Random Forests vs. Neural Networks; a Comparison for NO_2 Concentration Prediction

Lasse Vulto, Ezekiel Djeribi Stevens, Bennet Weiß

Institute for Marine and Atmospheric Research Utrecht, Utrecht University

Utrecht University

Institute for Marine and Atmospheric research Utrecht

1. Abstract

- In the light of emergent capabilities of machine learning (ML) techniques we apply a random forest (RF) and a feed-forward neural network (NN) to air pollution prediction based on meteorological variables, specifically predicting NO_2 concentration.
- To achieve optimal prediction, the two different ML approaches (RF and NN) are compared in terms of their predictive skill and computational efficiency.
- Meteorological data from a measurement station in Eindhoven between 2015 and 2017 is used for training, 2018 is used for testing.
- Both approaches produce similar results: RMSE is below 9 $\mu g/m^3$ and a correlation with observations of around 0.75. The observed variance is not fully represented by models as the explained variance hovers around 50%.
- The RF is at least 2 orders of magnitudes more computationally efficient than the NN which translates into a significantly faster hyperparameter optimization and training.
- This makes the RF the ML approach of choice for hourly NO_2 concentration prediction.

2. Introduction

- Air pollution, especially Nitrogen Dioxide (NO_2), poses significant global environmental and public health concerns (WHO, 2022). Figure 2 shows the Benelux region's elevated NO_2 concentration.
- NO_2 has adverse effects on human health and ecosystems, serving as a precursor to harmful pollutants such as ozone and particulate matter (EPA, 2023).
- Accurate NO2 concentration prediction is vital for informed decision-making and policy development. Traditional modelling methods rely on complex chemical transport models, demanding substantial resources (Zhang et al., 2023).
- Machine learning based approaches, specifically Neural Networks (NN) and Random Forests (RF) will be compared to each other for accuracy and efficiency.

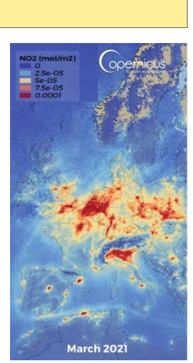


Figure 2. NO_2 concentrations for the EU region (Copernicus).

Figure 1. An AI artists impression of Machine learning based air quality modelling research (Dall-E)

3. Methods

Random Forest

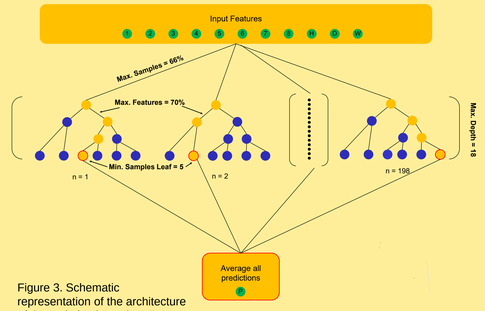


Figure 3. Schematic representation of the architecture of the optimized Random Forest.

Neural Network

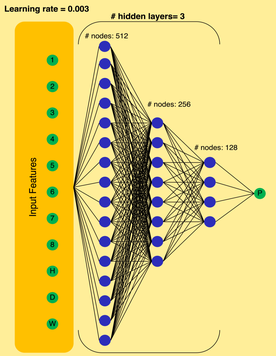


Figure 4. Schematic representation of the architecture of the optimized Neural Network.

4. Data

- The model training was carried out based on meteorological data from the years 2015-2017, with 2018 being an independent test set.
- Data courtesy of Rijksinstituut voor Volksgezondheid en Milieu (RIVM)
- In Figure 5, the location of the RIVM measurement station is illustrated, located in urban Eindhoven.
- 8 meteorological variables with an hourly resolution and 3 temporal variables were selected as input features (Figure 6).
- All timestamps with unphysical data (negative values) were corrected to 0, while timestamps with missing data (< 3% of dataset) were cleaned from the dataset

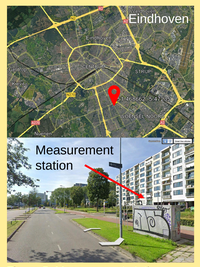


Figure 5. Measurement station location (Google Earth Pro).

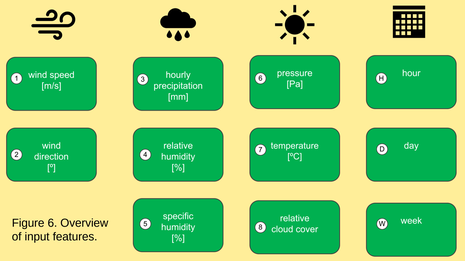


Figure 6. Overview of input features.

5. Results

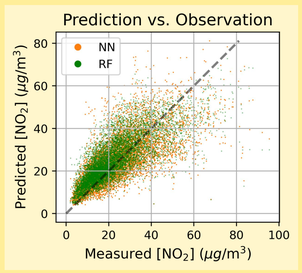


Figure 7. Comparison of hourly NN (orange) and RF (green) predictions for testing data (2018).

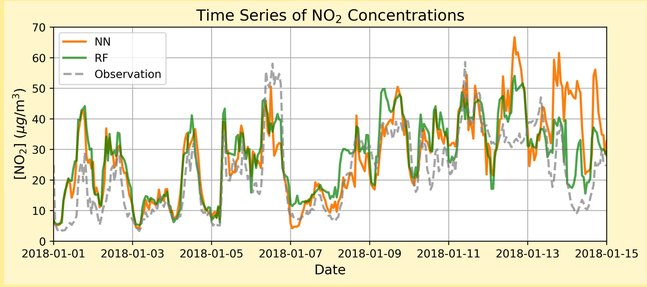


Figure 8. Observation (dashed) timeseries for the first 14 days of 2018 (testing data) compared to NN (orange) and RF (green) predictions.

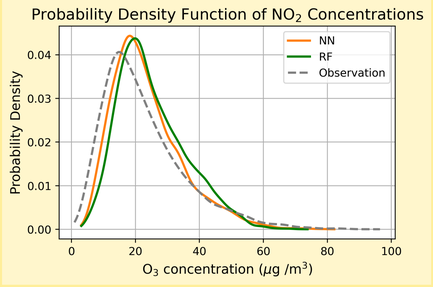


Figure 9. Probability distribution for observed values (dashed), NN (orange) and RF (green) predictions for testing data (2018).

Table 1. Statistical comparison between RF and NN predictions

Metric	Random Forest	Neural Network
RMSE ($\mu g/m^3$)	8.78	8.91
Correlation	0.76	0.74
Explained Variance	0.58	0.53
MAE ($\mu g/m^3$)	6.66	6.37

6. Discussion & Conclusion

How does the prediciton quality compare?

- Both the NN and the RF are able to reproduce basic features of the hourly NO_2 concentration with a RMSE of below 9 $\mu g/m^3$ and a correlation of around 0.75.
- However, these machine learning approaches fall short of reproducing the full variance as only about 50% of the observed variance is reproduced.

→ **Both ML approaches offer similar prediction quality.**

Did we find the optimal hyperparameters?

- For the RF, a random grid search (Figure 10, middle) was performed with 500 iterations, while the NN was optimized over 100 iterations by adaptive selection Figure 10, right).
- As adaptive selection is based on searching in regions where it found promising results, it can be considered more efficient and is therefore expected to compensate for the reduced number of iterations.
- The possible hyperparameter space was limited by computational resources.

→ **While the perfect hyperparameters were certainly not found, we do not expect significant improvement with more optimization iterations. Especially NNs with architecture outside of the hyperparameter space might offer further improvements.**

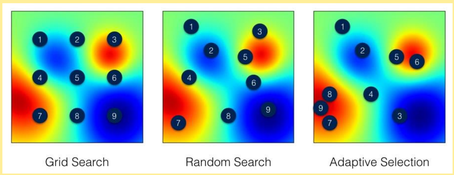


Figure 10: Different hyperparameter optimizing methods (Carnegie Mellon, 2018).

Is there a difference in computational cost?

Metric	Random Forest	Neural Network
Optimizing: 20 iterations	20 seconds	1 hour, 15 minutes
Optimizing: 100 iterations	1 minute, 30 seconds	7hours, 30 minutes
Total amount of optimization iterations	500	100
Training optimised network	< 1 second	14 minutes, 46 seconds

Table 2. Computational cost for the RF and the NN

So which ML method is best suited for NO_2 concentration prediction?

- Predicative skill is similar between RF and NN.
- RF is more efficient for both model training and testing.

→ **RF is a clear winner**

Our Data and analysis are free to access with a GPL 3.0 Licence.

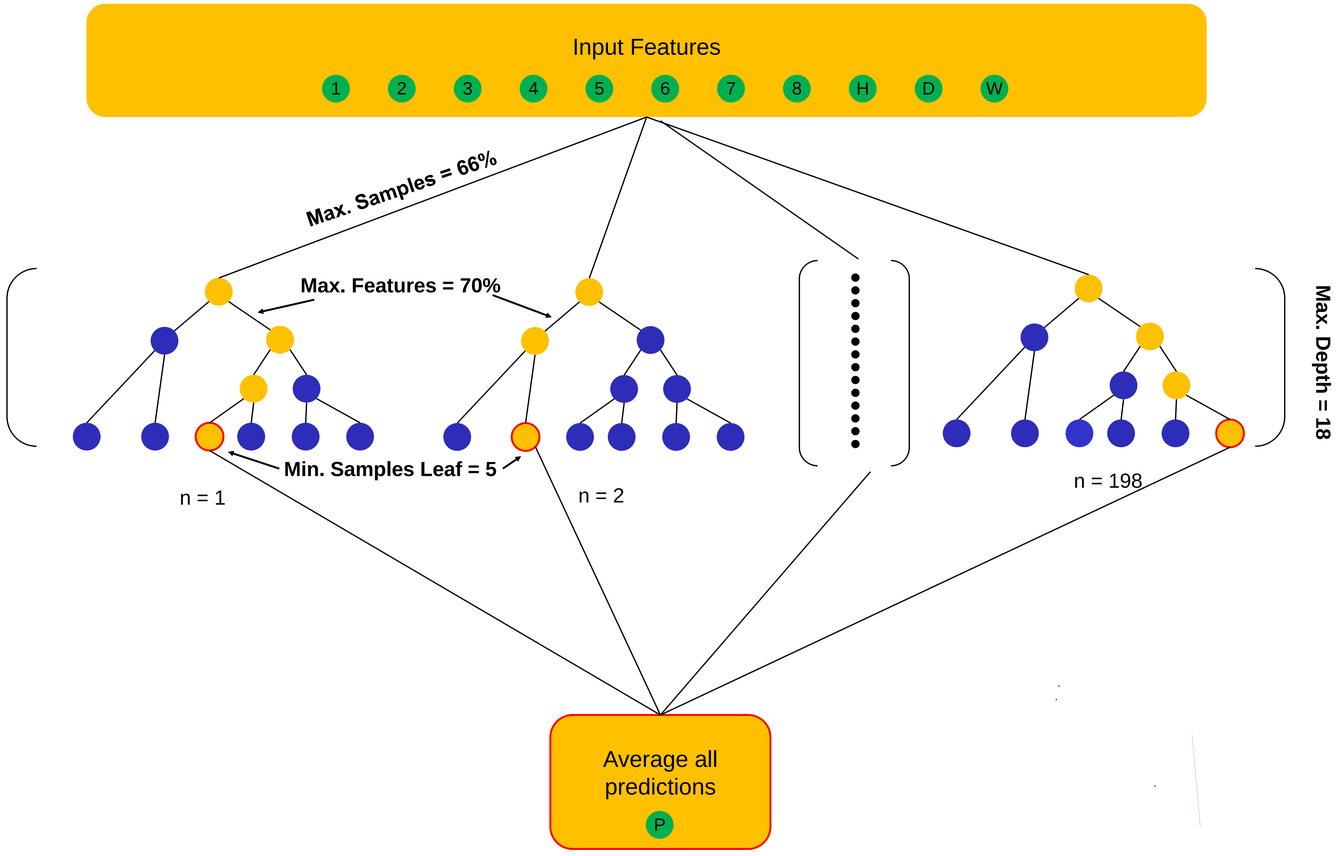
References

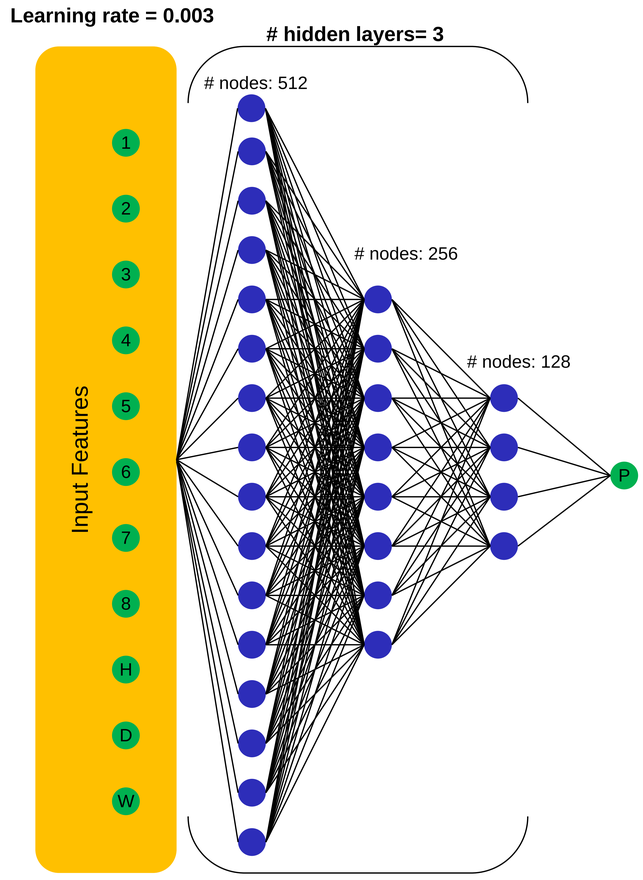
EPA. (2023). *Nitrogen dioxide (no2) pollution*. United States Environmental Protection Agency. Retrieved October 8, 2023, from https://www.epa.gov/no2-pollution

WHO. (2022). *Ambient (outdoor) air pollution*. World Health Organization

Zhang, D., Wang, Q., Song, S., Chen, S., Li, M., Shen, L., Zheng, S., Cai, B., Wang, S., & Zheng, H. (2023). Machine learning approaches reveal highly heterogeneous air quality co-benefits of the energy transition. *iScience*, *26*(9), 107652. https://doi.org/10.1016/j.isci.2023.107652

Chat GPT was used in this project. OpenAI. (2023). *Chatgpt*. Retrieved November 3, 2023, from https://chat.openai.com





Batch size = 512

Input Features

1 wind speed [m/s]

3 hourly precipitation [mm]

6 pressure [Pa]

H hour

2 wind direction [º]

4 relative humidity [%]

7 temperature [ºC]

D day

5 specific humidity [%]

8 relative cloud cover

W week

Metric	Random Forest	Neural Network
RMSE ($\mu g/m3$)	10.23	8.91
Correlation	0.64	0.74
Explained Variance	0.41	0.53
MAE ($\mu g/m3$)	7.79	6.37