2.4 Evaluating Hyperparameters

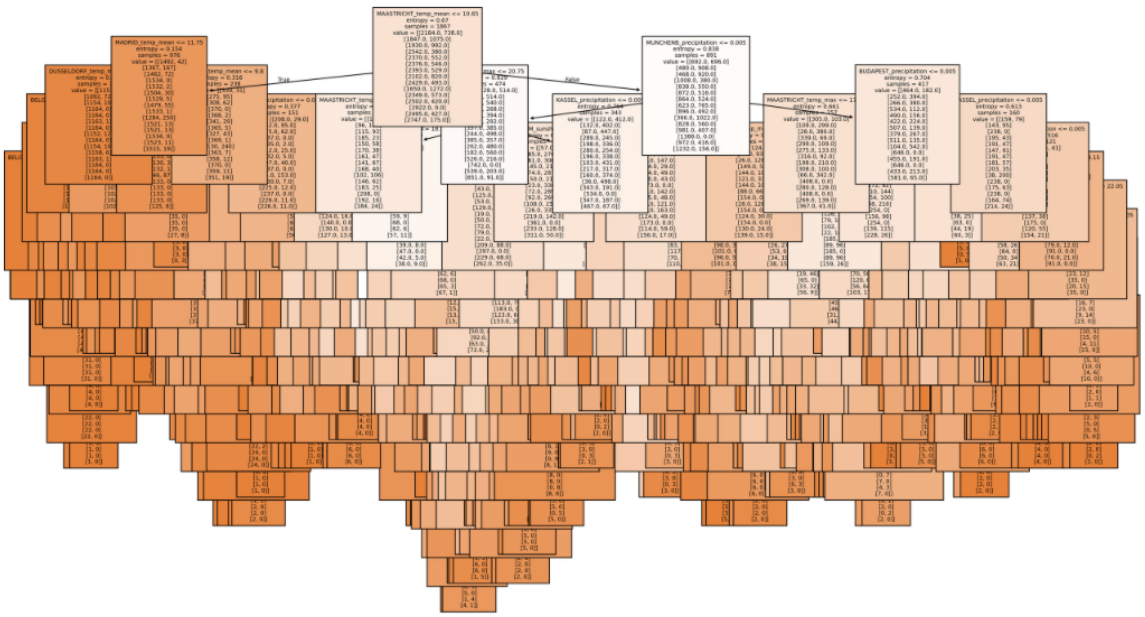
Part 1

Dataset	Accuracy before optimisation	Accuracy after optimisation
All Weather Stations and Only a Decade of Data	60.2%	56.5%
Single-Station Full Timeline Dataset	100%	99.9%

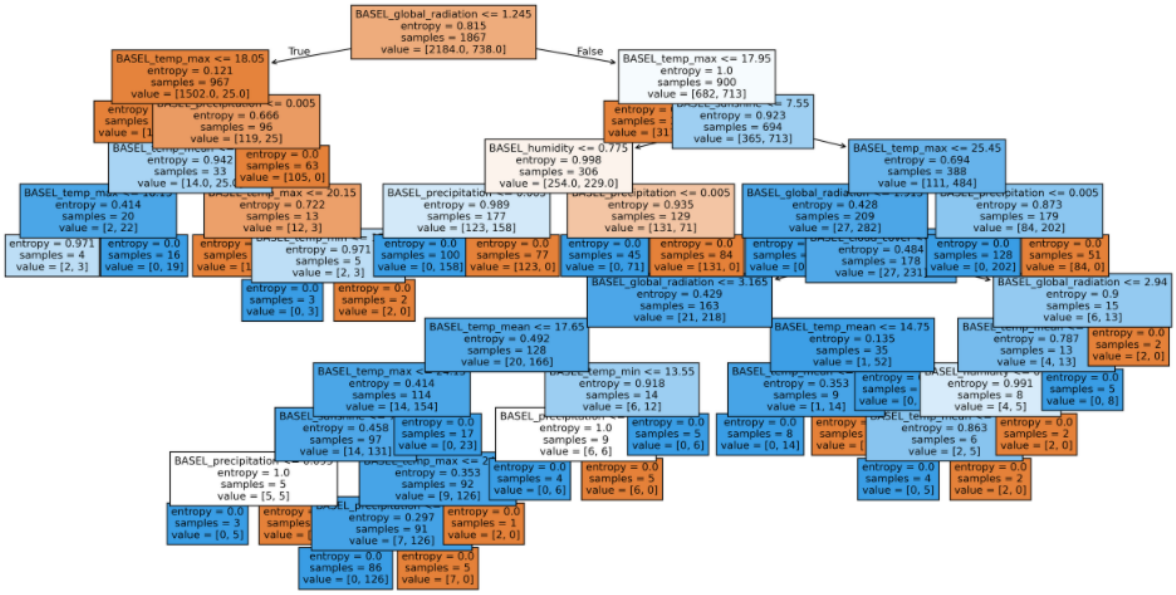
Before Optimisation

Baseline Random Forest

All Weather Stations and Only a Decade of Data



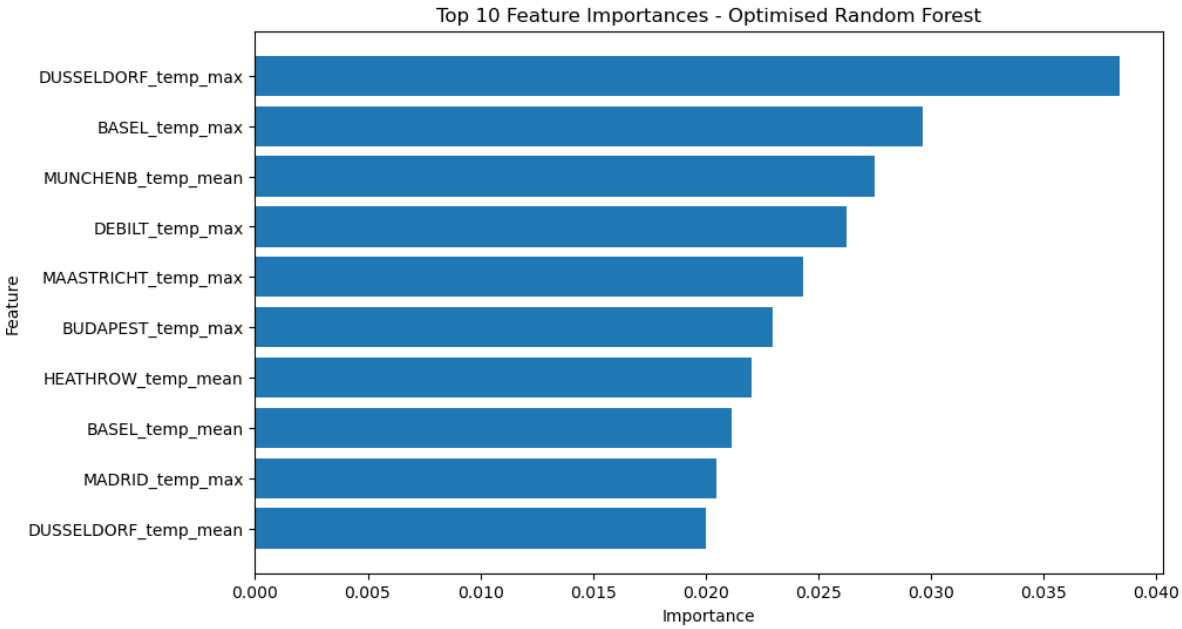
Single-Station Full Timeline Dataset



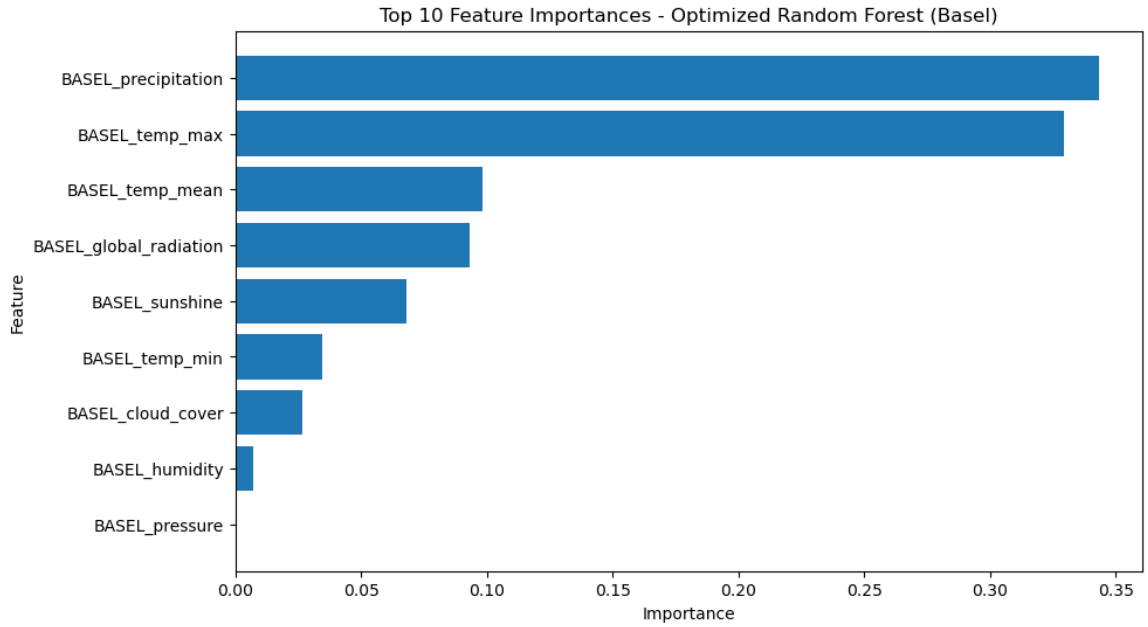
After Optimisation

Baseline Random Forest

All Weather Stations and Only a Decade of Data



Single-Station Full Timeline Dataset



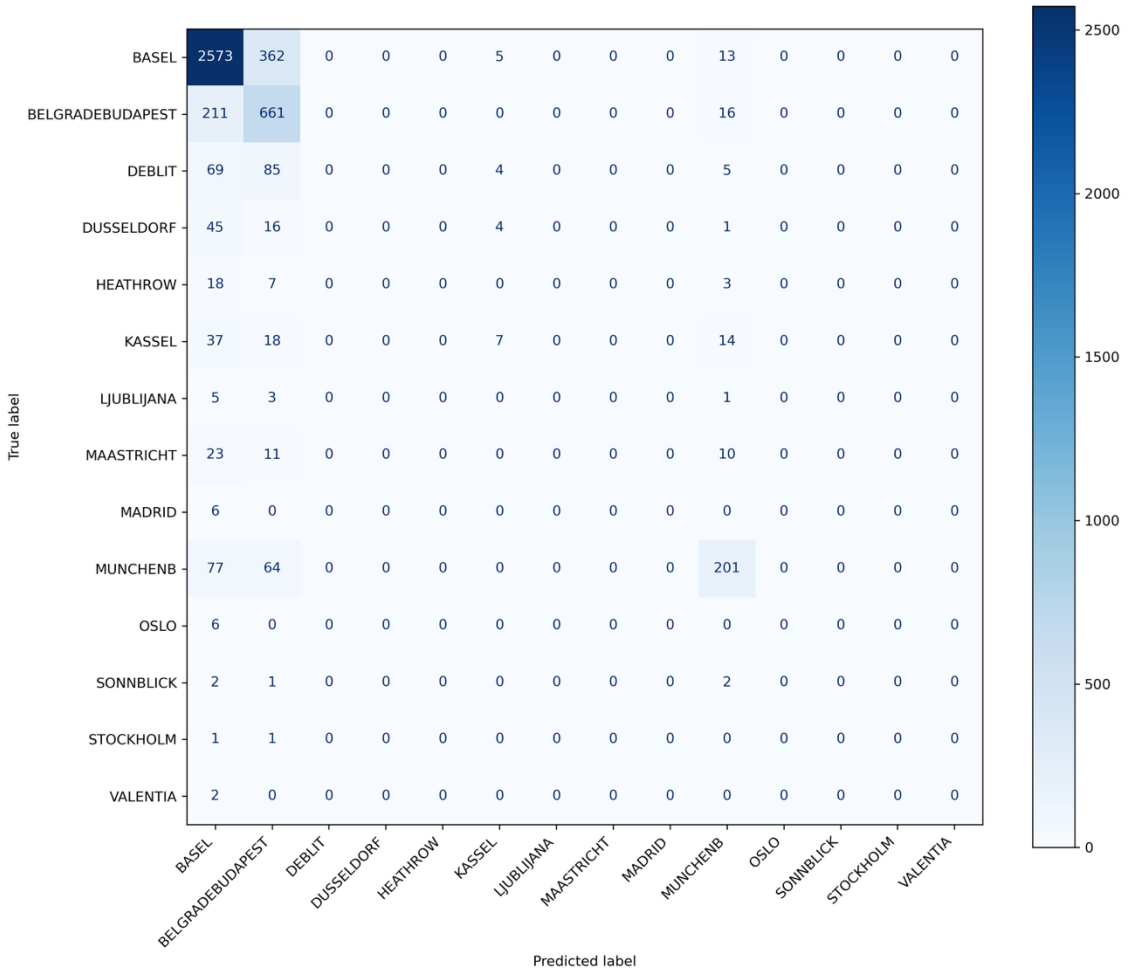
When comparing the results from Exercise 2.4 with those from Exercise 2.3, several observations emerge. For the multi-station dataset using only the first decade of data, the baseline model in Exercise 2.4 achieved a test accuracy of 0.602, which slightly decreased to 0.565 after hyperparameter optimisation. This indicates that tuning had minimal positive impact on predictive performance for this dataset, and the default Random Forest parameters were already reasonably effective. In terms of feature importance, there is a noticeable update from Exercise 2.3. Previously, global precipitation was the dominant variable across the top stations — Düsseldorf, Maastricht, and Basel — with temperature extremes also playing a key role. In Exercise 2.4, the top features across all stations over the decade are Dusseldorf_temp_max, Basel_temp_max, and MunchenB_temp_mean, confirming that temperature-related variables remain the most influential predictors of pleasant weather, while precipitation plays a relatively smaller role in this multi-station, first-decade dataset.

For the single-station dataset (Basel) covering the full timeline, the baseline model already achieved near-perfect accuracy (1.0), and the optimised model remained similarly high at 0.999, showing minimal change. Feature importance for Basel identified basel_precipitation, basel_temp_max, and basel_temp_mean as the primary drivers of the model's predictions. Overall, hyperparameter tuning had a limited effect on accuracy, but the analysis confirms that temperature- and precipitation-related features are consistently the most critical indicators for predicting pleasant weather, both across multiple stations over a decade and at a single station across the full historical timeline.

Part 2

In Exercise 2.2, the baseline CNN models struggled to learn meaningful patterns from the weather data. Trial 1 achieved an accuracy of only 9.15%, missing two stations, with

training and validation accuracies showing slight improvement but overall poor generalisation. Trial 2 performed worse, reaching 5.04% accuracy and failing to recognise three stations, with both training and validation losses diverging. Trial 3, despite reaching a higher accuracy of 19.34%, exhibited severe instability and likely overfitting, missing three stations and showing erratic validation performance. In Exercise 2.4, the Bayesian-optimised CNN achieved substantial improvements in accuracy, with training accuracy of 73.87%, validation accuracy of 74.26%, and test accuracy of 74.99%. However, the confusion matrix reveals that the model still did not recognise all stations, missing ten stations: OSLO, HEATHROW, LJUBLIJANA, DEBLIT, MADRID, MAASTRICHT, VALENTIA, STOCKHOLM, SONNBLICK, and DUSSELDORF. Despite this, the close agreement between training and validation accuracies suggests the model is not overfitting, and the optimisation has stabilised training and improved generalisation compared to the baseline CNNs. Overall, the Bayesian-optimised model represents a significant step forward in learning temporal patterns from the weather data, even though recognition of all stations remains incomplete.



Part 3

To effectively test and iterate, the data can be broken down by station, time window, and feature subset. Starting with a single station, such as Basel, allows focused analysis

of temporal patterns, faster iteration, and clearer insights into predictive variables. Shorter time windows or the first decade of multi-station data help reduce dataset size and highlight general trends before scaling up. Random Forest models work well for these smaller subsets, providing baseline performance and interpretable feature importance, while deep learning models, particularly CNNs, excel at capturing temporal dependencies across larger datasets. Observations from the Random Forest analyses indicate that single-station models achieve near-perfect accuracy, with precipitation, maximum temperature, and mean temperature emerging as the most influential features. Multi-station models rely heavily on maximum temperature across several locations. The CNN with Bayesian-optimized hyperparameters improved training, validation, and test accuracy (~74–75%), showing that careful tuning stabilises learning and mitigates overfitting compared to arbitrary configurations. However, the model did not recognise all stations in the multi-class classification task, with several stations missing from predictions, suggesting that some station-specific patterns remain difficult to capture. For Air Ambulance safety decisions, the most critical variables remain precipitation, temperature extremes, and solar radiation, as these factors strongly impact local flight conditions and risk assessment.