Practice 2 Performance Report

Data Analysis

Before we start, we will analyze the data distribution of our dataset, both the splits distribution and per-class distributions.

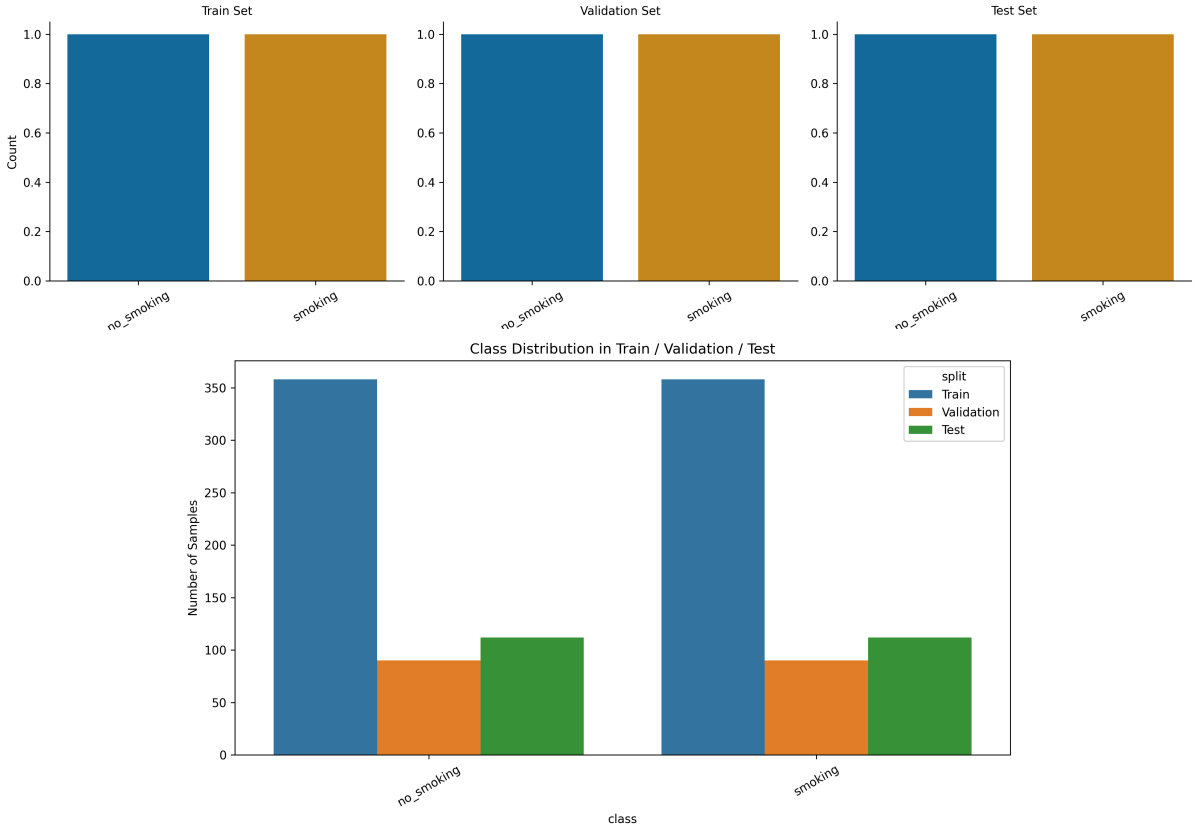

Figure 1: Class distribution between dataset splits

The dataset is **balanced**. Each split has the same ratio of **smoking** and **non-smoking** images in it. The dataset is divided in 80% for train and validation, and the rest 20% for test.

To each image of the dataset, we've manually added new data to it: **gender** and **category**:

- Genders: Man and Woman
- Categories: Random, Phone, Inhaler, Water and Cough

With this new information, we can visualize the distribution of the image set.

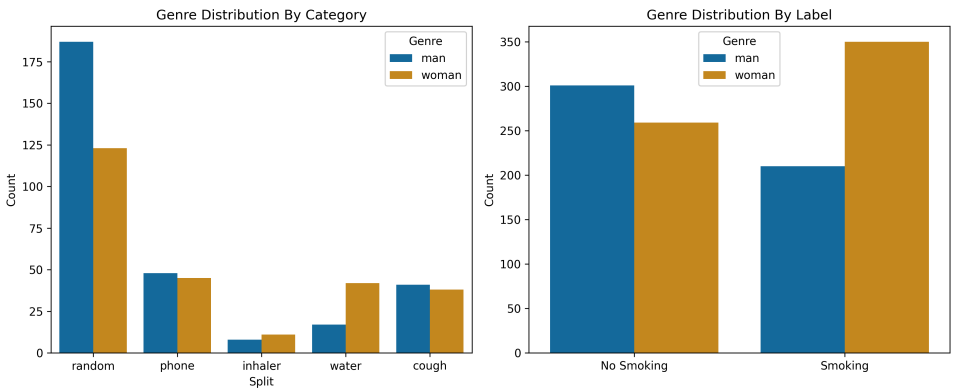

Figure 2: Gender distribution between categories and classes

We can see the ratio between **Man** and **Woman** in the image set is not balanced. There is a bigger women-to-men ratio in the **Smoking** image class in comparison to **Man**. Between the image categories, there is a higher amount of Man images without a clear action, while there are more Woman images of drinking water. Other categories are balanced.

Model training analysis

With 2-3 epochs we can already see the model converging to nearly 0. This result is very positive. However, when compared to the validation loss, we can see it plateaus at around 0.35. This divergence between training and validation loss suggests some degree of overfitting, which is common when fine-tuning pretrained models on smaller datasets.

Despite this, the validation accuracy performs well, starting at around 50% and quickly climbing to 87% by epoch 2, then gradually improving to just below 90% and stabilizing there.

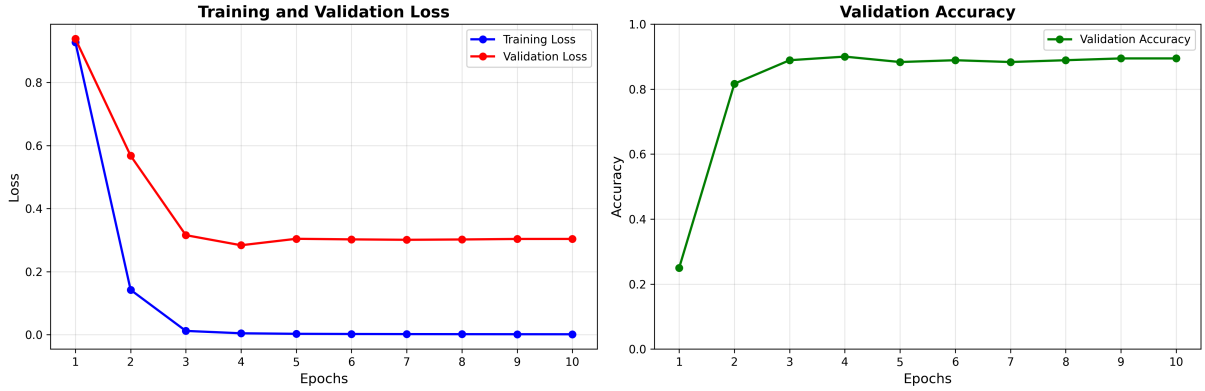

Figure 3: Loss and Accuracy curves

Model prediction analysis

Once the model has been trained, we can annalize in which categories the model tends to classify incorrectly, and see if the initial data distribution is affecting the final prediction capabilities.

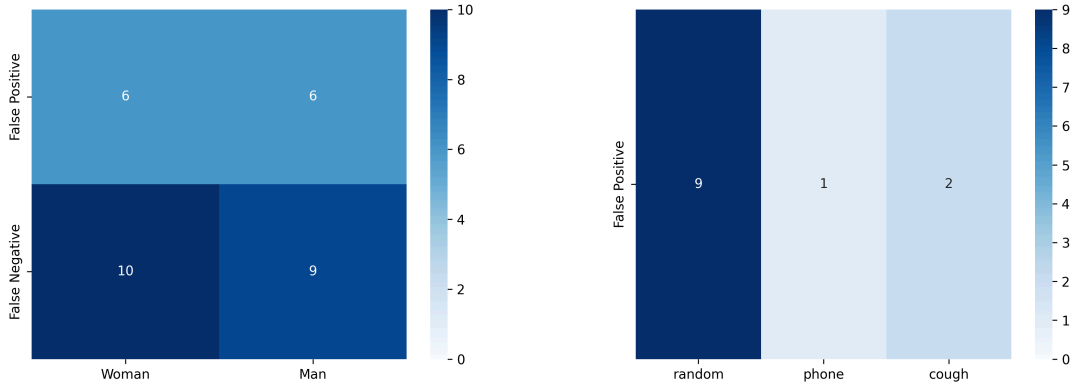

Figure 4: Distribution of wrongly predicted images

We can see that **False Negatives** (Smoking images classified as non-smoking) are more common that False Positives. The ratio of failure between genders is balanced, not showing a negative effect due to the unbalance shown in Figure 4. Between the test split wrong predictions, only 3 categories are present: Random, Phone and Cough. Most of the wrong predictions fall into the Random category, most likely due to more unpredictable person actions.

Figure 5: Wrongly classified smoking and non-smoking images

For the sample images in Figure 5, we can theorize they where wrongly classified for different reasons. For the smoking ones, the first ones cigarette partially blends in with the background building, making it harder to distinguish. The second one does not hold the cigarette with the hand, a common thing between most smoking images.

For the wrongly classified non-smoking images, the reason might not be as clear. We could theorize it is due to image composition and the persons placement. We saw how smoking images tended to have darker background colors, something partially visible in these images.

Model calibration

Once we have taken a look into a portion of the weaknesses of the model, to better assess the reliability, we can perform a model calibration analysis. The aim is to see how well the predicted probabilities of each class match the outcome. To do so, we aggregate along the X-axis the examples with similar predictions (in 10 bins in this case). Along the Y-axis, we represent the actual fraction of positives in each bin. The diagonal line represents perfect calibration.

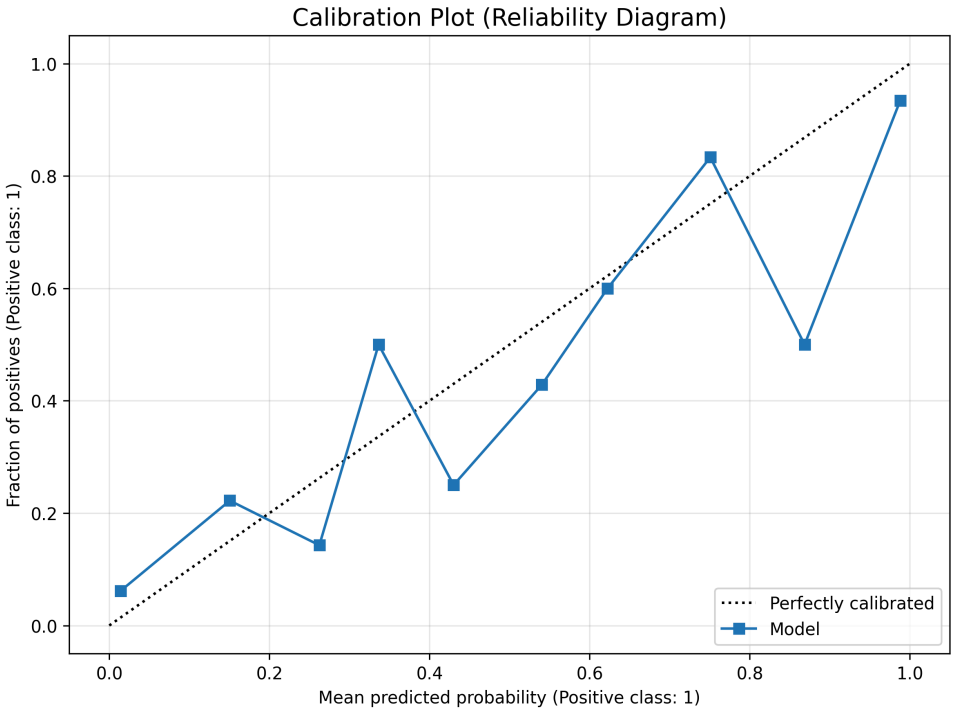

Figure 6: Model calibration curve

For this curve, we took the prediction of class 1 (smoker). The model is poorly calibrated overall. It is overconfident in the low-range predictions. Meanwhile, around the mid-range values, it seems to be underconfident in its predictions. The most concerning pattern is in the high-confidence region (0.7-1.0), where the model shows highly erratic behaviour. At 0.9, it drops to 0.5, which is concerning, as we would expect the accuracy to continue rising in line with the previous bins.