# Experiment 6: Training Schedule

In order to address the limitations that have been discerned 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 an 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** | **Early Stopping Used** | **Epochs** |
| Adam | 0.001 | 32 | None | None | 15 |
| Adam | 0.001 | 32 | None | None | 30 |
| Adam | 0.001 | 32 | ReduceLROnPlateau(factor=0.1, patience=3, min\_lr=1e-6) | None | 30 |
| Adam | 0.001 | 32 | ReduceLROnPlateau(factor=0.1, patience=3, min\_lr=1e-6) | None | 50 |
| Adam | 0.001 | 32 | ReduceLROnPlateau(factor=0.1, patience=3, min\_lr=1e-6) | Early Stopping(patience=5 delta=1e-4) | Up to 50 (may stop early) |
| Adam | 0.001 | 32 | ReduceLROnPlateau(factor=0.1, patience=3, min\_lr=1e-6) | Early Stopping(patience=8 delta=1e-4) | Up to 50 (may stop early) |

### Performance Metrics Comparison

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Metric** | **Initial Baseline Run (Epoch=15)** | **Tuned Run #1 (Epoch=30)** | **Tuned Run #2**  **(Epoch=30 with Scheduler)** | **Tuned Run #3**  **(Epoch=50 with Scheduler)** | **Tuned Run #4**  **(Epoch=50 with Scheduler and Early Stopping [Patience=5])** | **Tuned Run #5**  **(Epoch=50 with Scheduler and Early Stopping[Patience=8])** |
| Train Accuracy | 0.91 | 0.97 | 0.97 | 0.97 | 0.92 |  |
| Validation Accuracy | 0.86 | 0.83 | 0.88 | 0.87 | 0.85 |  |
| Test Accuracy | 0.87 | 0.88 | 0.87 | 0.87 | 0.87 |  |
|  |  |  |  |  |  |  |
| Apple F1 Score | 0.92 | 0.92 | 0.92 | 0.89 | 0.87 |  |
| Banana F1 Score | 0.89 | 0.88 | 0.88 | 0.88 | 0.88 |  |
| Orange F1 Score | 0.89 | 0.94 | 0.89 | 0.91 | 0.92 |  |
|  |  |  |  |  |  |  |
| Mixed Precision | 0.50 | 0.50 | 0.50 | 0.50 | 0.60 |  |
| Mixed Recall | 0.60 | 0.80 | 0.80 | 0.80 | 0.60 |  |
| Mixed F1 Score | 0.55 | 0.62 | 0.62 | 0.62 | 0.60 |  |
|  |  |  |  |  |  |  |
| Macro Average F1 Score | 0.81 | 0.87 | 0.83 | 0.83 | 0.82 |  |
| Weighted Average F1 Score | 0.87 | 0.88 | 0.87 | 0.87 | 0.87 |  |

### Observations & Insights

### Tuned Run #1

* In this run of 30 epochs, it had led to a slight increase in overall test accuracy, rising from 87% (baseline) to 88%, with some notable improvements in class-specific F scores, particularly for the previously underperforming “Mixed" class.
* Validation accuracy peaked at 86.67% at epoch 26, after which it began to decline slightly, suggesting that the model may have begun overfitting beyond this point. This further aligns with the trend of the steadily decreasing training loss while validation loss started increasing at around epoch 20, occasionally fluctuating from ~4.9 to 7.6.
* While the training accuracy continued to rise, reaching as high as 98% by epoch 25, it was observed that the validation accuracy remained constant within the 83–86% range. This reinforces the evidence of diminishing returns from prolonged training.
* The F1 score for the "Mixed" class improved from 0.55 to 0.62. This indicates that the extended training had allowed the model to better capture patterns for ambiguous classes. However, it is to note that the score for the “Mixed” class is still underperforming as compared to the other classes.
* Despite minor gains in test performance, the extended training duration also introduced fluctuations in validation loss and accuracy, highlighting a need for mechanisms like early stopping or learning rate scheduling to be able to stabilise training and prevent overfitting in the later epochs.

### Tuned Run #2:

* + The learning rate scheduler was first triggered at epoch 15, with the learning rate successfully reduced from 0.001 to 0.0001, followed by reductions to 0.00001 and eventually 0.000001. This scheduling behaviour aligned well with the plateauing of validation loss and allowed for more conservative and stable learning in the later stages of training.
  + Validation accuracy peaked at 90.00% around epoch 28, with consistent performance maintained in the 87–90% range from epoch 12 onward, showing more stable and gradual improvement compared to the earlier 30-epoch runs without scheduling where accuracy frequently plateaued earlier.
  + Training accuracy steadily increased to 97.50% by epoch 30. Notably, validation loss steadily decreased and remained in the range of 3.7–4.5 from epoch 10 onwards, rather than fluctuating erratically. This suggests that the learning rate reductions helped to stabilise the model’s learning and reduce overfitting risks.
  + Compared to the previous run without scheduling, class-level performance saw slight improvements in balance, with the "Mixed" class F1 score maintaining at 0.62, confirming that the scheduler did not hinder learning for minority classes. However, improvements were still limited, indicating that the class imbalance challenge remains unresolved.
  + Overall, the learning rate scheduler in this run provided a more structured and stable convergence pattern. Validation performance improved marginally and remained consistently strong throughout the training, though gains over the unscheduled 30-epoch run were minor.
  + These results suggest that the model begins to stabilise by epoch 15 and that most generalisation benefits are achieved by epoch 20–25. However, there is still potential to explore further generalisation improvements by extending training duration. Therefore, in the next tuned run, we will be extending the training to 50 epochs with the same scheduling parameters to evaluate whether longer training can lead to better performance, particularly in minority class recognition, before determining if early stopping should be enforced.

### Tuned Run #3:

* While maintaining an overall high training accuracy of 97.3% and a stable test accuracy of 87%, the performance plateaued past the mid-training point, indicating diminishing returns from extending to 50 epochs.
* Validation accuracy peaked at 89.17% around epoch 25 and remained in the 87–89% range, showing no further improvement despite continued learning.
* Validation loss was observed to have begun to plateau around epoch 20, staying within the range of 3.5 to 3.7 until the end of training. This suggests that the model had reached its generalisation peak early.
* The learning rate was progressively reduced from 0.001 to 0.0001, then to 0.00001, and finally to 0.000001 by epoch 26. While this ensured stability, smaller updates were insufficient to push the model beyond the performance achieved earlier.
* The macro F1 score was recorded at 0.83, similar to the 30-epoch result, suggesting that while the model was stable and that the added epochs did not resolve performance imbalance across classes.
* In all, this run confirms that by extending training to 50 epochs with a scheduler for LR improves training stability. However, it does not further enhance generalisation. With performance gains plateauing around epoch 25, the next run will implement early stopping to automatically halt training once validation loss stops improving, ensuring greater training efficiency without compromising model performance.

### Tuned Run #4

* + In this run, it was observed that training was halted early at epoch 15, just one epoch after the learning rate had dropped from 0.001 to 0.0001. This suggests that early stopping may have been triggered before the model can benefit from its refined learning rate adjustments.
  + A test accuracy of 87% was achieved, which is consistent with earlier tuned runs, suggesting that generalisation performance was preserved despite the shortened training duration.
  + Validation accuracy peaked at 85.83% around epoch 10 and remained relatively stable in the 83–85% range up to the point of stopping, reflecting moderate but consistent generalisation.
  + Validation loss was noted to have fluctuated between 5.1 and 5.9 in the later epochs, indicating some instability or possible noise, which may have contributed to the early stopping trigger despite ongoing learning potential.
  + Class-level performance remained strong for Apple, Banana, and Orange, with F1 scores ranging from 0.87 to 0.92. Banana experienced a minor drop in recall (0.83), which slightly impacted its overall F1.
  + The “Mixed” class had obtained a F1 score of 0.60, which is a balanced precision and recall outcome. While improved from the baseline, this result remains slightly lower than the consistent score of 0.62 observed in the previous longer runs.
  + The macro F1 score of 0.82 and weighted F1-score of 0.87 indicate a balanced performance across classes. However, further gains were likely possible with additional training under the refined learning rate.
  + The results obtained from this run validate the combination of a **scheduler with early stopping** as a possible strategy. However, it also suggests that **the current patience setting of 5 for the Early Stopping** may be **too aggressive**. As a result, cutting off training just as the scheduler had begun refining its learning rate, preventing further optimisation.
  + Therefore, the next tuned run will see an increase of the patience setting for the early stopping from 5 to 8. This is expected to allow the scheduler a chance to implement its refinement of the learning rate, while still capping any unnecessary training with the early stopping.

### Tuned Run #5

* + In this run, training halted at epoch 21 from the early stopping criteria, as compared to the previous run at epoch 15. This allows for a longer training window after the learning rate scheduler first activated at epoch 10, ensuring that the model had ample time to adapt under a reduced learning rate before termination.
  + The model was observed to have obtained a test accuracy of 87%, which is consistent with the highest performing runs, reflecting a strong generalisation performance.
  + Validation accuracy peaked at 90.00% by epoch 13 and remained stable in the 87–90% range until early stopping occurred, indicating effective learning under the combined influence of LR scheduling and delayed stopping.
  + Validation loss gradually stabilised between 3.3 and 3.6 after epoch 13, showing well-regulated learning without overfitting.
  + Class-wise performance remained strong for Apple, Banana, and Orange, with F1-scores ranging from 0.88 to 0.92. Banana showed a slightly lower recall (0.83), contributing to a modest dip in its F1.
  + The “Mixed” class achieved an F1 score of 0.62 (precision 0.50, recall 0.80), matching the best performance observed in previous runs. Therefore, confirming that the increased patience allowed the model to better capture minority class characteristics.
  + The macro F1 score of 0.83 and weighted F1 score of 0.87 also indicate a balanced performance across all classes with no overfitting, even with extended training.
  + In conclusion, this run demonstrates that increasing the early stopping patience to 8 strikes an optimal balance between training duration and model generalisation. It allowed the scheduler to guide learning deeper into convergence while still avoiding unnecessary overtraining.

### Final Recommendations