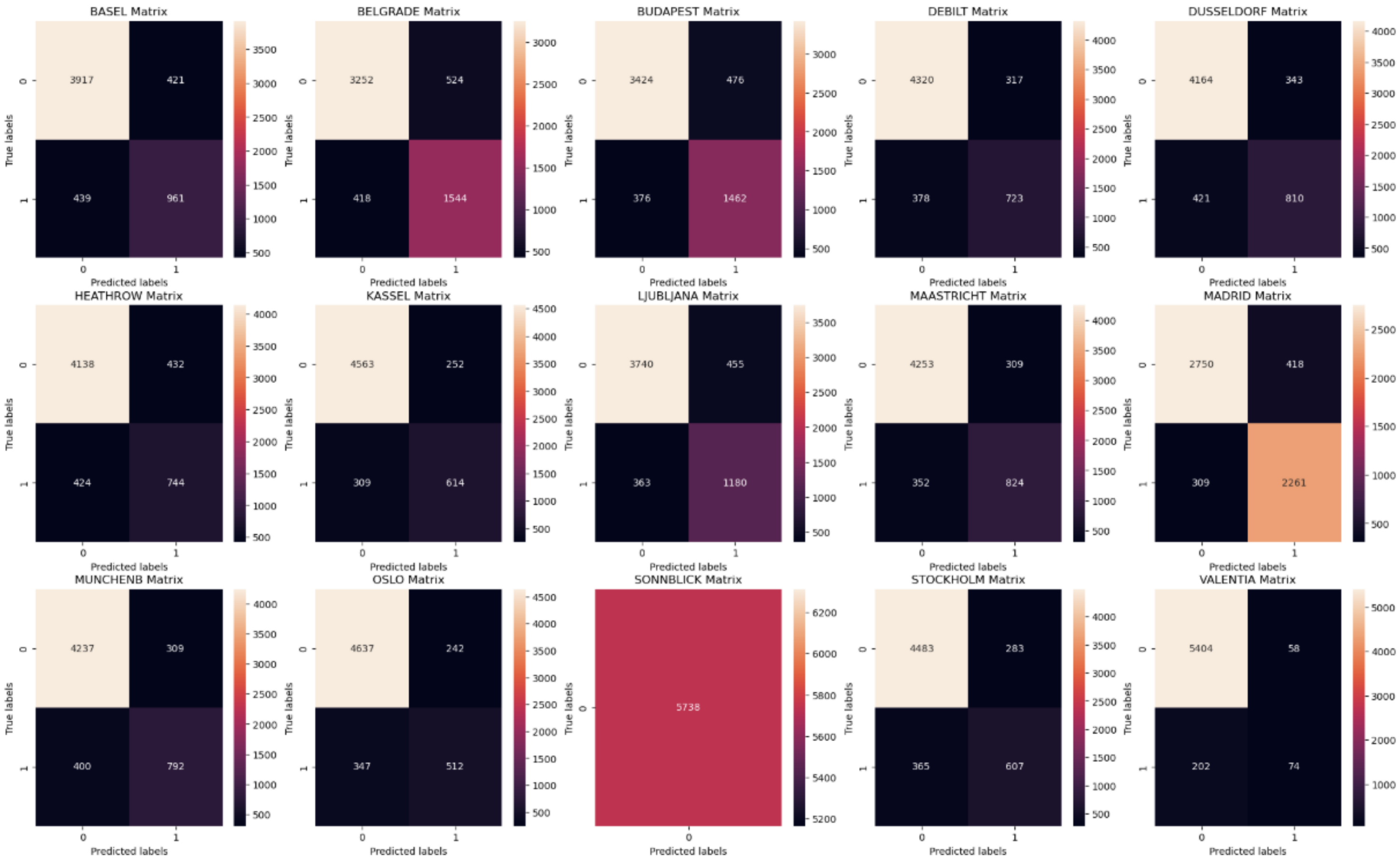
Exercise 1.4 Supervised Learning Algorithms Part 1

By Daniel Mueller

Supervised Learning Algorithm - K_Nearest Neighbor



Interpretation

1.1 Overall Performance of the Model

- The confusion matrices show varying degrees of prediction accuracy for different weather stations.
- Some stations exhibit a strong diagonal dominance (indicating good predictions), while others show misclassifications in both categories (pleasant vs. unpleasant weather).

1.2 Are Any Weather Stations Fully Accurate

- Sonnenblick appears to be highly inaccurate
  - Its confusion matrix contains only one class (0), meaning the model never predicted 1 (unpleasant).
  - This suggests the model is either overfitting to one class or the dataset lacks enough variation for that station.
- No station is fully accurate, as all others show some misclassification (off-diagonal values).

1.3 Is Overfitting Happening?

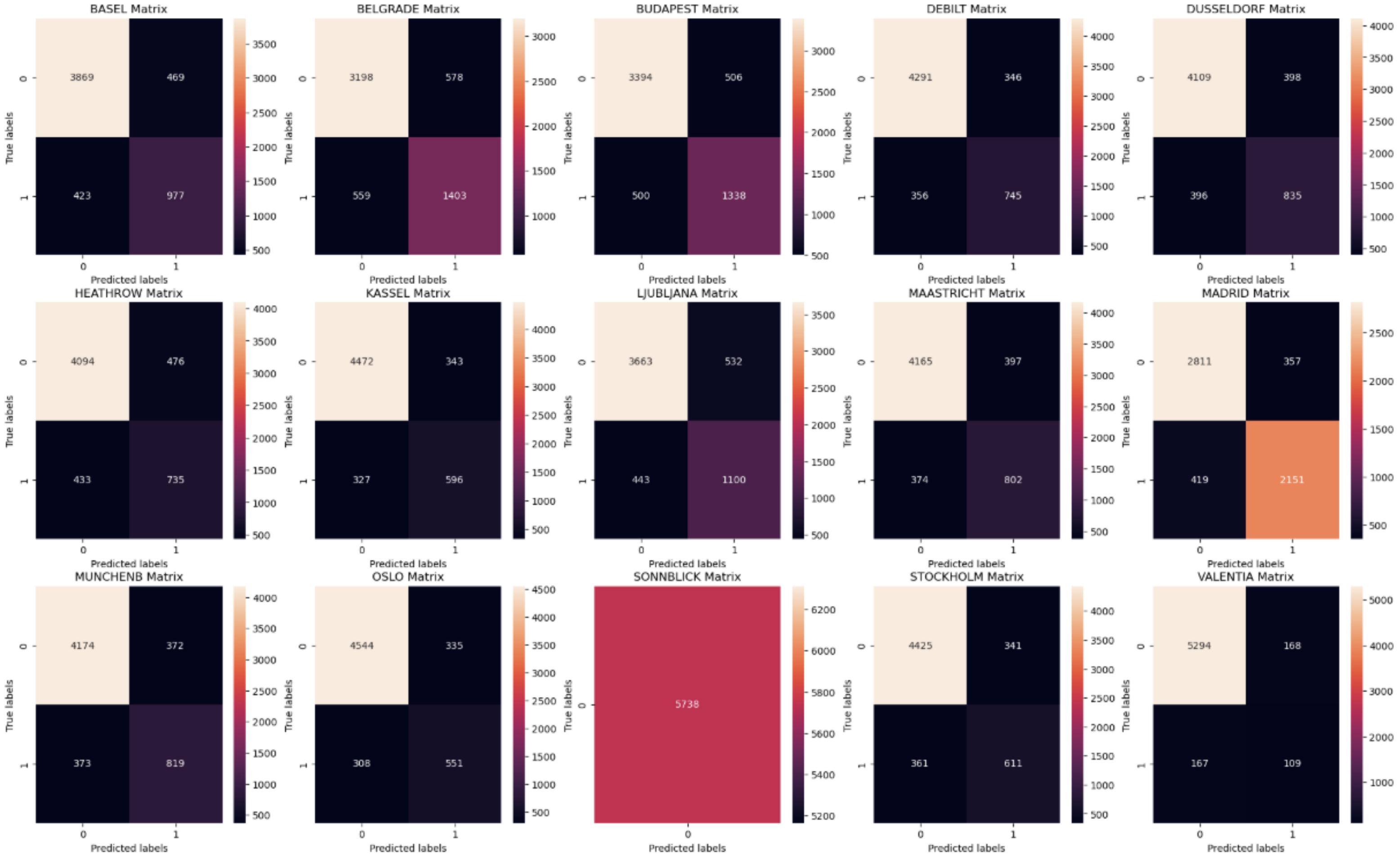
- Possible overfitting to the dominant class in some stations
  - Many matrices show a much higher count for 0-0 (true positives) compared to 1-1 (true negatives), suggesting a bias toward predicting "pleasant" weather.
  - This bias may indicate class imbalance in the dataset or overfitting to frequent weather patterns.
- Some stations have balanced misclassification errors, indicating more generalized predictions.

1.4 Stations Contributing to Overall Accuracy or Inaccuracy

- SONNBLICK is the most problematic: It completely fails to predict 1, which may significantly affect overall model accuracy.
  - Stations like Madrid and Valentina show a large number of 1-1 (true negatives), suggesting better performance in predicting unpleasant weather.
- Stockholm, Kassel, and Oslo have relatively balanced confusion matrices, indicating they contribute to a more stable overall model accuracy.

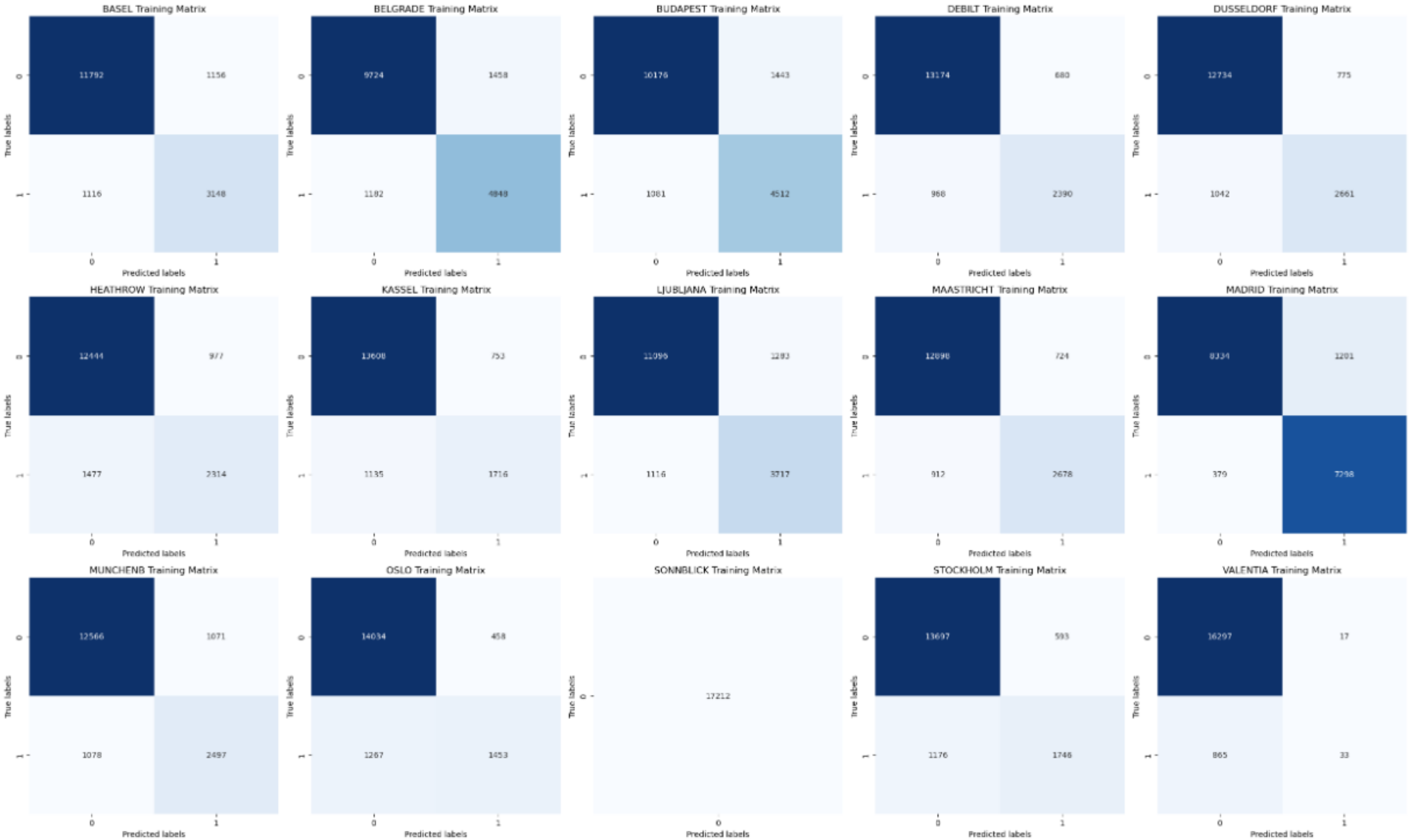
Exercise 1.5 Supervised Learning Algorithms Part 2

Supervised Learning Algorithm – Decision Tree

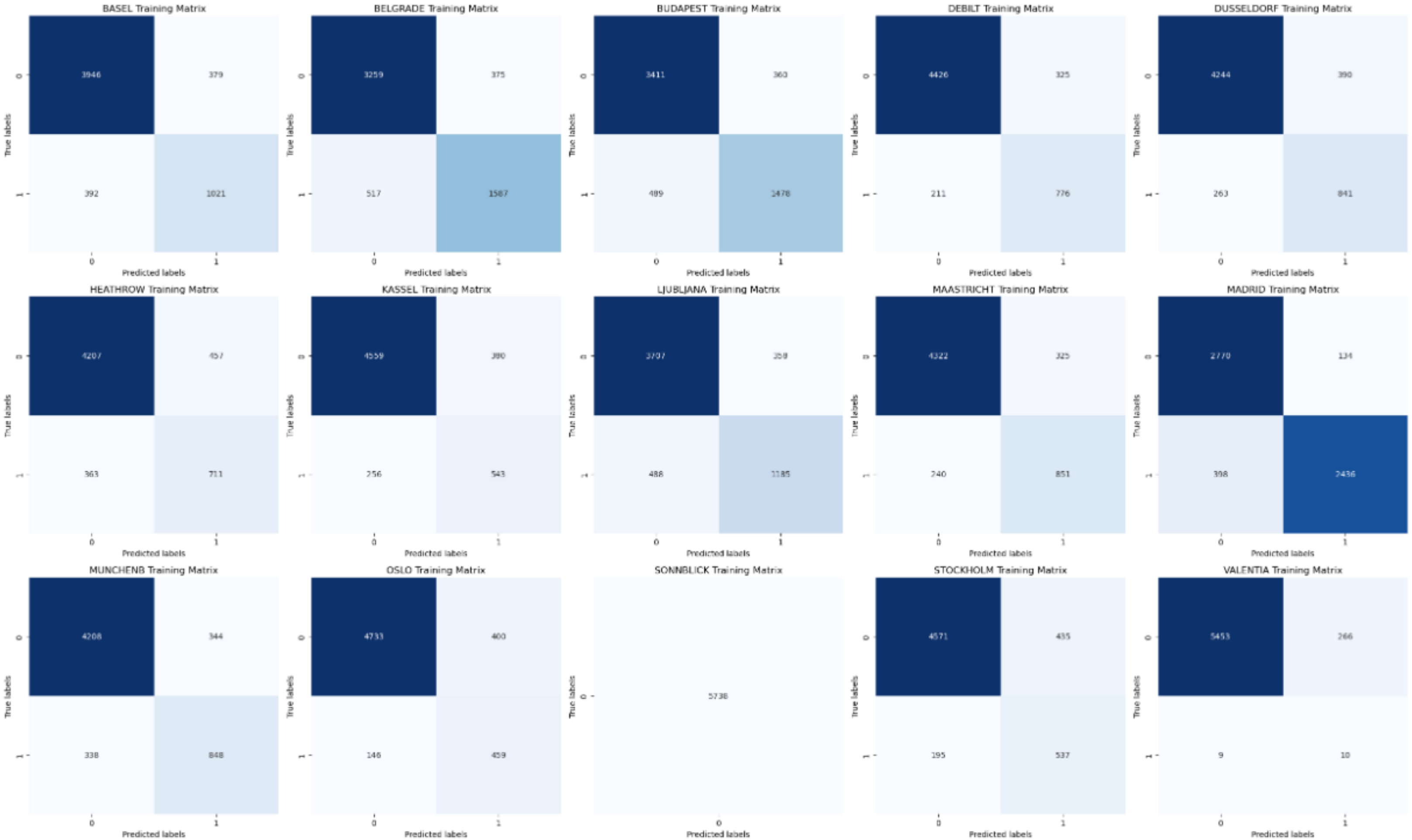


Supervised Learning Algorithm – Artificial Neural Networks (ANNS)

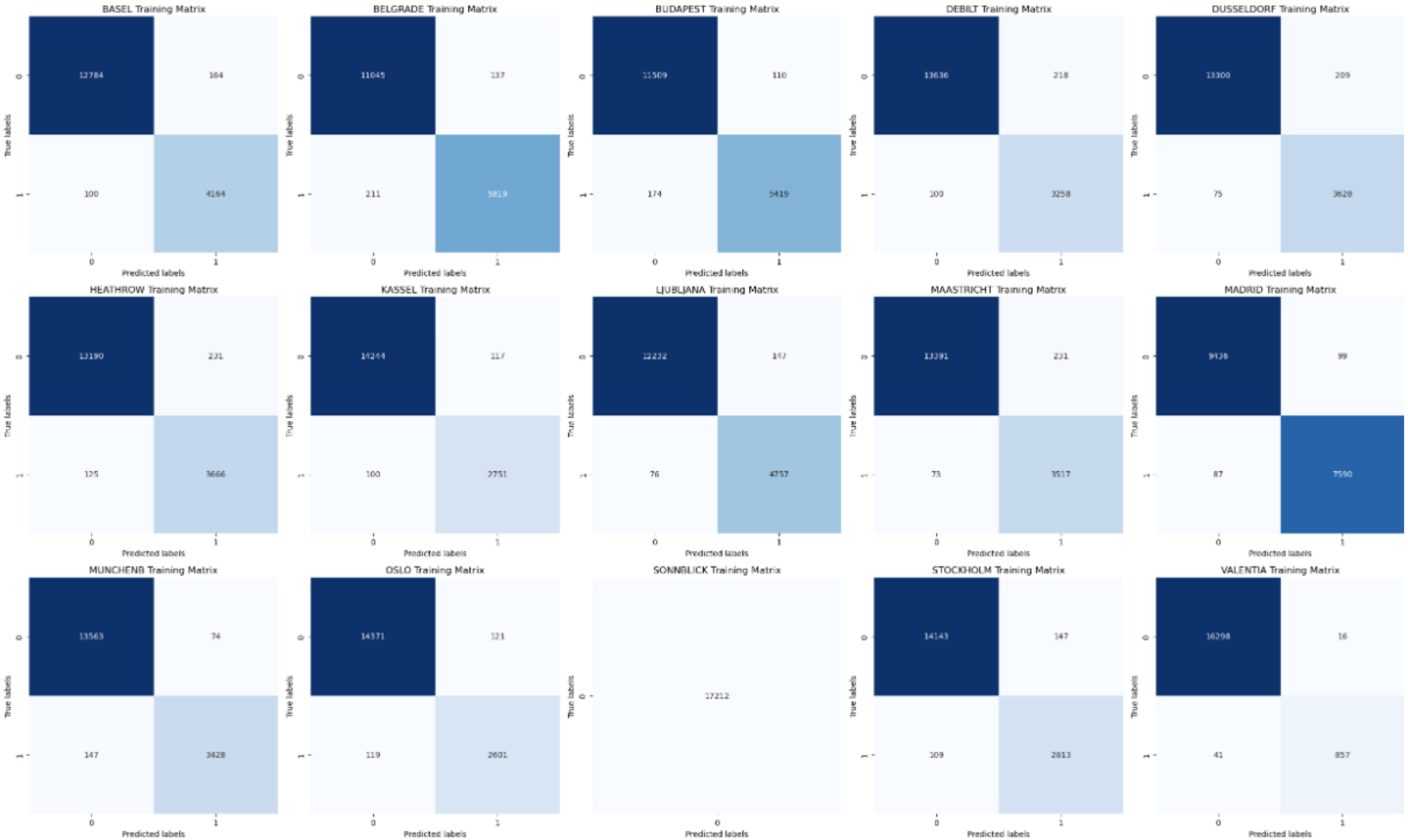
RUN 1 - Training



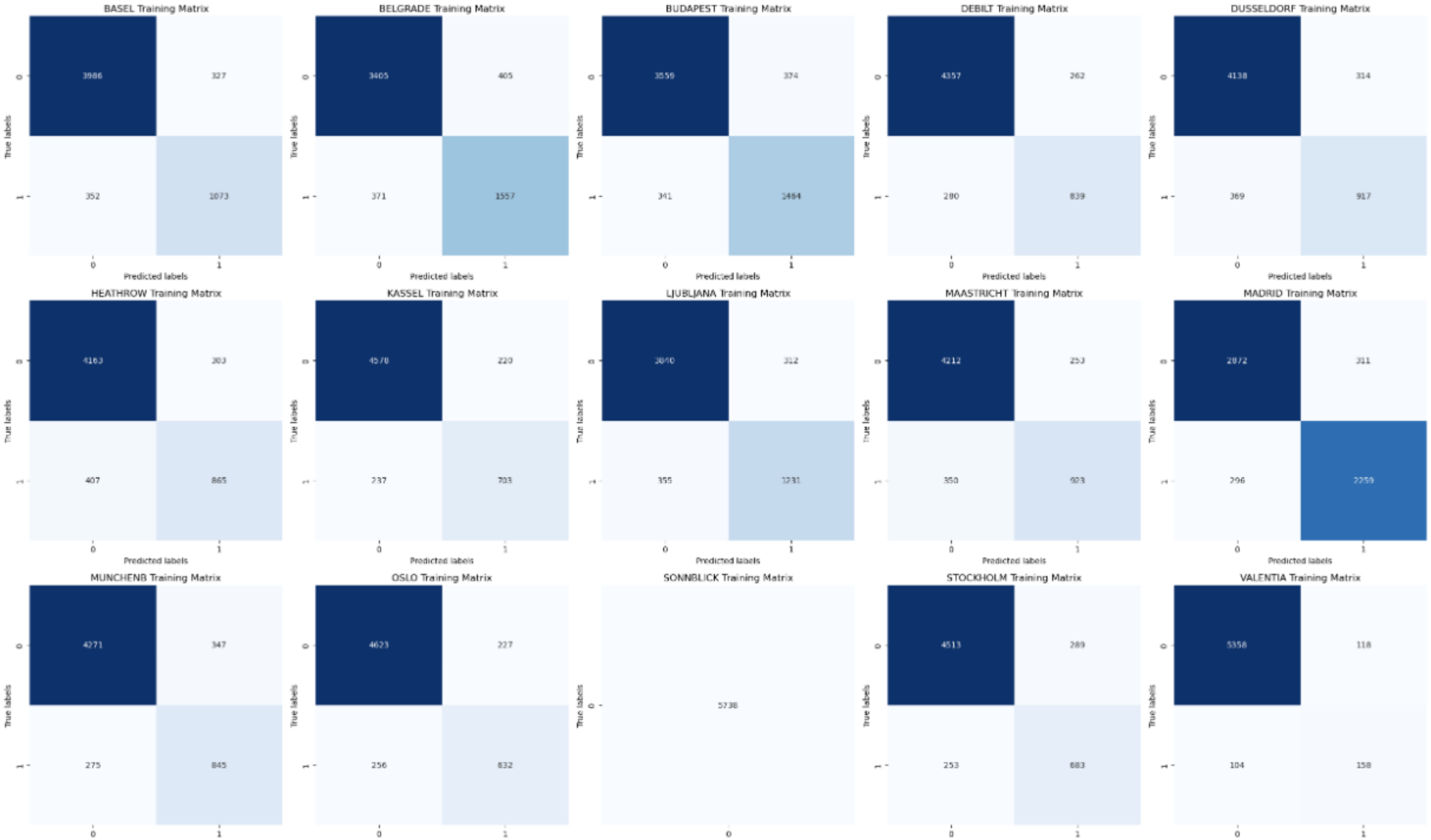
RUN 1 – Test


BASEL Training Matrix
BELGRADE Training Matrix
BUDAPEST Training Matrix
DEBILT Training Matrix
DUSSELDORF Training Matrix
HEATHROW Training Matrix
KASSEL Training Matrix
LJUBLJANA Training Matrix
MAASTRICHT Training Matrix
MADRID Training Matrix
MUNCHENB Training Matrix
OSLO Training Matrix
SONNBLICK Training Matrix
STOCKHOLM Training Matrix
VALENTIA Training Matrix
True labels
Predicted labels

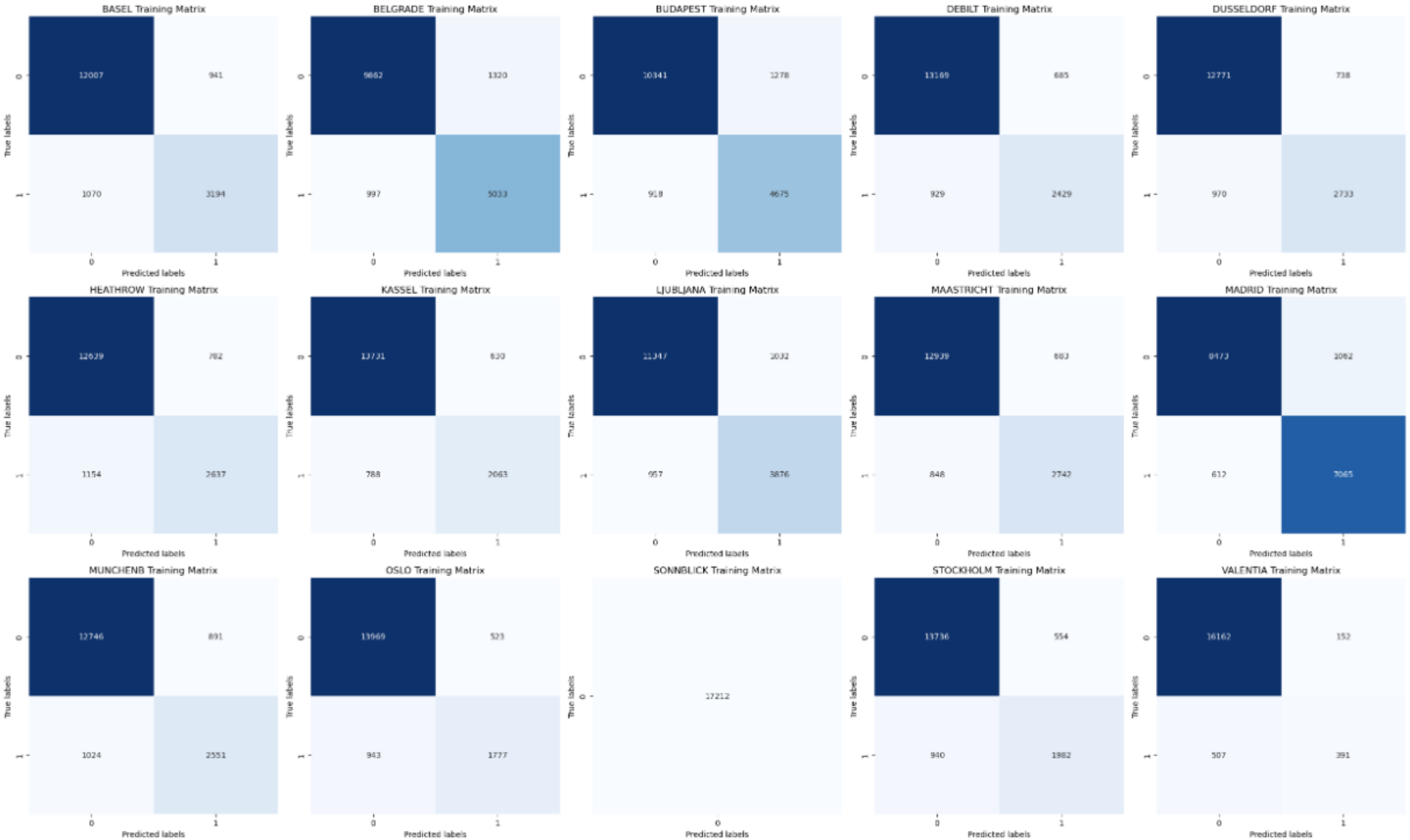
RUN 2 – Training



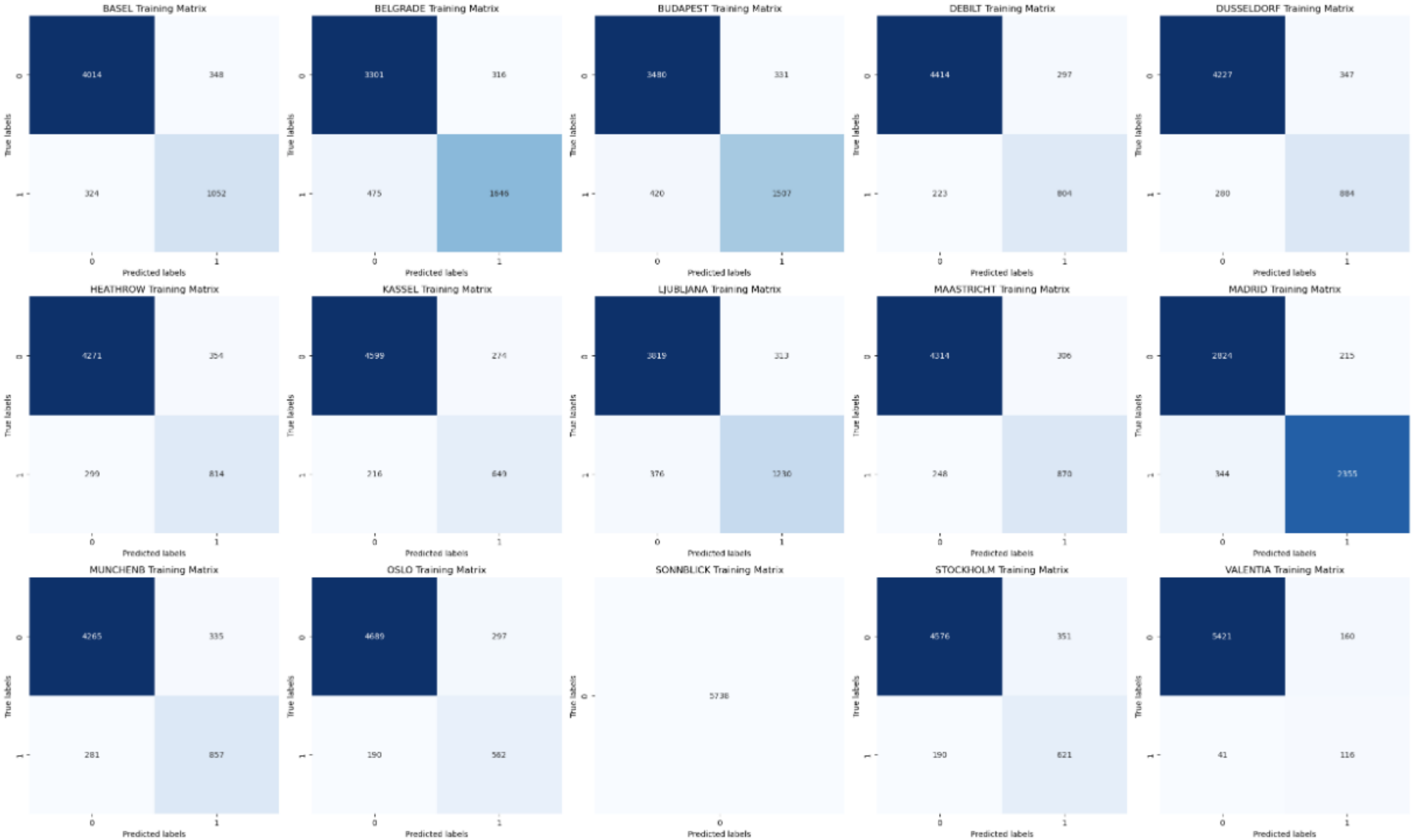
RUN 2 – Test



FINAL RUN – Training



FINAL RUN – Test



Interpretation

1.5 Which algorithm (including the KNN model from Exercise 1.4) best predicts the current data?

- The KNN model appears to be the most effective in predicting the current data.

1.6 Are any weather stations fully accurate? Is there any overfitting?

- No weather station is fully accurate.

1.7 Which features in the dataset contribute to overall accuracy?

- The dataset includes various weather-related features such as temperature, wind speed, and humidity. The complexity of these factors makes it challenging to achieve consistently high accuracy.

1.8 Which model would you recommend for ClimateWins?

- I recommend that ClimateWins use the ANN model. Although it requires more manual input and fine-tuning, ANN is well-suited for handling complex weather patterns. By learning from historical data, it can improve prediction accuracy over time. With human supervision and continuous optimization, ANN can be trained to deliver more precise forecasts based on station data.