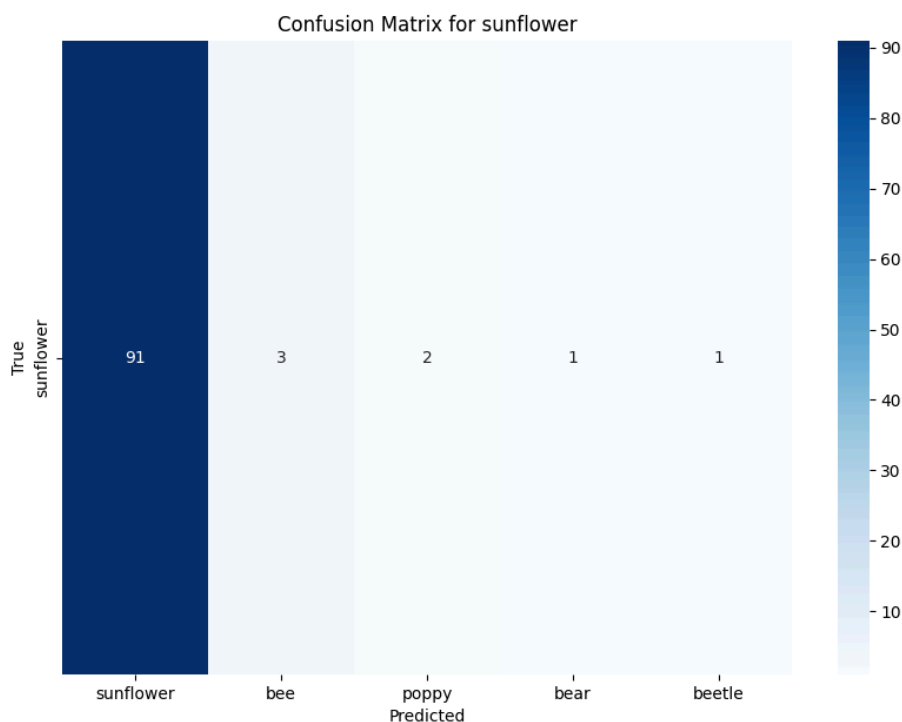
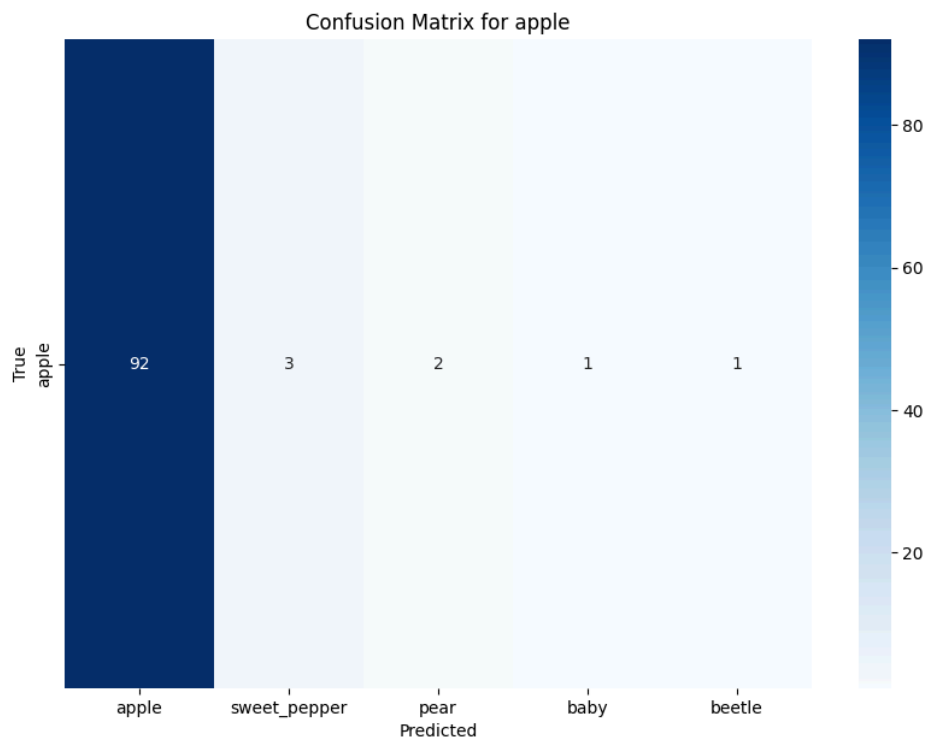


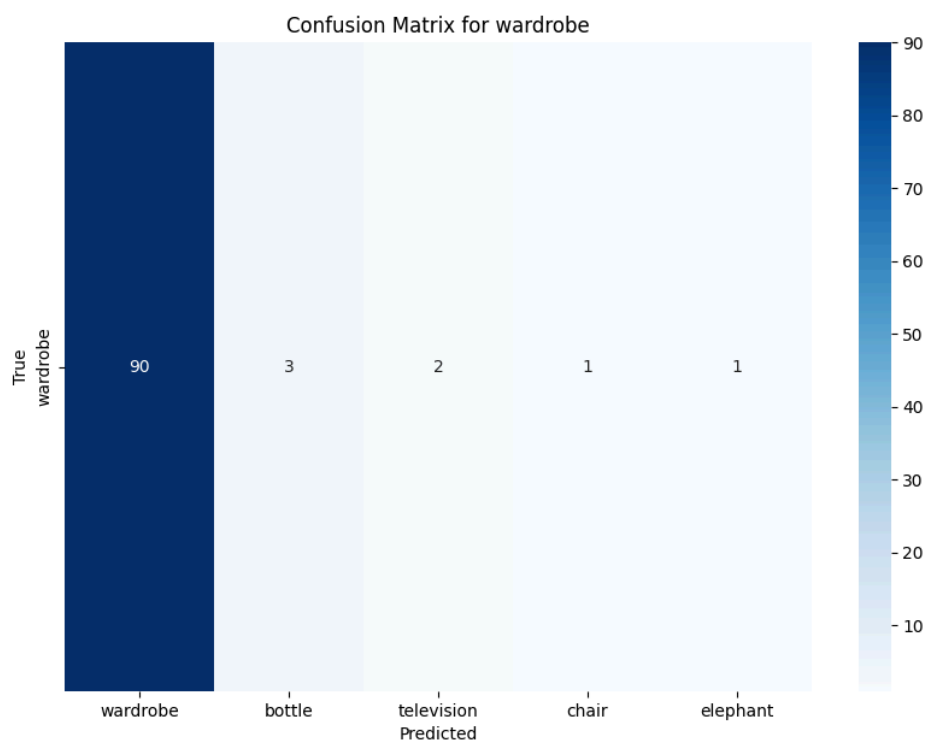
The best model - 66.30% prediction accuracy

Confusion matrices

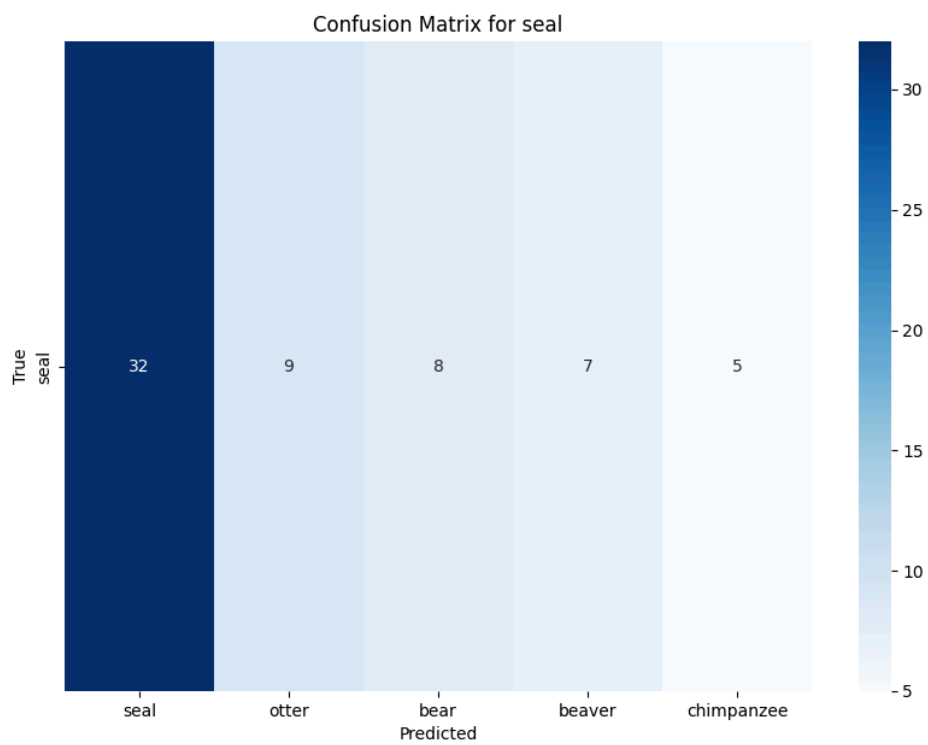
Confusion matrices for the three best predicted and three worst predicted classes are presented. The matrices show the 5 most frequently predicted classes by the model for a given class, so also FN cases.

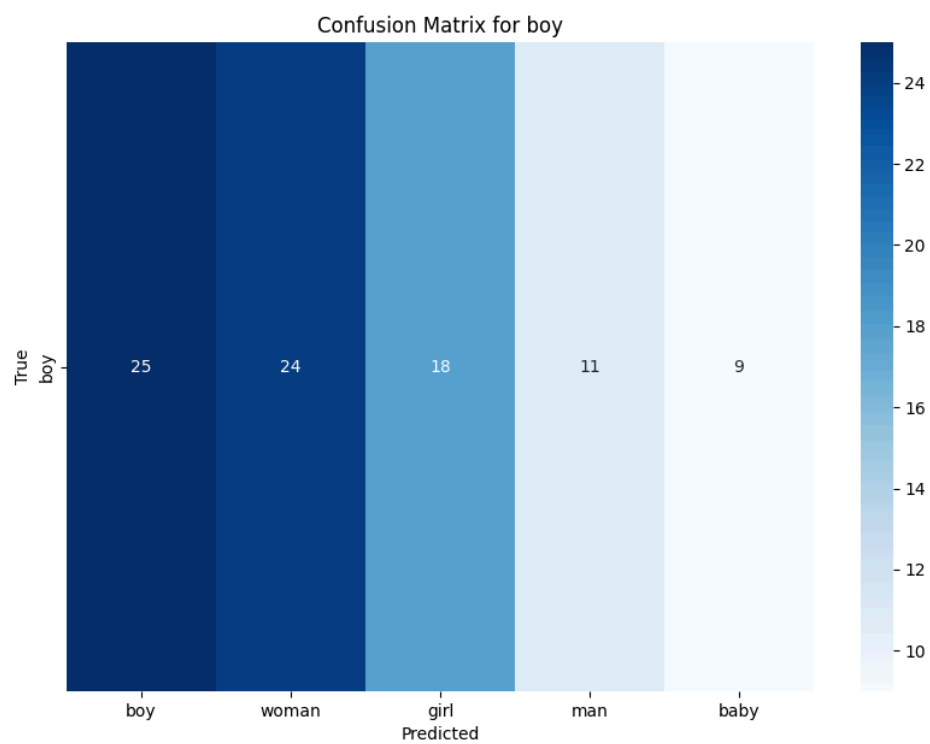
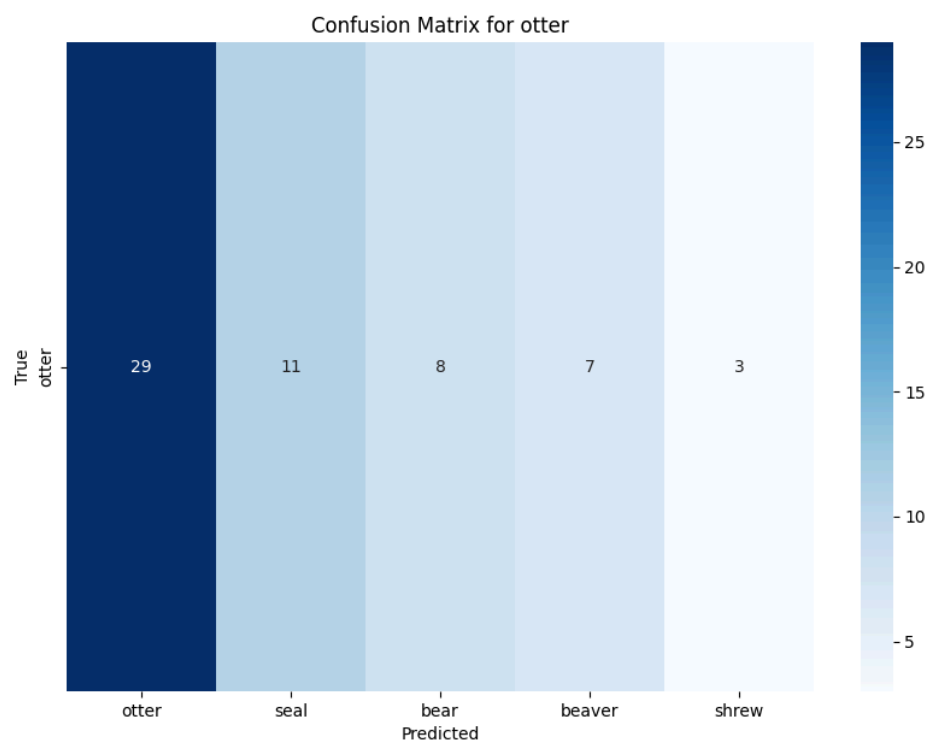
Top predicted classes: **Apple (92%)**, **Sunflower (91%)**, **Wardrobe (90%)**





Worst predicted classes: **Seal(32%), Otter(29%), Boy(25%)**





Evaluation metrics

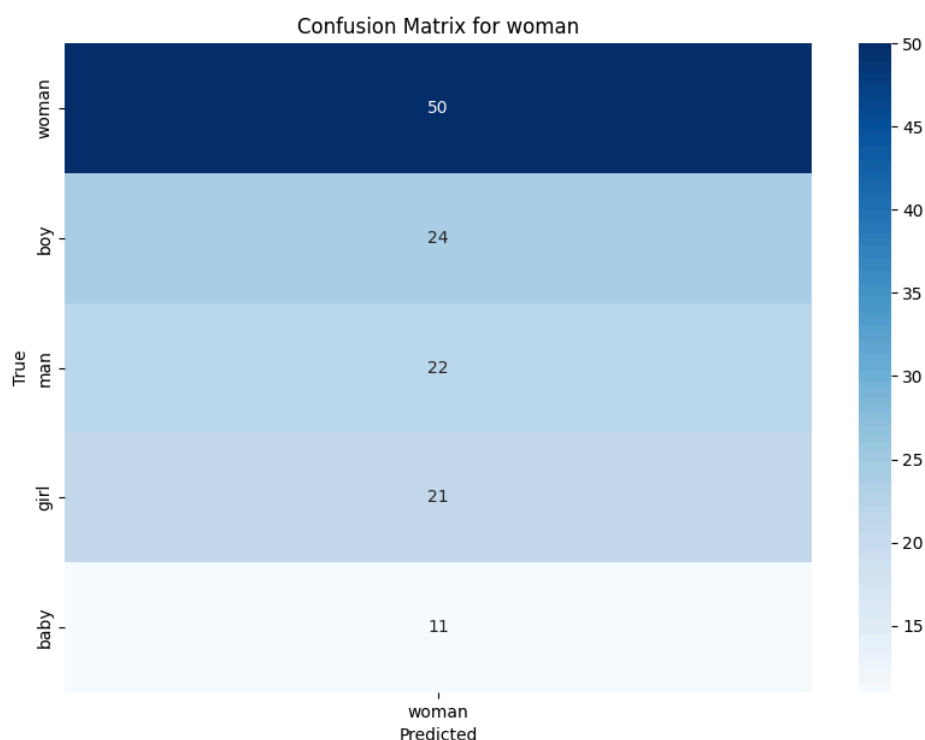
Precision Score - 0,675

$$Precision = \frac{TP}{TP+FP}$$

It determines how many of the cases predicted as positive are actually positive. To calculate precision, I used the weighted variant, i.e. precision was calculated for each class and weighted accordingly by the number of examples of a given class in the entire set.

Class with best precision: **Trout (0.94)**

The class with the least precision: **Woman (0.34)** - in the chart below you can see that the model often falsely predicts the woman class:



Recall - 0,663

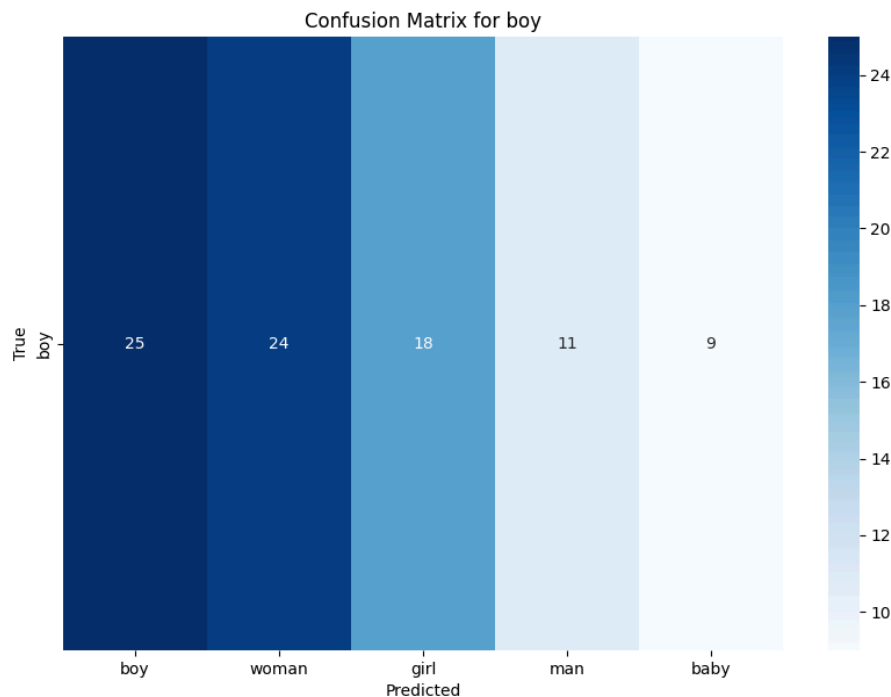
$$Recall = \frac{TP}{TP+FN}$$

It measures what proportion of all true positive cases the model detected. This is a key metric where missing a positive case could have serious consequences. I also used the weighted variant to calculate the recall.

Class with best recall: **Apple (0.92)**

Classes with the least recall: **Boy (0.26)** - as you can see in the chart below, for the Boy class the model predicts a lot of false negative (FN) classes

:



F1-Score - 0,665

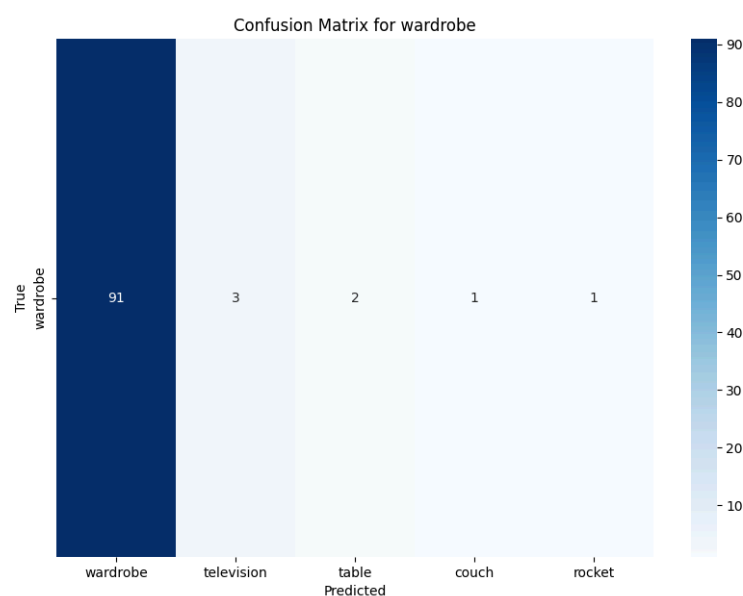
It is the harmonic mean of precision and sensitivity that takes both aspects into account. I also used the weighted variant to calculate it. F1-score is useful when we want to balance precision and recall, e.g. in situations where both false positives and false negatives are costly.

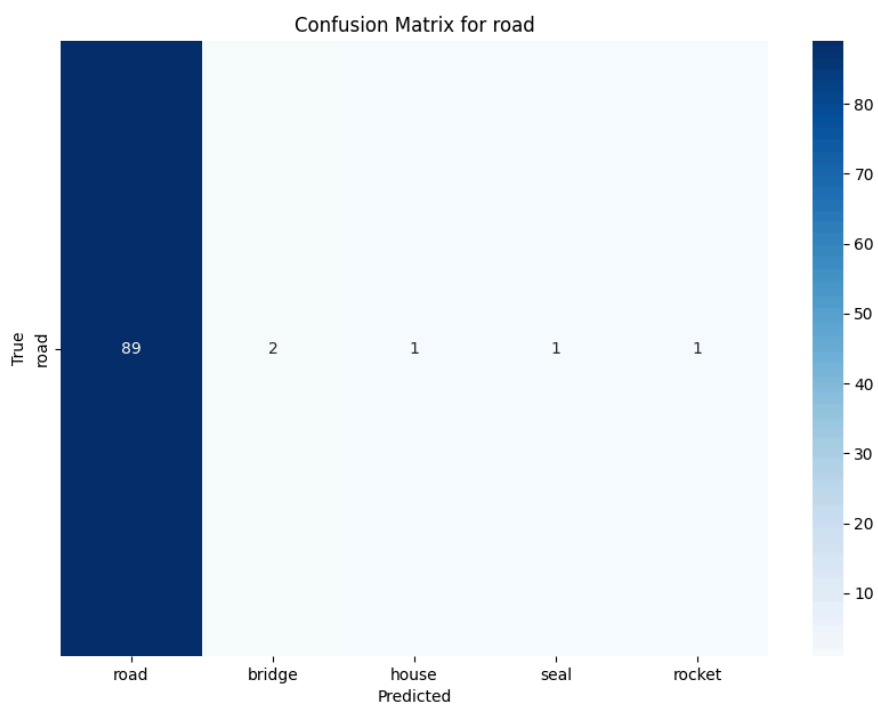
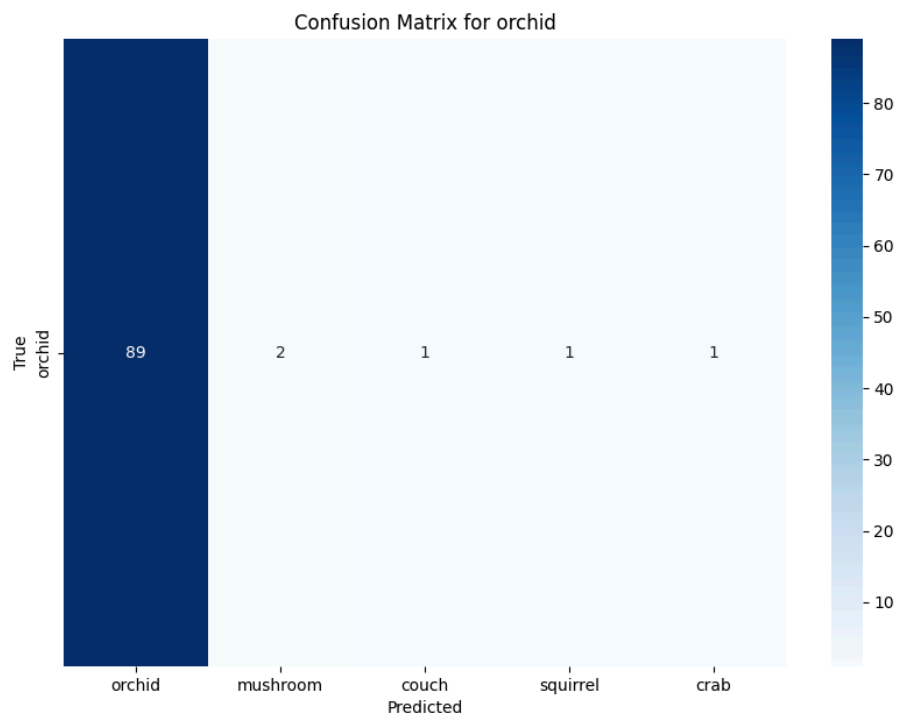
The second best model - 64.32% prediction accuracy

Confusion matrices

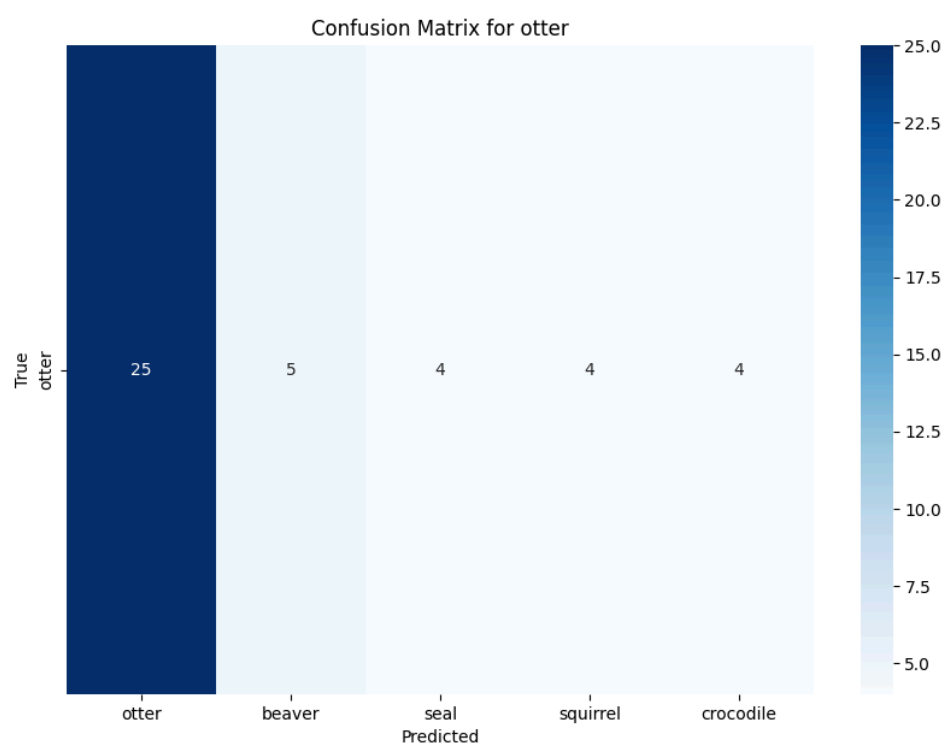
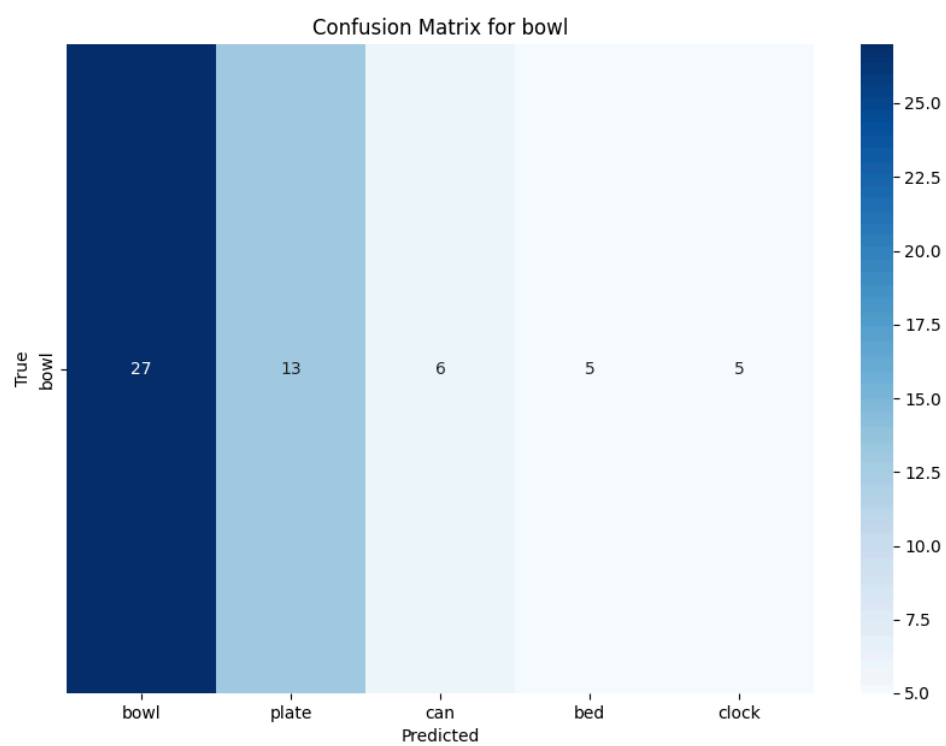
As before, confusion matrices are shown when a certain class was expected and the model classified the image to a different label. The matrices are shown for the three best predicted classes and the three worst predicted classes.

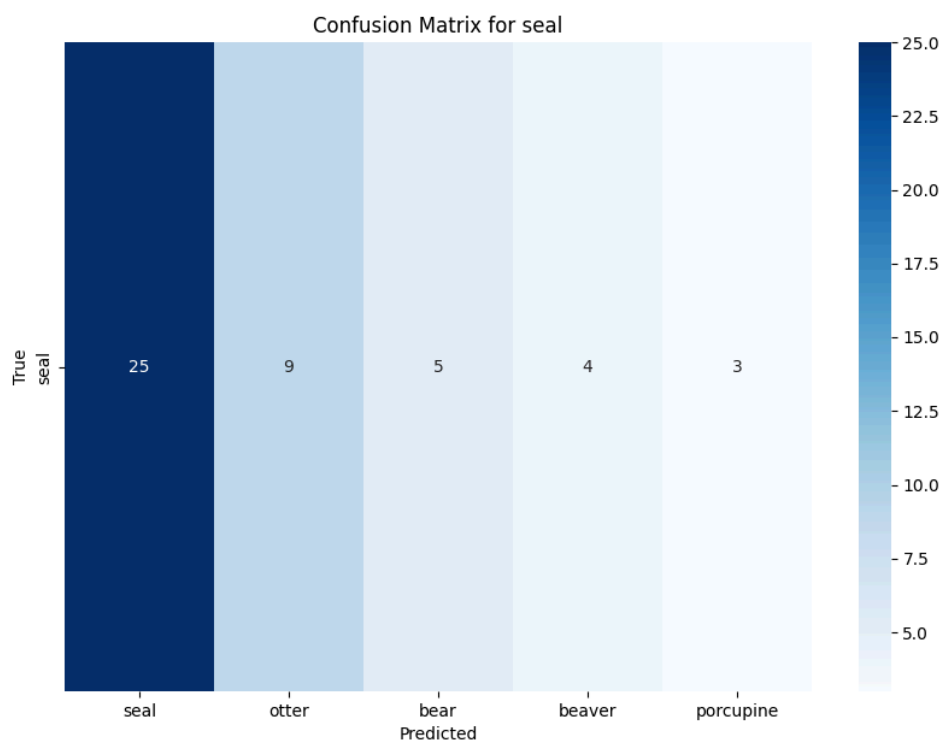
Best predicted classes: **Wardrobe (92%), Orchid(91%), Road(90%)**





Worst predicted classes: **Bowl(27%), Otter(25%), Seal(25%)**





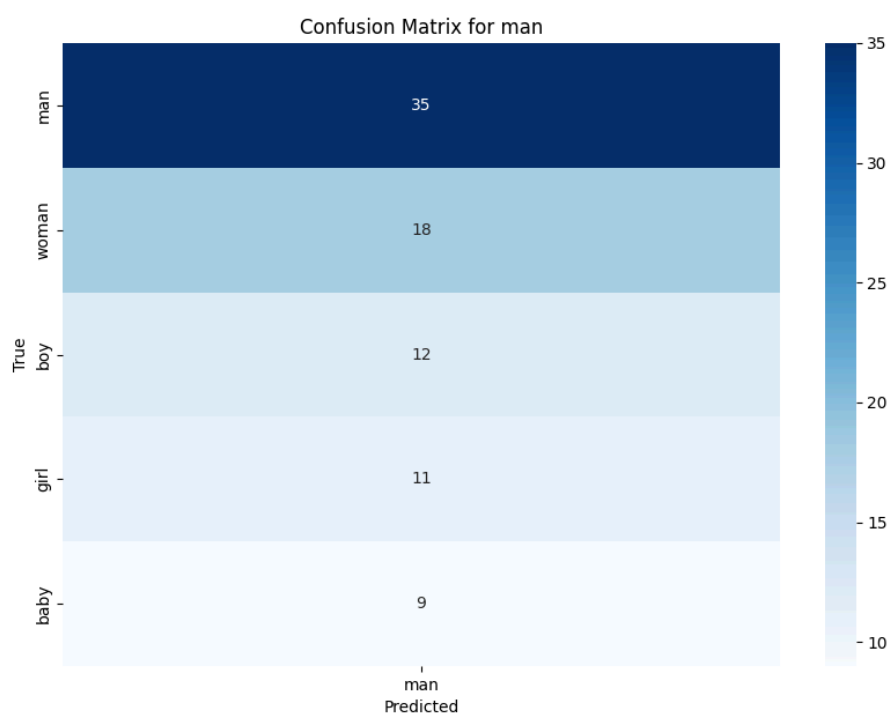
Evaluation metrics

Precision Score: 0.656

It was calculated in the same way as for the best model.

Class with best precision: skunk (0.90)

The class with the least precision: Man (0.27) - in the chart below you can see that the model often falsely predicts the Man class:

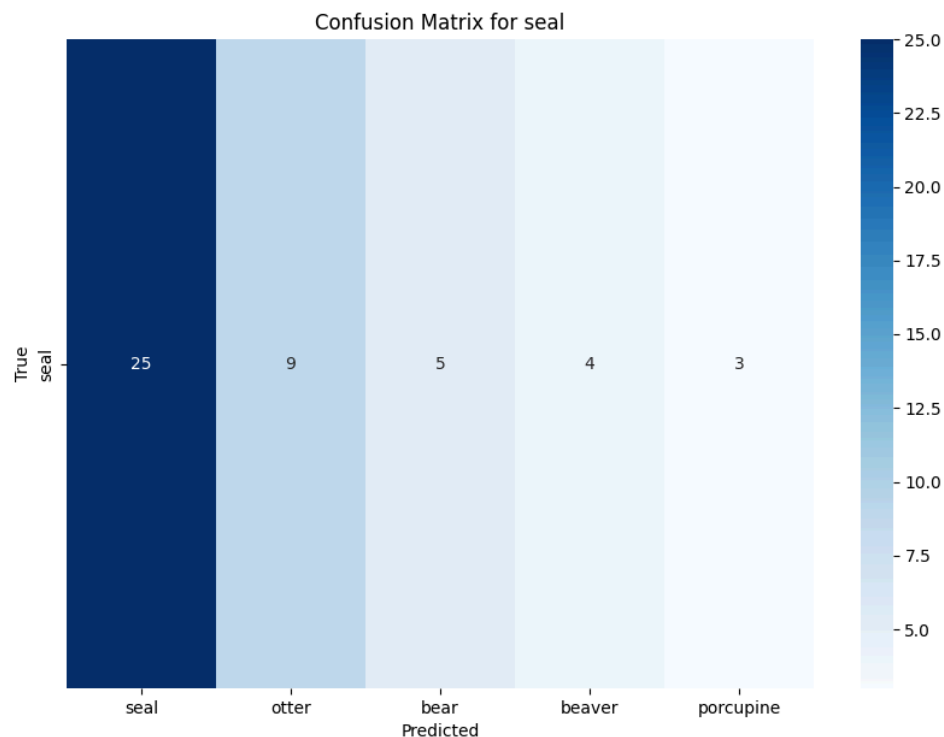


Recall: 0.643

Recall was calculated in the same way as for the best model.

Class with the best recall: Wardrobe (0.91)

Classes with the lowest recall: Seal (0.25) - as you can see in the chart below, for the Seal class the model predicts a lot of false negatives (FN) classes:

**F1-Score: 0.643**

The harmonic mean of precision and sensitivity for this model also came out similar to the precision and recall values. Therefore, the model achieves balanced performance and does not favor any of the measures, which is also due to well-balanced data, because each class in the CIFAR100 set has the same number of images.