

A COMPARATIVE STUDY OF CONVENTIONAL AND DEEP LEARNING CORRECTION FUNCTIONS FOR LOW- COST PARTICULATE MATTER MEASUREMENTS

Krzysztof Emerling, Katarzyna Schmidt, Patryk Sitarek
University of Silesia, Katowice, Poland
katarzyna.schmidt@us.edu.pl

Abstract

The goal of the project was to compare several machine learning models in improving the measurement accuracy of a low-cost optical PM2.5 particle sensor, whose performance deteriorates with increasing air humidity [1]. We aim to determine if the deployment of more complex models is justified by verifiable accuracy gains within our specific testing environment. Measurement data were collected by two identical monitoring stations for a few months in one city in Upper Silesia, Poland, with recordings taken every 10 minutes. Both stations were equipped with a heater to dry the air, but in one station the heater was permanently turned off, ensuring identical airflow in both stations. The dataset used in the project included the date and time of each measurement, readings from both stations measuring PM10 and PM2.5 levels, although only PM2.5 readings were used. The models utilized four input features - readings from a station without air drying combined with three other climatological measurements - to forecast the heated station's data.

Keywords: Air quality monitoring, PM2.5 particles, humidity compensation, machine learning, polynomial regression, random forest, DNN, time series analysis

Primerjalna študija konvencionalnih funkcij in funkcij globokega učenja za popravke meritev nizkocenovnih senzorjev trdnih delcev

Izvleček

Cilj projekta je bil primerjati več modelov strojnega učenja z namenom izboljšanja natančnosti meritev nizkocenovnega optičnega senzorja delcev PM2.5, katerega delovanje se poslabša z naraščajočo vlažnostjo zraka. Želimo ugotoviti, ali je uporaba kompleksnejših modelov upravičena s preverljivimi pridobitvami pri natančnosti v našem specifičnem testnem okolju. Podatki o meritvah so bili zbirani z dvema enakima merilnima postajama nekaj mesecev v enem od mest v Zgornji Šleziji na Poljskem, pri čemer so se meritve izvajale vsakih 10 minut. Obe postaji sta bili opremljeni z grelcem za sušenje zraka, vendar je bil na eni postaji grelec trajno izklopljen, kar je zagotovilo enak pretok zraka v obeh postajah. Podatkovni niz, uporabljen v projektu, je vključeval datum in čas vsake meritve ter odčitke z obeh postaj, ki sta merili ravni PM10 in PM2.5, čeprav so bili uporabljeni le odčitki PM2.5. Modeli so za napovedovanje podatkov z ogrevane postaje uporabljali štiri vhodne značilnosti – odčitke s postaje brez sušenja zraka v kombinaciji s tremi drugimi klimatološkimi meritvami.

Ključne besede: spremljanje kakovosti zraka, delci PM2.5, kompenzacija vlažnosti, strojno učenje, polinomska regresija, naključni gozd, DNN, analiza časovnih vrst

1 INTRODUCTION

In recent years, numerous solutions have been developed to monitor air quality with varying levels of precision and cost. When selecting a monitoring station, users are typically faced with a trade-off between low-cost sensors and high-precision instruments equipped with systems that compensate for the influence of environmental factors on measurement accuracy.

Among these factors, air humidity has been identified as having the greatest impact on measurement reliability. To ensure accurate measurements, professional monitoring stations often dry the air sample by heating it before analysis. However, this approach substantially increases both the construction complexity and operational costs of such stations. In Figure 1, the ratio of $PM_{2.5}$ measurements from the sensor without a heater to those from the sensor with a heater is shown as a function of relative humidity. If humidity had no effect on the measurements, both sensors (with and without the heater) would provide similar values (ratio should be close to one). However, it is clear that the measurements from the sensor without the heater exceed those from the heated sensor, especially at humidity levels above 50%.

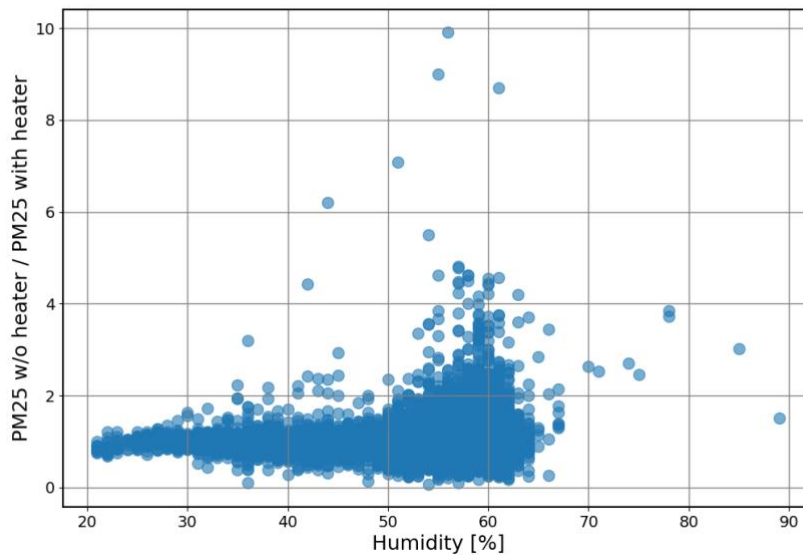


Figure 1: The ratio of $PM_{2.5}$ measurements from the sensor without a heater to those from the sensor with a heater is shown as a function of relative humidity.

To address these challenges, this project investigates data-driven modeling techniques capable of predicting the readings of high-end, humidity-compensated monitoring stations based on measurements from simpler, low-cost sensors. The target (dependent) variable is the $PM_{2.5}$ readings from the station with the heater activated. The input features consist of the unheated station measurements alongside three other climatological variables (temperature, humidity, and barometric pressure). The objective is to develop a mathematical calibration function that allows an unheated station to accurately simulate a heated one using software-based compensation.

2 CURRENT PROGRESS IN LOW-COST PARTICULATE MATTER SENSOR CALIBRATION

Recent research focusing on the calibration of low-cost particulate matter (PM) sensors has prioritized the mitigation of meteorological influences, particularly relative humidity (RH). In [2] authors have used a comparative analysis between computational Hanel growth models and low-cost thermal aerosol dryers for SDS011 sensors and got results showing that while hardware dryers effectively eliminate fog and hygroscopic growth artifacts regardless of location, mathematical corrections remain limited by the temporal and spatial stability of aerosol hygroscopic properties.

To address these physical limitations, in the study [3], authors have used a correction algorithm based on κ -Kohler theory applied specifically to the particle size distribution rather than total PM mass and got a significant reduction in measurement overestimation, bringing the error factor down from 5 to 1.05. Statistical and machine learning approaches have also seen significant advancement; for instance, in the research by Zusman et al. [4] the authors have used region-specific stepwise regression models for Plantower sensors and got high performance metrics ($R^2 = 0.80\text{--}0.97$), although they noted that climatological differences between cities severely limit the portability of a single universal model.

Furthermore, in the paper [5], the authors have used a hybrid model combining Stepwise Regression Analysis and Multi-Layer Perceptron (SRA-MLP) neural networks and got an accuracy increase of 42.5% to 86.5% compared to raw data. Finally, shifting toward autonomous systems, in the work by Casari and Po [6], the authors have used the 'Mith' framework to dynamically model hygroscopicity using historical rolling windows without relying on reference station data and got an R^2 improvement of 0.3 in challenging autumn conditions. These findings collectively highlight a shift from static linear calibrations toward dynamic, context-aware frameworks that can maintain reference equivalence in diverse environments.

In this study, we utilize both classical and deep machine learning methods to evaluate their respective capabilities in achieving reference equivalence for low-cost PM_{2.5} sensors. By comparing traditional statistical approaches with more advanced neural network architectures, our research aims to determine the extent to which increased model complexity translates into practical accuracy gains within our specific experimental setup.

3 METHODOLOGY

To collect the dataset, two custom stations were built based on a modified Nettigo Air Monitor design [7]. Each unit features an SDS011 particulate matter sensor [8] with an air intake coupled to a heating element (an SHT31 sensor [9] and a cermet resistor) enclosed in a reflective aluminum profile. This compact chamber utilizes airflow to reduce humidity without degrading measurement accuracy. For ambient monitoring, an SHT35 temperature/humidity sensor [9] and a BMP180 pressure sensor [10] are mounted externally to isolate them from the internal airflow. The system is controlled by a Wemos D1 Mini Pro board [11] with an ESP8266 micro controller [12]. While structurally identical, the stations use different operational logic: one dynamically activates the heater when relative humidity exceeds 63% (safeguarding the SDS011's 70% accurate operating threshold), whereas the other station's heater remains permanently deactivated. The dataset comprises 26456 measurement points collected between 2022-04-21 and 2022-10-31. The measurements were conducted in Upper Silesia, Poland, a region historically characterized by high levels of atmospheric pollution and industrial emissions. Dataset contains timestamp and meteorological data including humidity, temperature and pressure. Figure 2 shows the first five entries of the dataset, illustrating its structure and the types of parameters recorded.

	index	date	pm25_x	pm10_x	pressure	temperatureHeca_x	humidityHeca_x	pm25_y	pm10_y
0	1	2022-04-21 00:10:00	19.0	23.2	986	11.4	58	13.8	24.4
1	2	2022-04-21 00:20:00	16.7	20.6	986	11.2	53	14.4	27.0
2	3	2022-04-21 00:30:00	21.8	24.6	986	11.7	59	15.1	29.1
3	4	2022-04-21 00:40:00	21.8	25.8	986	11.3	57	18.0	33.7
4	5	2022-04-21 00:50:00	25.0	29.4	986	12.0	53	19.0	36.0

Figure 2: Five first data points in dataset

While using a heated low-cost station as a proxy ground truth effectively removes the confounding effects of humidity, it introduces several limitations. First, the thermal drying process causes volatile and semi-volatile pollution particles (such as ammonium nitrate, which is highly prevalent in Upper Silesia) to evaporate into gas, leading the heated station to underestimate the true particulate mass. Second, by artificially shrinking particles back to their dry core size, the setup measures a conditioned environment rather than the actual ambient air quality people breathe. Lastly, the physical stress of constant heating cycles accelerates the mechanical degradation and calibration drift of the heated sensor relative to the unheated control unit. Consequently, the heated station serves as an operational baseline to develop our mathematical calibration function, rather than an absolute, regulatory ground truth for real-world pollution.

The methods applied aimed to predict the readings from a station equipped with a heating element based on four features: readings from a station without air drying, humidity, temperature, and pressure. For modeling the data, various methods were tested, ranging from classical approaches to modern deep learning. Classical Methods included:

- Polynomial Regression**
 Polynomial regression [13] was employed to find potential nonlinear relationships between the input features and target variable. The optimal polynomial degree was found by testing nonlinear kernels from the second to the sixth degree.
- Random Forest**
 The Random Forest (RF) [14] algorithm is considered one of the most robust and versatile general-purpose regression methods [15]. Its strength lies in constructing an ensemble of multiple decision trees and computing the average of their individual predictions. This averaging mechanism significantly reduces variance and enhances model stability, effectively preventing overfitting and improving the generalization capability of the model structure and resistance to overfitting. For the Random Forest model, two variants were tested. The first variant was a baseline Random Forest using the default parameters provided by the Sklearn library [16]. The second variant involved hyper parameter tuning performed with GridSearchCV [17] using five-fold cross-validation to optimize the model's performance. Listing 1 shows the hyper parameter grid evaluated during the tuning process.

Listing 1: Hyper parameter grid used for tuning the Random Forest model

```
param_grid = {
    'n_estimators': [50, 100, 150, 200],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5],
    'min_samples_leaf': [1, 2],
    'max_features': ['auto', 'sqrt']
}
```

- Stacked Classical models**
 The stacking models were another attempt to obtain a more accurate model. According to [18], the approach used can improve the effectiveness of classical models through

their repeated execution or combination. Stacking involved generating the output from the first model and then adding it as another feature in the data passed to the next model. To maintain compatibility with other models, dedicated classes were created for each stack model. These classes implement the fit and predict methods, ensuring that stacked models can be seamlessly integrated into standard Sklearn pipelines. Listing 2 shows an example structure of such a class, which is representative of all stacked model implementations used in this project.

Listing 2: Example class structure used to implement stack models

```
class rf_lr_model:
    def __init__(self, rf_model, lr_model, lr_degree = 1):
        self.rf_model = rf_model

        self.lr_model = lr_model
        self.poly = PolynomialFeatures(degree= lr_degree)

    def fit(self, X_train, y_train):
        self.rf_model.fit(X_train, y_train)
        y_temp_pred = self.rf_model.predict(X_train)

        X_train_lr = X_train.copy()
        X_train_lr[:, 0] = y_temp_pred

        X_train_poly = self.poly.fit_transform(X_train_lr)
        self.lr_model.fit(X_train_poly, y_train)

    def predict(self, X_test):
        y_temp_pred = self.rf_model.predict(X_test)
        X_test_lr = X_test.copy()
        X_test_lr[:, 0] = y_temp_pred

        X_test_poly = self.poly.transform(X_test_lr)
        return self.lr_model.predict(X_test_poly)
```

Additionally, neural networks were employed using keras library [19]. The hyper parameters of the models were found using keras_tuner [20] with Bayesian optimization [21].

Final evaluation was conducted on the held-out test set using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE) and Coefficient of Determination (R^2). This approach allowed comparison of model types to select the best-performing model for predicting accurate readings of levels of PM_{2.5} particles.

It is important to note that MSE tends to be large even when there are few but significant outliers or large individual errors, because it squares the errors, disproportionately penalizing large deviations. Therefore, while MAE gives an average magnitude of errors treating all equally, MSE emphasizes the impact of larger prediction errors, which can be critical in assessing model performance when rare but large errors exist. This property of MSE makes it sensitive to occasional large prediction mistakes, highlighting them more than MAE does.

4 RESULTS

In order to determine the optimal polynomial regression model, a series of linear regression models enhanced with polynomial feature mappings of increasing degrees was trained and evaluated. Increasing the polynomial degree enabled a systematic assessment of the dataset's complexity and facilitated the identification of the configuration that achieved the best predictive performance. Table 1 presents the evaluation metrics for polynomial regression models of degrees from 2 to 6. The best performance was obtained using the fifth-degree polynomial model, which achieved the highest R^2 score and the lowest MSE among the tested

configurations: $R^2 = 0.891$, $MAE = 1.238$, $MSE = 5.792$. Figure 3 shows the relationship between the true target values and the predictions generated by the fifth-degree polynomial regression model. The plot provides a direct visual assessment of how closely the model's output aligns with the actual values.

Table 1: Results of polynomial regression models for different polynomial degrees.

Degree	R^2	MAE	MSE
2	0.871	1.419	6.822
3	0.884	1.313	6.134
4	0.888	1.259	5.919
5	0.891	1.238	5.792
6	0.886	1.221	6.038

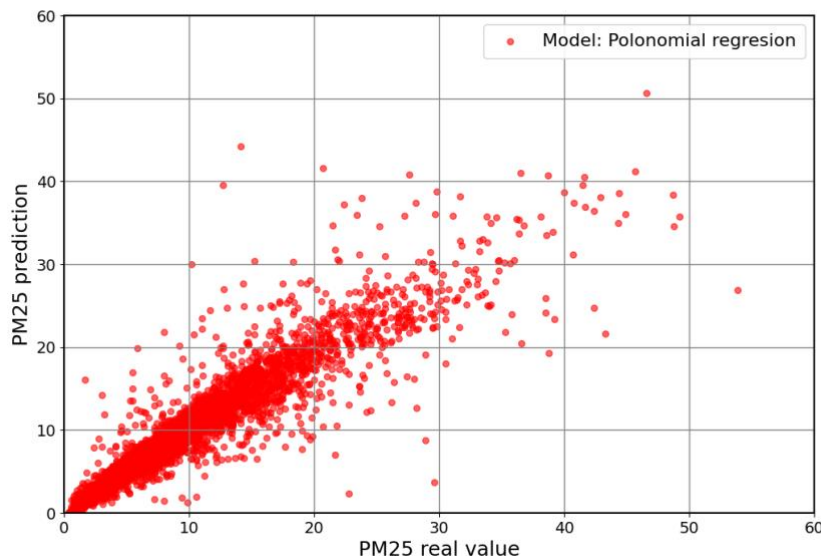


Figure 3: Predicted by the fifth-degree model values vs true values.

Obtaining results for the Random Forest models involved conducting two separate training processes. The first model was trained using the default parameters and served as the baseline for comparison. The second model was developed to identify the optimal set of hyper parameters using the GridSearchCV [17] tuning method. The following metrics were obtained for the discussed models: Baseline Random Forest model: $R^2 = 0.890$, $MAE = 1.151$, $MSE = 5.822$. Random Forest tuned with GridSearchCV: $R^2 = 0.898$, $MAE = 1.100$, $MSE = 5.410$.

Figure 4 compares the performance of the baseline Random Forest model with the version optimized using GridSearchCV. In both sub figures, the vertical axis represents the model predictions, while the horizontal axis displays the true target values. Points closer to the diagonal line indicate higher predictive accuracy.

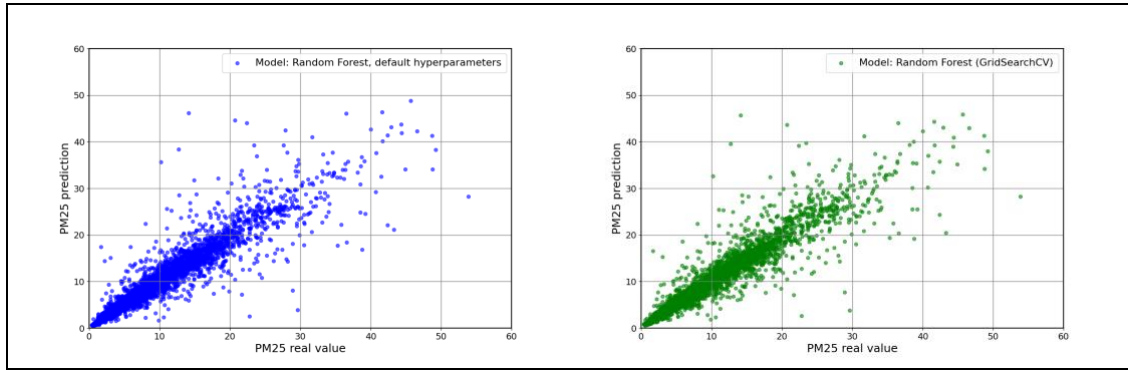


Figure 4: Comparison of true values and values predicted by the Random Forest Model, left: default hyper parameters, right: model tuned by the GridSearchCV.

The last group of classical models consisted of stacked models. The stacked calibration architecture employs a sequential two-stage modeling approach to enhance sensor data quality. In the primary stage, an initial model is trained to establish fundamental quantitative relationships, producing a set of preliminary predictions. These fitted values are then utilized as an additional input feature alongside the original raw measurements to train a more complex secondary model. In our stacked models, we utilized various combinations of polynomial regression, investigating different degrees, and the random forest regressor.

Owing to the substantial number of stacking models developed and evaluated, the result presented here focuses exclusively on the top-performing configuration. The selected model combines a Random Forest regressor with a second-degree Polynomial Regression model as part of the stacking architecture. This setup achieved the highest evaluation metrics among all tested stacking configurations: $R^2 = 0.899$, MAE = 1.099, MSE = 5.342

A simple dense neural network was trained as a baseline. The network structure consisted of four fully connected layers. The input and first hidden layer contained 128 units utilizing the Rectified Linear Unit (ReLU) activation function. This was followed by two additional hidden layers with 64 and 32 neurons, respectively, to refine feature representation before reaching a final output layer with a single neuron dedicated to predicting PM25 mass concentrations. The training process utilized the Adam (Adaptive Moment Estimation) optimization algorithm and the Mean Squared Error (MSE) as the primary loss function. Additionally, the Mean Absolute Error (MAE) was monitored as a key performance metric. The model was trained over 50 epochs with a batch size of 256. To mitigate the risk of overfitting and ensure the stability of the calibration model, an internal validation procedure was applied, reserving 20% of the training data as a validation set. Trained model achieved the following metrics: $R^2 = 0.888$, MAE = 1.229, MSE = 5.968. In Figure 5 the training curves and the relationship between the true target values and the predictions generated by the simple dnn model. The vertical axis represents the model predictions, while the horizontal axis displays the true values.

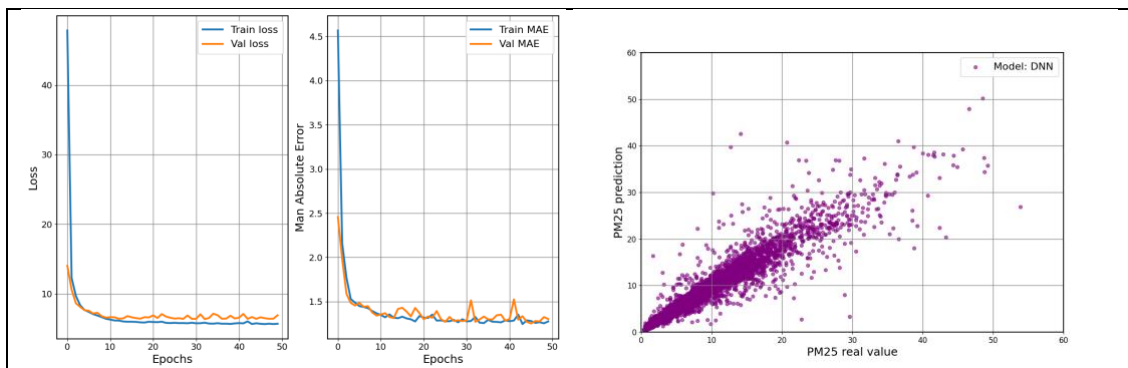


Figure 5: Left panel: Learning curves of simple DNN model: Loss vs number of epoch, mean absolute value vs number of epoch. Blue curves represent training, orange ones the validation. Right panel: Predicted by the model values vs true values.

The results for all next deep neural network models were obtained using a Bayesian optimization tuner. Despite the significantly greater time investment dedicated to tuning, the deep neural networks were unable to outperform the best stacking model. Among the dense models, the best results were achieved by rather small models not exceeding 3 layers - DNNs started to overfit very quickly.

Overfitting often results from overly complex models relative to the amount and diversity of training data available, causing them to memorize the training set. Incorporating residual connections into the Deep Neural Network (DNN) architecture proved to be an effective strategy for overcoming overfitting and improving overall model robustness. These connections facilitated the propagation of gradients through deeper layers, reducing the risk of vanishing or exploding gradients during training. The architecture accepts a 4-dimensional input vector and processes it through a series of three sequential residual blocks before reaching a single-unit linear output layer for continuous scalar prediction. To facilitate robust feature extraction and prevent gradient degradation, each residual block implements a dual-path topology: a primary computational path consisting of a fully connected (Dense) layer followed by a Dropout regularization layer to mitigate overfitting, and a parallel projection shortcut (Dense layer) that matches tensor dimensions. These paths are merged via element-wise addition (Add) and mapped through a non-linear activation function. The network structurally alters its latent dimensionality across these blocks - expanding from 4 to 32 units in the first block, compressing to a 16-unit bottleneck in the second block, and re-expanding to 32 units in the third block. While this complex architecture successfully avoided overfitting—as verified by the convergence of the training and validation learning curves—it resulted in negligible performance gains on the test dataset ($R^2 = 0.893$, MAE = 1.235, MSE = 5.693).

While recurrent neural networks with LSTM layers initially yielded highly promising results, they exhibited a strong tendency toward overfitting, indicating that further investigation and regularization are required to ensure generalizability.

5 CONCLUSION

This study evaluated several modeling approaches—ranging from traditional statistical methods to advanced machine learning and deep learning architectures—to predict station readings using climatological features. To facilitate a clear performance comparison, the evaluation metrics for all investigated models are summarized in Table 2.

Table 2: Performance Comparison of the Investigated Models

Model Architecture	R^2	MAE	MSE
5th-Degree Polynomial Regression	0.891	1.238	5.792
Baseline Random Forest	0.890	1.151	5.822
Tuned Random Forest (GridSearchCV)	0.898	1.100	5.410
Stacked Ensemble (Best Performing)	0.899	1.099	5.342
Simple DNN	0.888	1.229	5.968
Residual DNN	0.893	1.235	5.693

The empirical results offer critical insights into our primary research objective: determining the extent to which increased model complexity translates into practical accuracy gains within this experimental setup. It should be noted that this evaluation was strictly bounded by our specific local dataset; while the models were not validated against data from alternative geographic locations, spatial generalization across different microclimate fell outside the immediate scope

of this research. Within this localized framework, the ensemble approach utilizing stacked models achieved the highest overall performance ($R^2 = 0.899$, MAE = 1.099, MSE = 5.342), closely followed by the hyper parameter-tuned Random Forest regressor.

Interestingly, escalating structural complexity via deep learning did not yield proportional performance benefits. While the Residual DNN architecture ($R^2 = 0.893$) successfully mitigated some optimization degradations compared to the simple DNN ($R^2 = 0.888$), both deep neural networks were outperformed by the simpler, optimized tree-based ensembles. Furthermore, initial investigations into recurrent neural networks (RNNs) yielded highly promising metrics within the training phase; however, they exhibited a severe tendency toward overfitting, rendering them less generalizable even within our local context without further regularization.

In summary, for our specific environmental monitoring framework, the marginal accuracy gains provided by complex deep learning and recurrent structures do not justify their substantial computational overhead. Instead, optimized tree-based ensembles and stacked architectures offer the most robust, accurate, and practically viable solution for predicting the target station readings. Future work will focus on implementing advanced regularization techniques to stabilize the recurrent networks and exploring alternative feature engineering strategies to determine if these localized findings hold true across broader regional networks.

6 ACKNOWLEDGMENT

This work was supported by the AI2SEP project (No. 2023-1-PL01-KA220-HED 000166765).

References

- [1] R. Jayaratne, X. Liu, P. Thai, M. Dunbabin, and L. Morawska, "The influence of humidity on the performance of a low-cost air particle mass sensor and the effect of atmospheric fog," *Atmospheric Measurement Techniques*, vol. 11, no. 8, pp. 4883–4890, 2018.
- [2] B. Laquai and B. Kroseberg, "Comparison of a computational method for correcting the humidity influence with the use of a low-cost aerosol dryer on a SDS011 low-cost PM-sensor." 2021.
- [3] A. Di Antonio, O. A. Popoola, B. Ouyang, J. Saffell, and R. L. Jones, "Developing a relative humidity correction for low-cost sensors measuring ambient particulate matter," *Sensors*, vol. 18, no. 9, p. 2790, 2018.
- [4] M. Zusman *et al.*, "Calibration of low-cost particulate matter sensors: Model development for a multi-city epidemiological study," *Environment international*, vol. 134, p. 105329, 2020.
- [5] B. Liu, Q. Zhao, Y. Jin, J. Shen, and C. Li, "Application of combined model of stepwise regression analysis and artificial neural network in data calibration of miniature air quality detector," *Scientific reports*, vol. 11, no. 1, p. 3247, 2021.
- [6] M. Casari and L. Po, "MitH: A framework for mitigating hygroscopicity in low-cost PM sensors," *Environmental Modelling and Software*, vol. 173, p. 105955, 2024, doi: <https://doi.org/10.1016/j.envsoft.2024.105955>.
- [7] N. A. Monitor, "Nettigo air monitor - knowledge base." 2018-2023. Available: <https://www.air.nettigo.pl/baza-wiedzy/?lang=en>
- [8] *SDS011 laser PM2.5 sensor specification*. Nova Fitness Co., Ltd., 2015.
- [9] *Datasheet SHT3x-DIS humidity and temperature sensor*. Sensirion, 2022.

- [10] *BMP180 digital pressure sensor*. Bosch Sensortec, 2013.
- [11] Wemos.cc, “D1 mini pro - WEMOS documentation.” https://www.wemos.cc/en/latest/d1/d1_mini_pro.html, 2021.
- [12] *ESP8266 technical reference*. Espressif, 2020.
- [13] S. Brandt and S. Brandt, *Data analysis*. Springer, 1998.
- [14] L. Breiman, “Bagging predictors,” *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.
- [15] T. Zhu, “Analysis on the applicability of the random forest,” in *Journal of physics: Conference series*, IOP Publishing, 2020, p. 012123.
- [16] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [17] “GridSearchCV.” https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html, 2019.
- [18] B. Pavlyshenko, “Using stacking approaches for machine learning models,” in *2018 IEEE second international conference on data stream mining & processing (DSMP)*, IEEE, 2018, pp. 255–258.
- [19] F. Chollet *et al.*, “Keras.” <https://keras.io>, 2015.
- [20] T. O'Malley *et al.*, “KerasTuner.” <https://github.com/keras-team/keras-tuner>, 2019.
- [21] “BayesianOptimization tuner.” https://keras.io/keras_tuner/api/tuners/bayesian/.

Krzysztof Emerling - holds an engineering degree in Computer Science and is currently pursuing a Master's degree in Applied Computer Science. His research and technical interests focus on artificial intelligence, machine learning, computer graphics, game engine development, and C++/Python programming. Beyond academia, he applies modern technologies through personal software development, competitive chess, and calligraphy.

Patryk Sitarek - Graduate in Computer Science (B.Eng.) and Micro and Nanotechnology (M.Sc.). He is professionally involved in the IT consulting sector, focusing on the design and optimization of large-scale cloud data processing. Beyond his core data engineering tasks, he enjoys exploring AI tools and adapting them for data processing.

Katarzyna Schmidt - She holds a Ph.D. in experimental physics and has been actively involved in heavy-ion collision experiments at intermediate energies (with Texas A&M University) as well as high-energy particle physics research at CERN. Her current research focuses on applying machine learning techniques to experimental physics. She is a lecturer at the University of Silesia in Poland.