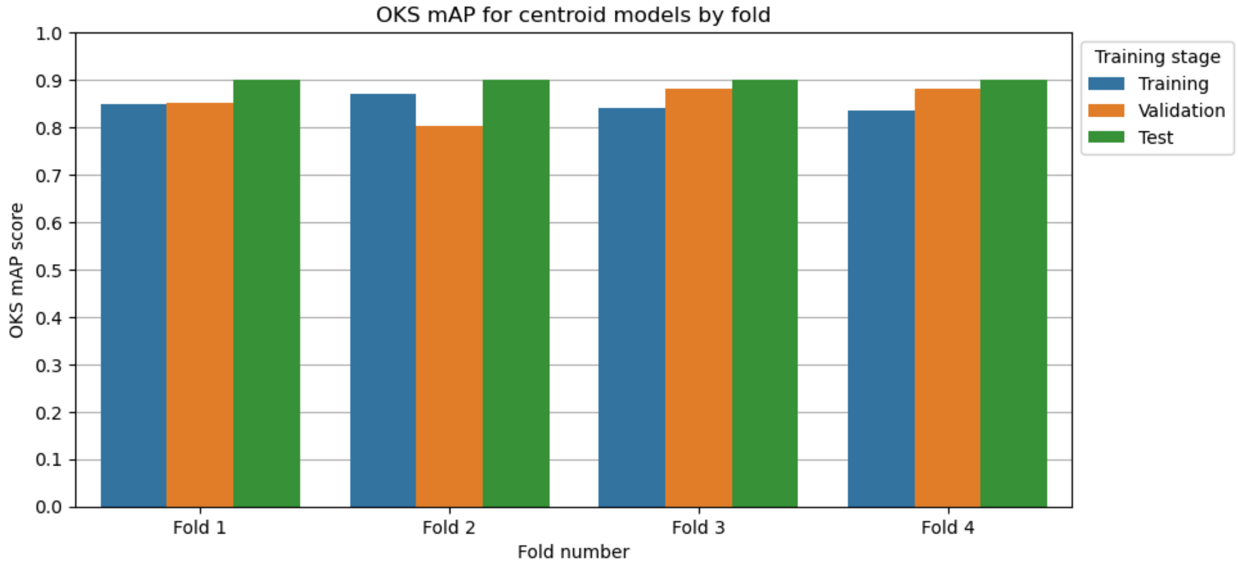
Following the initial testing, the client expressed an interest in seeing the results of training a pose estimation model focusing solely on single ants pulling on the leaf tips to assess the maximum possible accuracy that could be achieved in this scenario. A SLEAP dataset containing 104 labelled frames with one labelled ant instance per frame was used as the data source for the following tests. Only single ant instances that were pulling on the leaf tip in the experiment environment were chosen for labelling in this case, consistent with the client's request.

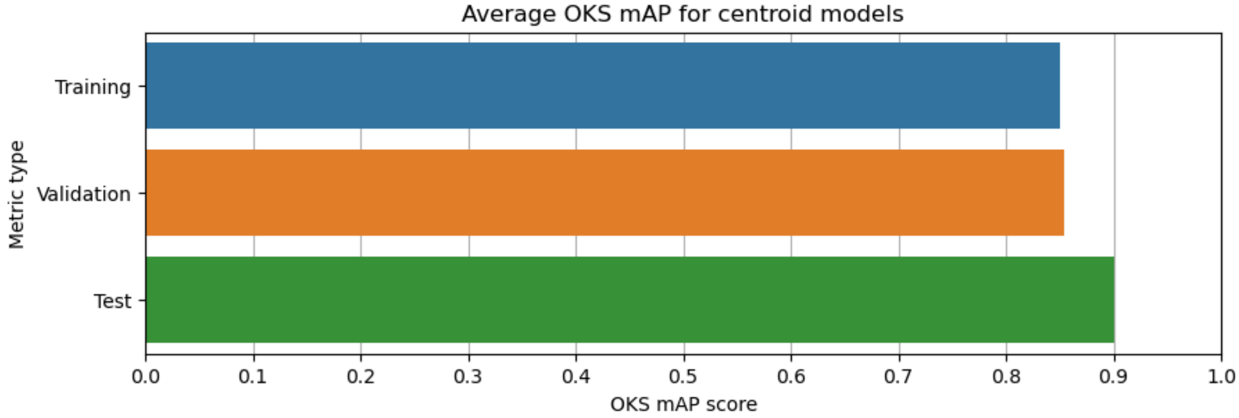
The dataset was randomly shuffled and a holdout fraction consisting of 20% of the labelling dataset was kept aside as a test dataset for validating the models after training. With the remaining frames in the dataset, a 4-fold cross-validation technique was used to split the dataset and train pose estimation models. The choice of a 4-fold split was primarily due to time constraints since previous experience showed that models could take approximately 30-60 minutes to train even on high-end hardware. The hyperparameters for the centroid and centred instance models were configured as follows based on the best performing models from previous testing, with all other hyperparameters set to their default values,

- Centroid model
    - Anchor part: LowerThorax
    - Maximum stride: 16
    - Number of filters: 32
- Centred instance model
    - Anchor part: LowerThorax
    - Maximum stride: 32
    - Number of filters: 32

Each model was trained until the validation loss failed to improve consistently and the performance metrics for each fold were averaged together to calculate the overall metrics for the training, validation, and testing stages of the model training process. The key metrics for evaluating model performance are the OKS mAP score, which is used to determine the accuracy of the centroid and body part location predictions, and the visibility precision and recall, which can be used to assess how accurately each model correctly identifies whether each body part for each instance is visible in the frame.

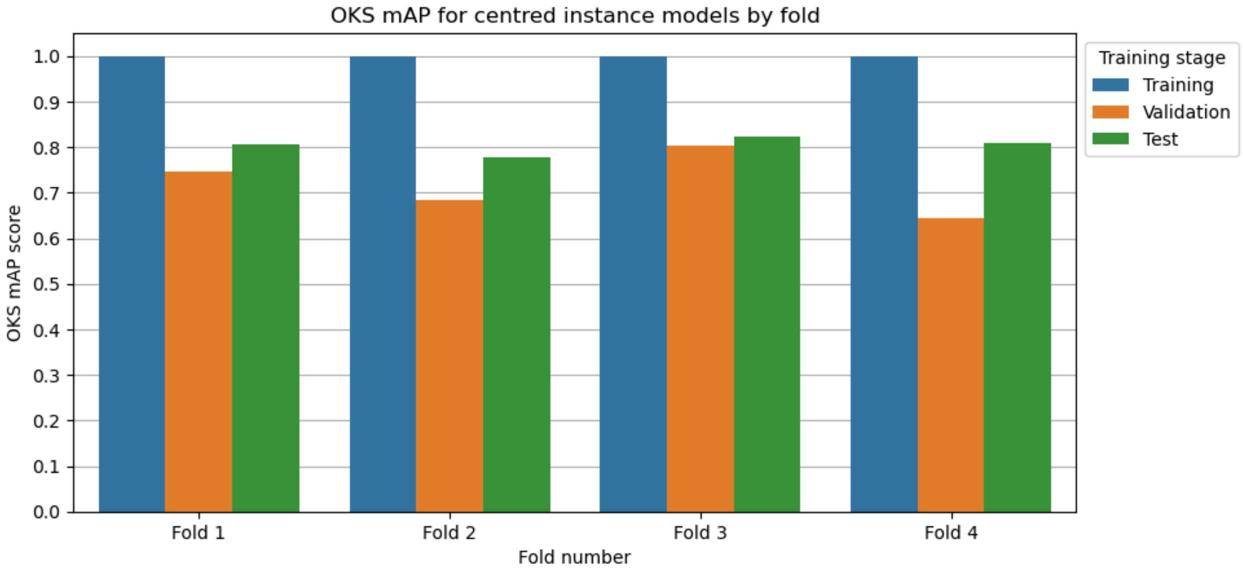
Centroid Location Accuracy

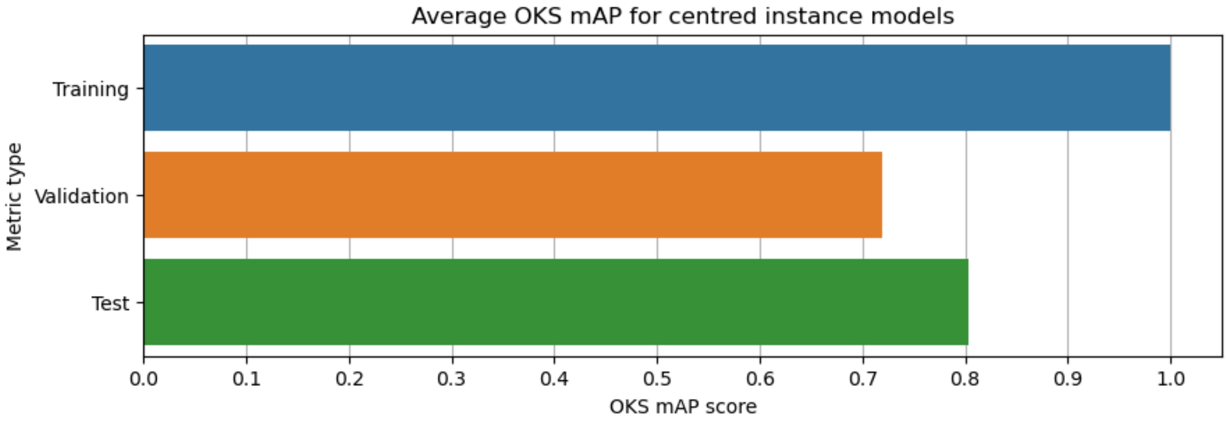




The average OKS mAP score for all of the centroid models is very high for each phase of the training process, with an average OKS mAP score of approximately 0.85 for the training and validation stages and 0.9 for the testing phase. The OKS mAP scores for each training phase in each fold are consistent with each other, with a difference of less than 0.05 from the average OKS mAP for each phase. This indicates that the effect of choosing different folds does not overly affect the model performance. Additionally, the difference in the average OKS mAP scores between the training and validation phases is very small, indicating that it is unlikely that overfitting is occurring. Overall, the high accuracy of the centroid models may be explained by the consistent locations of ant instances engaging in leaf pulling behaviour in the training videos, as the tip of the leaf that the ants are pulling on is always in the middle-left region of the frame.

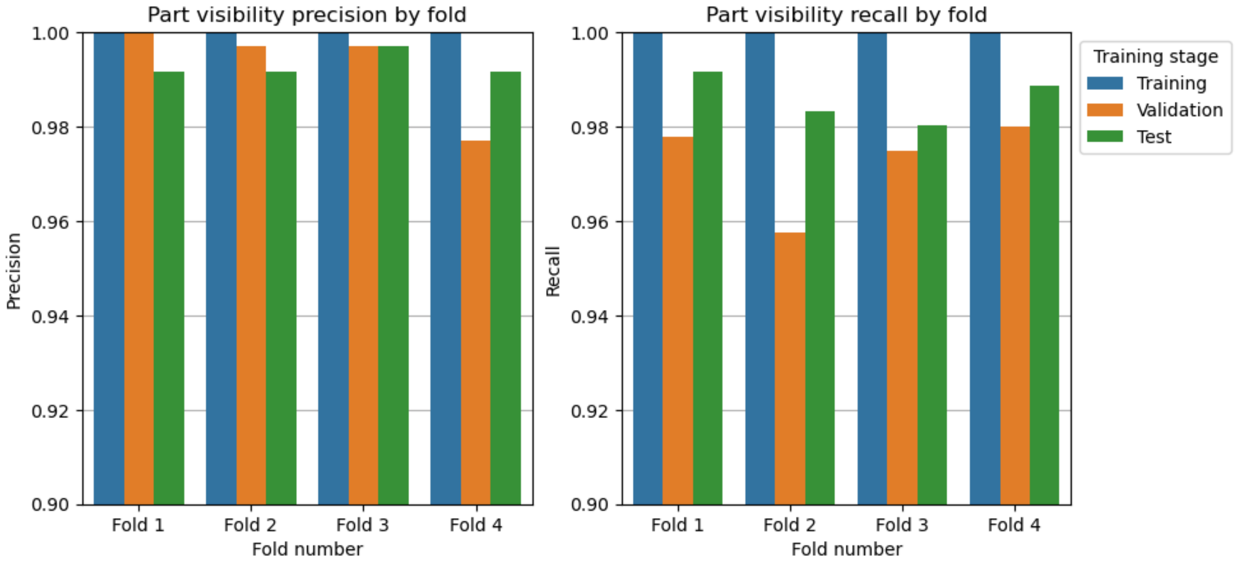
Centred Instance Pose Accuracy

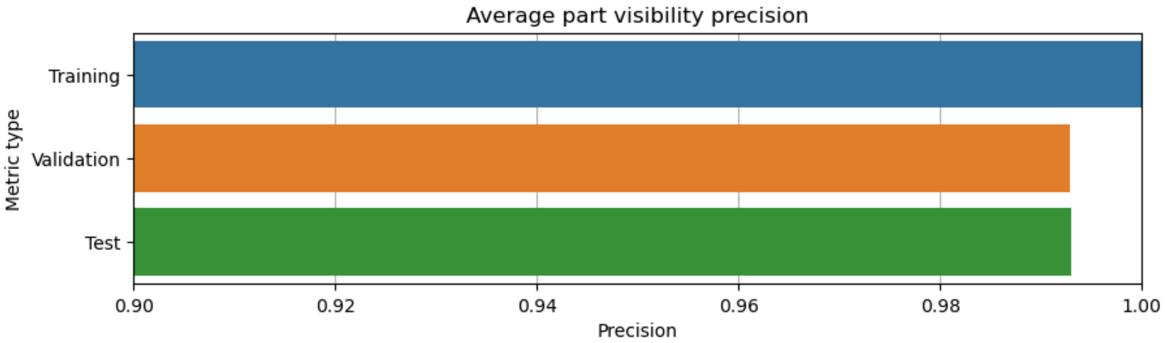


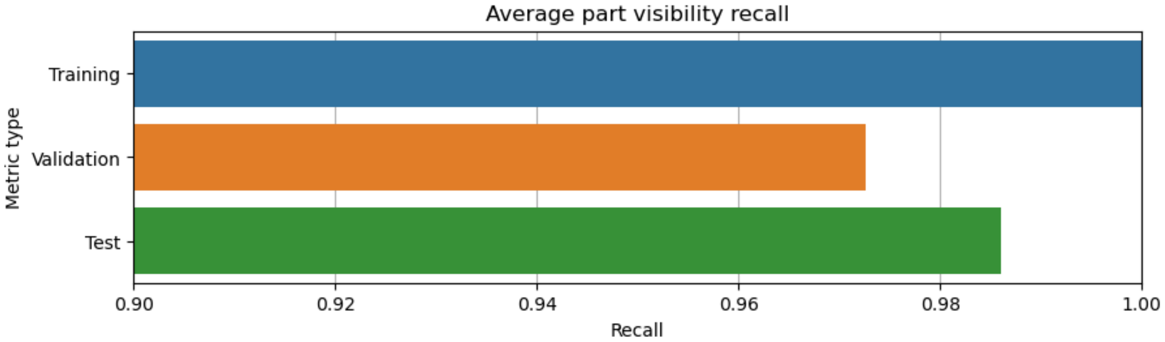


The average OKS mAP score for the centred instance models is somewhat lower than for the centroid model though still reasonably high, measuring approximately 0.72 for the validation stage and 0.8 for the testing stage. Notably, the OKS mAP score for the training phase of each fold was 1.0 which, when compared to the validation score, strongly indicates that overfitting is occurring. Given that this result occurred for all of the folds, it is highly likely that more training data are required to reduce overfitting and achieve more accurate inference, which would require more time to label.

Part Visibility Accuracy







The precision when determining the visibility of body parts in a frame is exceptionally high for each stage of the training process with >0.99 precision on average across all folds. This is consistent across all folds except for Fold 4, which may be due to how the split was generated. Overall, the high precision indicates that models trained using the training data for this test are unlikely to generate a spurious prediction for a body part that is not visible in the input video frame. The average recall is somewhat lower, with an average validation recall of approximately 0.97 and an average testing recall of approximately 0.985. This suggests that the models are slightly more cautious in generating a prediction when there is less indication that the body part is actually visible.