# Experiment 6: Optimizer & Learning Rate Tuning

In order to address the limitations that we have discern from the initial base run, this experiment intends to investigate the impact of training duration and learning rate adaptation on model generalisation. Specifically, in comparing the training for 30 epochs versus 50 epochs so as to identify if continued training leads to diminishing returns or exacerbates overfitting. In addition, we will be attempting to implement a learning rate scheduler to allow the model to begin with a higher learning rate for faster convergence and then gradually reduce it to refine learning in later epochs.

We also intend to incorporate early stopping criteria based on validation loss to terminate training once performance stagnates or deteriorates, ensuring computational efficiency and minimising overfitting. As such, with these modifications in place, the aim of this experiment is to enhance the model's ability to retain strong generalisation performance while avoiding unnecessary or harmful over-training.

### Experiment Parameters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Optimizer** | **Learning Rate** | **Batch Size** | **Scheduler Used** | **Epochs** |
| Adam | 0.001 | 32 | None | 15 |
| Adam | 0.001 | 32 | None | 30 |
| Adam | 0.001 | 32 | None | 50 |

### Performance Metrics Comparison

|  |  |  |
| --- | --- | --- |
| **Metric** | **Initial Baseline Run (Epoch = 15)** | **Tuned Run #1 (Epoch =30)** |
| Train Accuracy | 0.94 |  |
| Validation Accuracy | 0.89 |  |
| Test Accuracy | 0.90 |  |
|  |  |  |
| Apple F1 Score | 0.92 |  |
| Banana F1 Score | 0.92 |  |
| Orange F1 Score | 0.94 |  |
|  |  |  |
| Mixed Precision | 0.57 |  |
| Mixed Recall | 0.80 |  |
| Mixed F1 Score | 0.67 |  |
|  |  |  |
| Macro Average F1 Score | 0.86 |  |
| Weighted Average F1 Score | 0.90 |  |

### Observations & Insights

##### Tuned Run #1:

* Both runs achieved identical train (94%) and test (90%) accuracy, indicating consistent model capacity across learning rates.
* The baseline run (LR = 0.001) slightly outperformed the tuned run (LR = 0.0005) in validation accuracy (89% vs 87%), suggesting better generalization within the same epoch budget.
* Overall, for Apple, Banana and Orange classes, there is a slight notable improvement in F1 scores
* Whereas for the Mixed class, the baseline had a higher F1 score (0.67 vs 0.60), with a notably higher recall (0.80 vs 0.60), suggesting better sensitivity to true positives.
* **Slightly lower macro F1 score (0.85 vs 0.86)** indicates that the model's performance **varied more across different classes**, particularly underperforming in the "Mixed" class.
* The model may have been **more cautious** (higher precision) but **less sensitive** (lower recall) to detecting certain classes.
* This suggests that the **tuned run prioritized precision over recall**, which may be suitable in contexts where false positives are more costly than false negatives.
* It could also imply that the **learning rate was too small** for the model to fully optimize its performance across all classes within the same training duration.

### Final Recommendations