



Opening the Black Box: Explainable machine learning techniques for air quality sensor calibration

Miriam Chacón-Mateos¹, Eduardo Herrera-Carrión¹, Marc Golder², Katja Mannschreck², Ulrich Vogt³, Sebastian Diez⁴, Tobias Grein¹, Joschka Kieser⁵, Sven Reiland⁵, Nina Gaiser¹, and Markus Köhler¹

¹Institute of Combustion Technology, German Aerospace Center, Stuttgart, Germany

²Technical Faculty, University of Applied Science of Heilbronn, Germany

³Institute of Combustion and Power Plant Technology, University of Stuttgart, Stuttgart, Germany

⁴Centro de Investigacion en Tecnologías para la Sociedad, Universidad del Desarrollo, Santiago, Chile

⁵Institute of Vehicle Concepts, German Aerospace Center, Stuttgart, Germany

Air pollution remains a major environmental and public health challenge. The World Health Organization (WHO) estimates that air pollution is associated with 9 million premature deaths annually. Low-cost sensors (LCS) are a promising complement to regulatory monitoring because they can deliver high frequency, hyperlocal air quality data. However, LCS data quality is affected by limitations of the measuring principle, sensor drift/aging, cross-sensitivities to other compounds, and meteorological influences like temperature (T) and relative humidity (RH), which can undermine reliability and stakeholder trust. In recent years, machine learning (ML) has been widely explored and applied to LCS data to correct systematic biases in raw sensor signals and improve the accuracy of the measurements, yet the frequent lack of explainability of black-box models can further reduce transparency and confidence in the post-processed sensor data.

In the context of the MoDa project and in collaboration with UrbanAirLab project of the University of Applied Sciences in Heilbronn, this study aims to create an explainable ML calibration workflow for LCS NO₂ measurements to enhance transparency of calibration models. The dataset consists of 1-min raw data with a co-location period from 01.06.2025 to 20.11.2025 in a regulatory measurement station located in Heilbronn (urban background). First, an exploratory data analysis (EDA) is carried out, which includes time synchronization of LCS and reference data, handling of missing values, outlier detection with Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and resampling to hourly averages. Then different calibration models are trained including as input parameters the working and auxiliary electrode signals of the NO₂ sensor as well as external data such as T, RH and O₃ data. The tested models include Multiple Linear Regression (MLR), Support Vector Regressor (SVR), Random Forest Regressor (RF), eXtreme Gradient Boosting (XGBoost), and Artificial Neural Network (ANN). The performance evaluation is carried out using the relative expanded uncertainty as suggested in DIN CEN TS 17660-1 and also other standard metrics such as RMSE, MAE, R², and bias.

The results of these metrics suggest that RF provides the best overall performance (RMSE = 5.50

$\mu\text{g}/\text{m}^3$, MAE = 3.93 $\mu\text{g}/\text{m}^3$, $R^2 = 0.69$; Pearson $r = 0.83$) and near-zero mean bias. XGBoost performs similarly (RMSE = 5.62 $\mu\text{g}/\text{m}^3$, $R^2 = 0.69$), followed by ANN (RMSE = 5.76 $\mu\text{g}/\text{m}^3$, $R^2 = 0.67$).

Explainable ML techniques are implemented in a second step as an auditing layer to support data quality assurance and control (QA/QC). These include Permutation Feature Importance (PFI) to screen which predictors most affect out-of-sample performance by measuring the score drop after removing each feature, SHapley Additive exPlanations (SHAP) for global and local attributions, and Individual Conditional Expectation (ICE) and Partial Dependence (PDP) Plots to summarize average effects while exposing heterogeneity and interaction patterns. Because predictors such as T and RH are often correlated in co-location datasets, we also use Accumulated Local Effects method to obtain more reliable effect estimates under feature dependence.

By combining reproducible calibration models with systematic explainability, this work supports more transparent QA/QC practices and contributes to creating transferable workflows for deploying LCS for air-quality monitoring.