DOI: 10.17951/b.2025.80.0.265-283

ANNALES
UNIVERSITATIS MARIAE CURIE-SKŁODOWSKA
LUBLIN – POLONIA

VOL. LXXX SECTIO B 2025

JAROSŁAW BERNACKI
https://orcid.org/0000-0002-4488-3488
Department of Computer Science and Systems Engineering
Wrocław University of Science and Technology
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
jaroslaw.bernacki@pwr.edu.pl

Particulate Matter Forecasting Using Hybrid Autoencoders: The Role of Meteorological Data

Modelowanie i prognozowanie pyłu zawieszonego z wykorzystaniem hybrydowych autoenkoderów i danych meteorologicznych

Abstract: Accurate forecasting of particulate matter (PM) concentrations is crucial for effective air quality management and public health protection. In this study, we propose two deep learning-based forecasting models: a convolutional-recurrent autoencoder and a hierarchical autoencoder. Both models are trained and evaluated using historical data on $PM_{2.5}$ and PM_{10} concentrations from multiple monitoring stations in Poland. To assess the influence of meteorological conditions on prediction accuracy, the models are tested in two variants: one using historical PM concentrations and meteorological features such as temperature, wind speed, wind direction, and air pressure, and another that uses only historical PM data. The results clearly show that the inclusion of weather data significantly improves forecasting performance, with lower MAE, and MSE values observed across all test sites. The models trained with meteorological inputs consistently outperform their counterparts trained on PM data alone. The results are also compared with the baseline model. These findings highlight the importance of environmental context in air pollution forecasting and demonstrate the potential of autoencoder-based approaches for this task.

Keywords: air pollution; forecasting; deep models; particulate matter

Abstrakt: Prognozowanie stężeń pyłu zawieszonego ma kluczowe znaczenie dla efektywnego zarządzania jakością powietrza i ochrony zdrowia publicznego. W opisanym badaniu proponujemy dwa modele prognostyczne oparte na uczeniu głębokim: autokoder konwolucyjno-rekurencyjny oraz autokoder hierarchiczny. Oba modele są trenowane i oceniane z wykorzystaniem danych historycznych dotyczących stężeń pyłu $PM_{2.5}$ i PM_{10} z kilku stacji monitorujących jakość powietrza w Polsce. Aby ocenić wpływ warunków meteorologicznych na dokładność prognoz, modele testowano w dwóch

wariantach – jeden wykorzystuje historyczne stężenia pyłu zawieszonego oraz dane meteorologiczne, takie jak temperatura, prędkość i kierunek wiatru oraz ciśnienie powietrza, a drugi wyłącznie historyczne dane dotyczące pyłu zawieszonego. Wyniki pokazują, że uwzględnienie danych pogodowych znacząco poprawia skuteczność prognozowania, z niższymi wartościami MAE i MSE zaobserwowanymi we wszystkich lokalizacjach testowych. Modele trenowane z wykorzystaniem danych meteorologicznych konsekwentnie przewyższają swoje odpowiedniki trenowane wyłącznie na danych dotyczących pyłu zawieszonego. Uzyskane wyniki zostały porównane z wynikami modelu bazowego. Obserwacje te podkreślają znaczenie kontekstu środowiskowego w prognozowaniu zanieczyszczenia powietrza oraz pokazują potencjał podejść opartych na autokoderach.

Słowa kluczowe: zanieczyszczenie powietrza; prognozowanie; modele głębokie; pył zawieszony

INTRODUCTION

Particulate matter (PM) is an air pollutant that consists of small solid and liquid particles suspended in the atmosphere. The most commonly distinguished are the PM_1 with particles of a diameter less than 1 micrometer, $PM_{2.5}$ with particles with a diameter of less than 2.5 micrometers, and PM_{10} with particles with a diameter of less than 10 micrometers. These microscale particles can penetrate deep into the lungs, and in the case of PM_1 and $PM_{2.5}$, also into the bloodstream, which poses a serious health risk, leading to respiratory and cardiovascular diseases and cancer. Due to its harmful health effects, PM is one of the most important environmental problems in the world. Many countries, especially those with a high level of urbanization and industrial development, are trying to reduce PM emissions by introducing strict regulations, developing air purification technologies, and monitoring air quality. Effective forecasting of PM concentration is crucial for taking action to prevent its negative impact on public health and improving the quality of life of city dwellers (Polichetti et al., 2009).

In Poland, PM, especially $PM_{2.5}$ and PM_{10}, is a serious environmental and health problem. Its high concentrations in the air occur mainly during the heating season when intensive home heating contributes to the emission of pollutants from chimneys, as well as a result of road transport and industry. According to reports by the WHO and national institutions, Poland is one of the countries with the highest level of air pollution in the European Union, which negatively affects the health of citizens, especially in cities with high population density (Jasiński et al., 2021). Health effects, such as respiratory and cardiovascular diseases and an increased number of premature deaths, are directly related to exposure to PM. In response to this problem, Poland has taken a number of actions, including the introduction of air quality standards, the modernization of heating systems, and the development

of an air quality monitoring network. Despite this, effective reduction of PM emissions remains one of the main challenges for the country's environmental protection policy. The $PM_{2.5}$ and PM_{10} (but also other air pollutants) concentrations are monitored on an ongoing basis by the Chief Inspectorate of Environmental Protection (GIOS)[1], which provides data online (Penkała et al., 2023, Połednik, 2022).

Technical challenges at air quality monitoring stations often result in missing data. To tackle this problem, an effective strategy is to estimate the air pollutants' missing values, which is a task closely tied to forecasting. A variety of studies have been conducted on forecasting air pollutant concentrations (Sowka et al., 2019), utilizing diverse methodologies such as machine learning (Gryech et al., 2024), deep learning (Swetha et al., 2024), and hybrid models (Borah et al., 2024).

In this study, we focus on forecasting the concentrations of $PM_{2.5}$ and PM_{10} using advanced deep learning techniques. Specifically, convolutional-recurrent autoencoders and hierarchical autoencoders models are employed to predict future PM levels based on historical data, including meteorological variables such as temperature, wind speed, wind direction, and air pressure. The choice of these models is motivated by their ability to capture complex nonlinear patterns in air quality data, which traditional methods may struggle with. To evaluate the performance of these models, commonly used metrics such as mean absolute error (MAE), and mean squared error (MSE) are used, ensuring a comprehensive assessment of prediction accuracy. The experiments are conducted on the historical data collected from the following Polish cities: Zielonka (Bory Tucholskie), Zakopane, Częstochowa, Gliwice, and Szczecin.

The forecasting simulations are evaluated twofold: one using historical PM concentrations and meteorological features such as temperature, wind speed, wind direction, and air pressure, and another that uses only historical PM data. The use of meteorological data alongside historical PM concentrations leads to significantly improved forecasting results. Models trained with weather variables consistently achieved lower prediction error values compared to those using only PM historical concentrations. This demonstrates that atmospheric factors play a key role in shaping air pollution dynamics, and their inclusion enhances the models' ability to capture complex temporal patterns and variability in PM levels. In addition, we implemented a state-of-the-art random forest model and tested it analogously to the proposed autoencoders, comparing its performance with and without meteorological inputs under the same experimental setup.

[1] https://www.gov.pl/web/gios-en

CONTRIBUTION

The main contribution of this study lies in the application of advanced deep learning models for forecasting PM concentrations, specifically $PM_{2.5}$ and PM_{10}, using meteorological data. This research evaluates the effectiveness of convolutional-recurrent and hierarchical autoencoder architectures in capturing complex nonlinear dependencies in air pollution data. Meteorological variables are crucial in this context, as weather conditions strongly influence pollutant dispersion, transport, and accumulation. For example, wind speed and direction can disperse pollutants or carry them across regions, significantly altering local concentration levels. By comparing models trained with and without meteorological inputs, the study highlights the added value of utilizing atmospheric variables in prediction tasks. A thorough evaluation using standard performance metrics such as MAE and MSE on historical data demonstrates the potential of the proposed approach to improve the accuracy of air quality forecasts and support environmental and public health decision-making.

ORGANIZATION OF THE PAPER

The paper is organized as follows. The next section describes previous and related works. Section 3 presents proposed deep learning methods for forecasting the concentration of $PM_{2.5}$ and PM_{10}. In section 4, the experimental evaluation of the proposed methods is performed. Finally, the last section concludes this work. Everywhere in the paper, bold font in formal descriptions denotes matrices or vectors.

PREVIOUS AND RELATED WORK

In Li et al. (2020), a hybrid CNN-LSTM model combining a convolutional neural network (CNN) and an LSTM neural network is developed to forecast the $PM_{2.5}$ concentration for the next 24 hours in Beijing. This model takes advantage of both networks: CNN effectively extracts air quality-related features, and LSTM takes into account the long-term temporal context of the data. The model is fed with the air quality data of the past 7 days, and the output is the forecast of the next day's $PM_{2.5}$ concentration. This study compares four models: univariate LSTM, multivariate LSTM, univariate CNN-LSTM, and multivariate CNN-LSTM. The results show that the proposed multivariate CNN-LSTM model achieves the best performance with low error and short training time.

The study presented by Mauricio et al. (2024) investigates the use of a transformer-based model for forecasting coarse particulate matter (PMCO) concentrations in Mexico City over several horizons. The transformer outperformed ARIMA and LSTM models across multiple stations, though challenges remain for short-term forecasts and areas with high variability near industrial zones.

The work depicted by Harishkumar et al. (2020) determines the concentration of PM in the atmosphere to improve public health. In this study, machine learning models were used to predict the concentration of particulate matter based on data from the air quality monitoring system in Taiwan from 2012 to 2017. These models showed better performance compared to traditional forecasting approaches. The results were evaluated using statistical measures such as RMSE, MAE, MSE, and R^2, which confirms their higher accuracy and usefulness in air quality forecasting.

In Won et al. (2021), the hygroscopic properties of PM, which affect light scattering and absorption are analyzed. This study explores the behavior of coarse PM (CPM) and fine PM ($PM_{2.5}$) under varying weather conditions and their influence on visibility. A censored regression model was developed to analyze the relationship between PM concentrations and meteorological factors, enabling the calculation of the optical hygroscopic growth factor and hygroscopic mass growth. These metrics were applied to $PM_{2.5}$ field measurements using low-cost sensors in two distinct regions. The results indicate that CPM and $PM_{2.5}$ concentrations negatively impact visibility, with relative humidity (RH) significantly modulating these effects. The modeled hygroscopic growth factors aligned closely with observed values, particularly under haze and mist conditions. Adjusting $PM_{2.5}$ concentrations for RH based on visibility-derived growth factors notably improved the accuracy of low-cost PM sensors. This study underscores the importance of accounting for PM-meteorology interactions in both visibility forecasting and sensor calibration.

The paper presented by Kowalski et al. (2020) evaluates the feasibility of using modern neural networks to forecast PM_{10} concentrations over a 24-hour horizon. This paper analyzes various error measures and compares them with other prediction methods. The results demonstrate that the proposed convolutional neural network can be an effective tool for detailed air quality forecasts.

In the work depicted by Tong et al. (2024), a transformer-based model for spatiotemporal forecasting of $PM_{2.5}$ concentrations is proposed. Using aerosol optical depth data from California, the model outperforms LSTM and other baselines across multiple error metrics, demonstrating strong robustness and stability in complex prediction scenarios.

The study described by Nidzgorska-Lencewicz (2018) used artificial neural networks to forecast hourly PM_{10} concentrations in the Tricity area (Poland) during the winter periods of 2002/2003–2016/2017. Models based on measurement data and

meteorological factors achieved satisfactory results (R^2 ranging from 0.452 to 0.848), confirming the effectiveness of ANNs in predicting air pollution levels.

The paper presented by Qin et al. (2014) deals with the forecasting of PM concentrations to support regulatory planning, emission reduction, and early warning system implementation. The authors propose the CS-EEMD-BPANN model, which combines gray correlation analysis (GCA), ensemble empirical mode decomposition (EEMD), cuckoo search algorithm (CS), and backpropagation neural networks (BPANN). The unique element is the use of GCA to identify potential predictors of PM among other air pollutants (CO, NO_2, O_3, SO_2) and meteorological variables (wind speed and direction, temperature, humidity, pressure). The analysis results indicate that CO, NO_2, and SO_2 are strongly related to PM, and including these predictors significantly improves the model performance, which confirms the effectiveness of the proposed approach.

In Czernecki et al. (2021), the application of machine learning methods to short--term forecasting of PM_{10} and $PM_{2.5}$ concentrations in four large Polish agglomerations, based on 10 years of measurement data from background, traffic, and industrial stations is presented. Four models were tested: stepwise regression, random forests, XGBoost, and neural networks, demonstrating that XGBoost performed best, while stepwise regression performed worst. The results confirm the high utility of machine learning algorithms in predicting air pollution episodes and emphasize the importance of meteorological variables in air quality modeling.

The study described by Rogula-Kozłowska et al. (2014) examines seasonal changes in $PM_{2.5}$ concentration and chemical composition at three locations in Poland (Katowice, Gdańsk, and Diabla Góra). The results indicate very high $PM_{2.5}$ concentrations, especially in winter in Katowice, where the dominant components were carbon compounds, secondary aerosols, and benzo(*a*)pyrene (BaP), posing the greatest health risk during the heating season.

In the work presented by Vovk et al. (2024), a hybrid approach that combines the EMEP4PL chemical transport model with random forest to improve the accuracy of $PM_{2.5}$ estimates in Poland, where CTMs alone often underestimate concentrations, is proposed. The model integrates outputs with meteorological, emission, land use, and temporal predictors, and achieved substantially higher accuracy ($R^2 = 0.71$ vs. 0.38 for EMEP4PL), reduced bias, and better detection of severe pollution episodes.

The study described in Tariq et al. (2023) addresses the challenge of forecasting fine particulate matter ($PM_{2.5}$) concentrations in underground subway stations, where air quality is strongly affected by nonlinear dynamics and uncertainty. The authors propose a hybrid framework that combines a gated probabilistic transformer with a genetic algorithm-based quantile scheduling approach, improving both early warning accuracy and ventilation control. Compared to conventional methods, their

model achieves narrower prediction intervals, higher accuracy in sequential health risk assessment, and measurable reductions in $PM_{2.5}$ levels, demonstrating its potential for smart building applications.

The study by Kujawska et al. (2022) compares various machine learning models for forecasting PM_{10} concentrations in Lublin, using data from 2017–2019, including meteorological and chemical parameters. The artificial neural network achieved the best results, achieving $R = 0.89$, effectively predicting the risk of PM_{10} exceedances.

In Cichowicz et al. (2020), the analysis of PM_{10} concentrations in the winter (December–January) in 2009–2015 in Poland showed that low wind speeds are conducive to higher levels of air pollution. The study included data from three monitoring stations in the region, enabling the identification of smog episodes (so-called black smog) and verification of measurements at the station located near a large incineration plant. The results confirm the influence of wind direction on the movement of pollutants or their local deposition.

In the paper presented by Zeng et al. (2023), a hybrid forecasting model that combines CEEMDAN decomposition with a deep transformer architecture to improve long-term $PM_{2.5}$ prediction is proposed. By introducing a specialized embedding layer and a non-autoregressive multi-step decoder, the model reduces error accumulation and significantly outperforms conventional RNN-based approaches on public datasets.

The study presented by Kryza et al. (2019) discusses an air quality forecasts for Poland using the WRF-Chem model, comparing the baseline approach (GENEMIS) with the heating degree day (HDD) method. Both approaches yielded similar results, with HDD better representing concentrations on warmer days, while the baseline method tended to overestimate them.

The work by Ramentol et al. (2023) concerns the prediction of nitrogen dioxide NO_2 concentrations using machine learning methods, using historical pollutant data and meteorological variables. The study focuses on the city of Erfurt (Germany) and includes modeling of temporal dependencies using embedded variables that allow the model to take into account local events such as traffic or specific occasions. Experiments show that the proposed model achieves better forecasting accuracy compared to other models, especially during periods when traditional seasonal patterns change, such as holidays.

In Du et al. (2022), a new hybrid $PM_{2.5}$ concentration forecasting system, which also takes into account the evaluation of health effects and economic losses, is discussed. First, an efficient data analysis method was used to extract the main features of the $PM_{2.5}$ dataset, minimizing the influence of noise. Then, the Harris hawk optimization algorithm was introduced to tune the extreme learning machine (ELM) model, which achieved high prediction accuracy. The health effects and economic costs of pollution were estimated based on the predicted $PM_{2.5}$ values. Experiments

were conducted on the real data from Beijing, Tianjin, and Shijiazhuang, and the results showed that the proposed system can support environmental management, emission reduction, and health prevention, and can be applied to various fields such as $PM_{2.5}$ health research.

The work proposed by Kouziokas et al. (2020) introduces a novel support vector machine (SVM) kernel by combining the transformed weight vector of a particle swarm neural network (ANN) optimized by the particle swarm algorithm (PSO) into a Bayesian-optimized SVM kernel for the task of forecasting the concentration of PM_{10} as time series. The proposed model shows higher forecast accuracy compared to the traditional ANN and SVM-optimized models, which is confirmed by the experimental results.

The study presented in Sharma et al. (2020) presents a hybrid deep learning CLSTM model combining a convolutional neural network (CNN) with an LSTM network to predict hourly total suspended particle (TSP) concentrations in Australia. The CLSTM model has been extensively tested and shows better performance than an ensemble of five other machine learning models.

PROPOSED METHODS

In this section, we describe the proposed deep learning architectures for air pollutant concentration prediction. The proposed methods include a convolutional-recurrent autoencoder and a hierarchical autoencoder. All the proposed models are used to predict the PM concentration, including the weather conditions, such as the temperature, wind speed, wind direction, and air pressure. More formally, the $\boldsymbol{x}_t$ is the input data, containing PM historical concentration $\boldsymbol{x}^{conc}$, temperature $\boldsymbol{x}^{temp}$ expressed in °C, wind speed $\boldsymbol{x}^{wind}$ (km/h), wind direction $\boldsymbol{x}^{wdir}$ (degrees), and air pressure $\boldsymbol{x}^{pres}$ (hPa) in a time step t, and is defined as follows:

$$\boldsymbol{x}_t = [\boldsymbol{x}^{conc}, \boldsymbol{x}^{temp}, \boldsymbol{x}^{wind}, \boldsymbol{x}^{wdir}, \boldsymbol{x}^{pres}]$$

The output of the model is a tensor of shape (1,1), representing a single forecasted value of PM_{10} or $PM_{2.5}$ concentration at the next time step. Let us define the sigmoid and hyperbolic tangent functions. The sigmoid function is defined as eq. 1 and the hyperbolic tangent function is defined as eq. 2.

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{1}$$

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{2}$$

CONVOLUTIONAL-RECURRENT AUTOENCODER

The convolutional-recurrent autoencoder architecture integrates the advantages of convolutional (for spatial feature extraction) and recurrent (for modeling temporal dependencies) layers (Zheng & Zhang, 2023). The model architecture consists of two main parts: an encoder and a decoder. In the encoder part, the data passes through two convolutional (1D) layers that filter local patterns along the time axis within the features, allowing the capture of local dependencies between different variables (e.g. the effect of wind speed on pollutant concentration). The result of this operation is then processed by the xLSTM unit, which models global temporal dependencies, producing internal representations of reduced dimension.

In a classic autoencoder, the goal of the decoder is to reconstruct the input, and the output is a reconstruction of the same input. In the proposed case, the decoder will be configured to generate predictions for the next values instead of reconstructing the entire input sequence. This is realized by utilizing the linear layers used in the decoder, which allow the model to directly predict the values of future time steps. Thus, in this version, the output is the future values of the time series. The model is trained using the mean squared error (MSE) as the loss function, and the optimization procedure seeks to minimize this loss by reducing the difference between the forecasts and the observed values. This allows the model to learn a representation of the features specific to pollution and weather data, which is crucial for successful predictions. In the prediction process, the network receives historical data as input and generates predictions for subsequent values in the time series. The structure of the considered autoencoder is presented as Fig. 1.

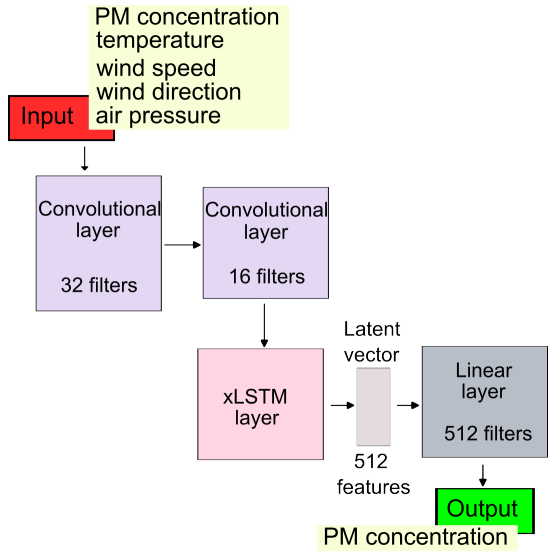

Fig. 1. The structure of the convolutional-recurrent autoencoder (Source: Author's own study)

HIERARCHICAL AUTOENCODER

Hierarchical autoencoder is a deep learning model designed for temporal data analysis and forecasting (Tran et al., 2021). The model uses two main components: an encoder and a decoder, which work together to compress and reconstruct the input data. The encoder transforms the input data, consisting of a time series of air pollutant concentration, temperature, wind speed, and direction values, into a lower-dimensional representation. This process allows for capturing the most relevant features of the data while eliminating redundancy and noise.

The encoder consists of several linear layers interspersed with ReLU activation functions that reduce the dimensions of the input data until a bottleneck is reached. At this point, the data is compressed to its most critical representations. The decoder then reverses the encoding process, transforming the compressed representation back to the original dimensions of the input data using successive linear and activation layers. The final decoder layer applies a sigmoid activation function, which ensures the stability and interpretability of the results.

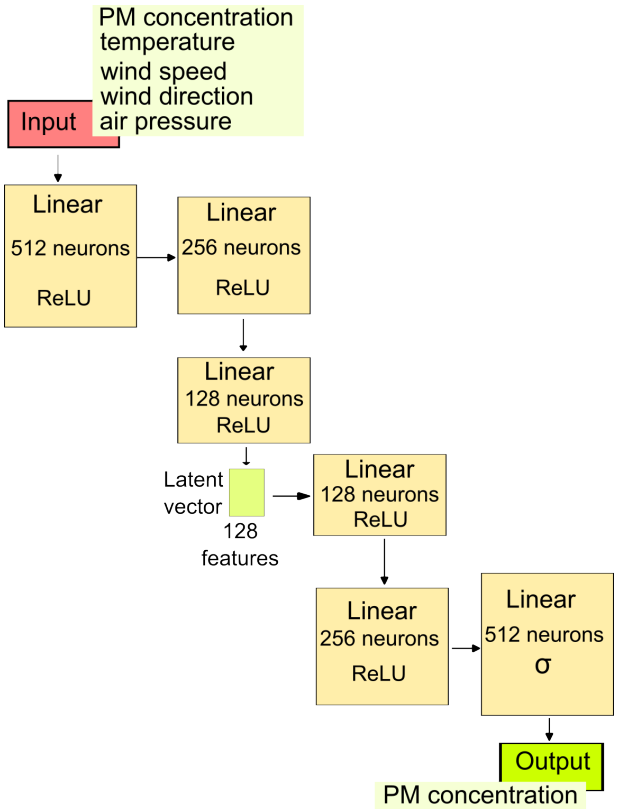

Fig. 2. The structure of the hierarchical autoencoder (Source: Author's own study)

The model is trained using a mean squared error (MSE) loss function, which measures the difference between the original input data and its forecasts. The Adam optimizer updates the network weights, minimizing the forecasting error. The training process involves iteratively tuning the network parameters, allowing the model to efficiently learn a representation of the input data. Once training is complete, the model is able to generate predictions for new data by reconstructing its key features based on the learned structure.

In the case of pollutant concentration forecasting, the model generates output by generating a time sequence that is shifted in time, i.e. the model generates the next value based on the previous one. The graphical structure of the proposed autoencoder is presented as Fig. 2.

EXPERIMENTAL EVALUATION

In this section, we evaluate the performance of the proposed methods through forecasting simulations based on real-world data obtained from the GIOS. This dataset includes historical concentration values (in $\mu g/m^3$) for various air pollutants. Specifically, our goal is to train the proposed models using historical $PM_{2.5}$ and PM_{10} concentration data from several Polish cities (referred to as the original dataset), while utilizing relevant meteorological variables. The predicted values are then compared with the recorded concentrations to assess forecasting accuracy. The two proposed models are a convolutional-recurrent autoencoder (AE 1) and a hierarchical autoencoder (AE 2), referred to in parentheses in the evaluation tables for clarity.

Meteorological data were included in the training and consist of the following variables:

- temperature (°C)
- wind speed (km/h)
- wind direction (degrees)
- atmospheric pressure (hPa)

These data were retrieved from the Meteostat platform[2], which aggregates historical weather information from numerous global sources. The meteorological records span the same period as the pollution data – from January 1, 2022, to December 31, 2022. The selected meteorological features were chosen for their completeness and relevance to PM concentration modeling. For example, temperature influences atmospheric chemical reactions; wind affects pollutant dispersion;

[2] https://meteostat.net/en/

and pressure plays a role in temperature inversions that can trap pollutants near the surface, thereby increasing concentration levels. Utilizing these variables allows the models to better reflect the physical processes governing PM behavior. As temperature, we used the daily mean.

For this study, we selected data from the year 2022 because it provided a complete and consistent set of measurements, including both PM concentrations and meteorological variables, necessary for model training and evaluation. Using a single recent year allowed us to thoroughly test and compare the performance of different models under realistic conditions, while avoiding complications from missing or inconsistent data in earlier years. Although 2022 reflects lower pollution levels due to pandemic-related reductions, the methodology is general and can be applied to previous years with higher PM concentrations to assess model robustness. Both the PM concentration data from the GIOS and the meteorological dataset from Meteostat were fully complete, with no missing values, ensuring consistent input for model training and evaluation.

The forecasting models and baseline algorithms were implemented in Python using the PyTorch framework with CUDA support. Statistical analyses were performed using MATLAB. All experiments were conducted on a Gigabyte Aero notebook equipped with an Intel Core i7-13700H CPU, 32 GB of RAM, and an NVIDIA RTX 4070 GPU with 8 GB of VRAM.

COLLECTED DATA AND AREA DESCRIPTION

The evaluation was based on historical concentrations of $PM_{2.5}$ and PM_{10} measured at the air quality monitoring stations in Poland. The locations of the monitoring stations are reported with geographic coordinates (latitude φ, longitude λ) in the WGS84 reference system. The analyzed stations are the following:

- Zielonka Bory Tucholskie (KpZielBoryTu; 53,662117° N; 17,934017° E)
- Zakopane (MpZakopaSien; 49,2936° N; 19,9601° E)
- Częstochowa – SlCzestoZana (50,801918° N; 19,106961° E – $PM_{2.5}$) / SlCzestoBacz (50,836389° N; 19,130111° E – PM_{10})
- Gliwice/Knurów – SlGliwicMewy (50,279481° N; 18,655736° E – $PM_{2.5}$) / SlKnurJedNar (50,233167° N; 18,655722° E – PM_{10})
- Szczecin (ZpSzczec1Maj; 53,712114° N; 16,692517° E)

In Częstochowa and Gliwice/Knurów, $PM_{2.5}$ and PM_{10} are monitored at different stations located in close proximity. For instance, SlCzestoZana and SlCzestoBacz are situated near each other, as are SlGliwicMewy and SlKnurJedNar. While

the short distance between these stations suggests a degree of comparability, it should be noted that even small spatial separations (e.g. 500 m) may lead to variations in PM concentrations depending on local emission sources. Figure 3 shows the geographic distribution of the monitoring stations.

The monitoring stations were selected based on the completeness and continuity of their data records for the study period, ensuring no substantial gaps in pollutant or meteorological measurements. In addition, the chosen stations represent diverse environmental contexts in Poland: for instance, Zielonka Bory Tucholskie (KpZielBoryTu) reflects rural background conditions, Zakopane (MpZakopaSien) represents a mountainous area, while Częstochowa (SlCzestoBacz / SlCzestoZana), Gliwice/Knurów (SlGliwicMewy / SlKnurJedNar), and Szczecin (ZpSzczec1Maj) capture urban or industrial settings. This diversity allows the models to be evaluated under heterogeneous air quality conditions and enhances the robustness and generalizability of the proposed forecasting method.

Fig. 3. The location of the air quality monitoring stations in Poland used in the experiment (Source: Author's own study)

TRAINING AND META PARAMETERS

Each model was trained in two configurations. The first uses full input data, including PM concentrations along with temperature, wind speed, wind direction, and air pressure. These inputs are organized into tuples consisting of one value from each variable. Each model takes five consecutive daily observations (i.e. five historical time steps for each variable) as input and generates a single forecast for the next day's PM concentration. The models are trained to minimize the error between the forecasted and observed values using only the PM concentration as the target label.

The second configuration excludes meteorological data and uses only historical PM values as input. Both autoencoders take the five most recent daily observations as input and produce a forecast for the next day. The choice of a five observations was made to effectively capture short-term temporal dependencies in the PM data while maintaining a sufficient number of training samples. Smaller windows would miss important patterns, whereas larger windows would reduce the number of independent samples and increase the risk of overfitting.

Model hyperparameters were selected experimentally. All models were trained using the Adam optimizer with a learning rate of 0.001 and mean squared error (MSE) as the loss function. An early stopping mechanism was applied to prevent overfitting, terminating training once the validation loss no longer decreased; in our experiments, this occurred at around 1000 epochs. The dataset was chronologically divided into three subsets following a 60–20–20 split: the first 60% of the observations were used for training, the next 20% for validation, and the final 20% for independent testing.

All input data (except wind direction) were normalized to the [0, 1] interval. The correctness of normalization was verified by checking the min–max ranges and inspecting histograms before and after scaling. No data distortion was observed (figures omitted for brevity). Specifically, wind direction values (in degrees) were converted into two components, $\sin(\theta)$ and $\cos(\theta)$, where θ is the wind direction angle. This representation avoids the artificial discontinuity between 0° and 359°, which would otherwise be incorrectly mapped to the extreme ends of the interval [0, 1] in standard min–max scaling.

BASE MODEL – STATE-OF-THE-ART ALGORITHM

As a baseline, we used a random forest algorithm (RF), which is an ensemble method based on decision trees. The algorithm creates multiple independent regression trees, each of which learns to predict PM concentrations based on different random subsets of data and features. The final forecast is calculated as the average of all trees' results, reducing the risk of overfitting and improving prediction stability.

The model utilized both historical PM concentration values and meteorological data: temperature, wind speed, wind direction, and atmospheric pressure. Sine and

cosine encoding was used for the wind direction variable to preserve its angular nature. The algorithm also uses PM values from previous days/hours as lag features to capture temporal dependencies.

The RF model was run with both historical PM values alone and, in an extended version, with meteorological data (temperature, wind, pressure). This allowed us to compare the impact of atmospheric conditions on the accuracy of particulate matter concentration forecasts.

EVALUATION RESULTS

To evaluate the performance of the forecasts generated by both the proposed and state-of-the-art methods, based on historical PM concentrations and meteorological data, we calculate the following error metrics: mean absolute error (MAE, eq. 3), and mean square error (MSE, eq. 4), which are defined below:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - x_t}{n}\right| \tag{3}$$

$$MSE = \frac{1}{n}\sum_{t=1}^{n}\left(x_t - y_t\right)^2 \tag{4}$$

where $\boldsymbol{x}_t$ is the actual value, $\boldsymbol{y}_t$ is a predicted value, and n is the number of input observations. The results for $PM_{2.5}$ forecasting experiments are presented in Tab. 1.

Tab. 1. Error measure values for the proposed methods from all the air quality monitoring stations for the $PM_{2.5}$ fraction. The AE 1*, AE 2*, and RF* denote the proposed and state-of-the-art methods trained without including the meteorological data (Source: Author's own study)

Air station	Error measure	AE 1	AE 2	AE 1*	AE 2*	RF	RF*
Zielonka Bory Tucholskie	MAE	0.150	0.160	0.230	0.230	0.190	0.210
	MSE	0.040	0.040	0.120	0.110	0.070	0.090
Zakopane	MAE	0.110	0.120	0.190	0.190	0.210	0.230
	MSE	0.020	0.030	0.100	0.100	0.060	0.080
Częstochowa	MAE	0.210	0.190	0.290	0.260	0.220	0.240
	MSE	0.060	0.070	0.140	0.140	0.100	0.120
Gliwice	MAE	0.250	0.210	0.330	0.280	0.240	0.260
	MSE	0.090	0.130	0.170	0.200	0.190	0.210
Szczecin	MAE	0.200	0.220	0.280	0.290	0.250	0.270
	MSE	0.070	0.100	0.150	0.170	0.130	0.150

Analysis of the quality of $PM_{2.5}$ concentration forecasts for five measurement stations shows a clear advantage of autoencoders over random forest. Analysis of the results shows that the autoencoder-based models (AE 1 and AE 2) generally outperform the RF in terms of both MAE and MSE, particularly for the stations in

Zakopane and Zielonka Bory Tucholskie. It can be seen that for most locations, the autoencoders achieve slightly lower forecast errors, suggesting their better ability to capture patterns in the temporal data. When training the models without meteorological data, the error values are higher. The data from the station in Gliwice proved to be the most difficult to model, where all methods generated significantly higher errors, and the RF model, in particular, was characterized by a large increase in MSE. The best results, on the other hand, were obtained in Zakopane, where all methods, especially the autoencoders, showed relatively low MAE and MSE values. Overall, it can be concluded that the proposed AE-based solutions are more stable and outperform the classical random forest algorithm in most cases. Table 2 presents the results of PM_{10} forecasting evaluation.

Tab. 2. Error measure values for the proposed methods from all the air quality monitoring stations for PM_{10} fraction. The AE 1*, AE 2*, and RF* denote the proposed and state-of-the-art methods trained without including the meteorological data (Source: Author's own study)

Air station	Error measure	AE 1	AE 2	AE 1*	AE 2*	RF	RF*
Zielonka Bory Tucholskie	MAE	0.200	0.150	0.280	0.220	0.180	0.200
	MSE	0.050	0.080	0.130	0.150	0.310	0.330
Zakopane	MAE	0.120	0.150	0.200	0.220	0.180	0.200
	MSE	0.020	0.040	0.100	0.110	0.070	0.090
Częstochowa	MAE	0.200	0.180	0.280	0.250	0.210	0.230
	MSE	0.060	0.060	0.140	0.130	0.090	0.110
Knurów	MAE	0.230	0.190	0.310	0.260	0.220	0.240
	MSE	0.070	0.070	0.150	0.160	0.120	0.140
Szczecin	MAE	0.240	0.220	0.320	0.290	0.250	0.270
	MSE	0.080	0.060	0.160	0.130	0.090	0.110

The results clearly show that utilizing meteorological data into the training process improves the accuracy of air quality forecasting. Models AE 1 and AE 2, which used both historical PM concentrations and meteorological variables, achieved lower MAE and MSE error values compared to the versions based solely on historical PM data. Model AE 2 performed particularly well, producing the lowest errors in most locations, especially in Knurów and Szczecin. Importantly, random forest without meteorological data performed worse than the version with additional input variables, with differences particularly visible in Zakopane and Szczecin. The most difficult conditions to predict were in Szczecin, where all models achieved higher errors, although the use of meteorological data continued to reduce MAE and MSE. The best forecasts were observed for Zakopane, where MSE values were the lowest among all stations. In summary, integrating meteorological data increases forecast accuracy and clearly supports autoencoder models, which prove to be more effective than the classic random forest.

CONCLUSION

In this study, two deep learning architectures based on autoencoders were proposed for forecasting PM concentrations: a convolutional-recurrent autoencoder and a hierarchical autoencoder. Both models were evaluated using historical data on $PM_{2.5}$ and PM_{10} concentrations from several monitoring stations. To assess the impact of additional inputs, the models were tested in two configurations: one using only historical PM data, and another using meteorological variables such as temperature, wind, and humidity. The experimental results clearly indicate that the inclusion of weather-related features significantly improves forecasting accuracy across all evaluated metrics and locations. Both proposed models achieved strong performance, with the weather-informed versions consistently outperforming their simplified counterparts. These findings confirm the importance of meteorological context in modeling air pollution dynamics and demonstrate the effectiveness of deep autoencoder-based approaches in this domain.

Future work will focus on extending the proposed models by integrating additional environmental variables such as solar radiation, as well as testing their performance in different geographical regions. Moreover, a comprehensive comparison with other state-of-the-art forecasting methods is planned to further validate the robustness and generalizability of the proposed approach.

REFERENCES

Borah, J., Nadzir, M.S.M., Cayetano, M.G., Majumdar, S., Ghayvat, H., & Srivastava, G. (2024). AiCareAir: Hybrid-ensemble Internet of Things sensing unit model for air pollutant control. *IEEE Sensors Journal*, *99*. https://doi.org/10.1109/JSEN.2024.3397735

Cichowicz, R., Wielgosiński, G., & Fetter, W. (2020). Effect of wind speed on the level of particulate matter PM_{10} concentration in atmospheric air during winter season in vicinity of large combustion plant. *Journal of Atmospheric Chemistry*, *77*, 35–48. https://doi.org/10.1007/s10874-020-09401-w

Czernecki, B., Marosz, M., & Jędruszkiewicz, J. (2021). Assessment of machine learning algorithms in short-term forecasting of PM_{10} and $PM_{2.5}$ concentrations in selected Polish agglomerations. *Aerosol and Air Quality Research*, *21*, 200586. https://doi.org/10.4209/aaqr.200586

Du, P., Wang, J., Yang, W., & Niu, T. (2022). A novel hybrid fine particulate matter ($PM_{2.5}$) forecasting and its further application system: Case studies in China. *Journal of Forecasting*, *41*, 64–85. https://doi.org/10.1002/for.2785

Gryech, I., Asaad, C., Ghogho, M., & Kobbane, A. (2024). Applications of machine learning & Internet of Things for outdoor air pollution monitoring and prediction: A systematic literature review. *Engineering Applications of Artificial Intelligence*, *137*, 109182. https://doi.org/10.1016/j.engappai.2024.109182

Harishkumar, K., Yogesh, K., Gad, I., & Doreswamy, N. (2020). Forecasting air pollution particulate matter ($PM_{2.5}$) using machine learning regression models. *Procedia Computer Science*, *171*, 20572066. https://doi.org/10.1016/j.procs.2020.04.221

Jasiński, R., Galant-Gołębiowska, M., Nowak, M., Ginter, M., Kurzawska, P., Kurtyka, K., & Maciejewska, M. (2021). Case study of pollution with particulate matter in selected locations of Polish cities. *Energies*, *14*(9), 2529. https://doi.org/10.3390/en14092529

Kouziokas, G.N. (2020). SVM kernel based on particle swarm optimized vector and Bayesian optimized SVM in atmospheric particulate matter forecasting. *Applied Soft Computing*, *93*, 106410. https://doi.org/10.1016/j.asoc.2020.106410

Kowalski, P., Sapała, K., & Warchałowski, W. (2020). PM_{10} forecasting through applying convolution neural network techniques. *International Journal of Environmental Impacts*, *3*(1), 31–43. https://doi.org/10.2495/EI-V3-N1-31-43

Kryza, M., Werner, M., & Dore, A.-J. (2019). Application of degree-day factors for residential emission estimate and air quality forecasting. *International Journal of Environment and Pollution*, *65*(4), 325–336. https://doi.org/10.1504/IJEP.2019.103748

Kujawska, J., Kulisz, M., Oleszczuk, P., & Cel, W. (2022). Machine learning methods to forecast the concentration of PM_{10} in Lublin, Poland. *Energies*, *15*(17), 6428. https://doi.org/10.3390/en15176428

Li, T., Hua, M., & Wu, X. (2020). A hybrid CNN-LSTM model for forecasting particulate matter (PM2.5). *IEEE Access*, *8*, 2693326940. https://doi.org/10.1109/ACCESS.2020.2971348

Mauricio-Alvarez, L.-E., Aceves-Fernandez, M.-A., Pedraza-Ortega, J.-C., & Ramos-Arreguin, J.-M. (2024). Evaluation of a transformer-based model for the temporal forecast of coarse particulate matter (PMCO) concentrations. *Earth Science Informatics*, *17*, 3095–3110. https://doi.org/10.1007/s12145-024-01330-6

Nidzgorska-Lencewicz, J. (2018). Application of artificial neural networks in the prediction of PM_{10} levels in the winter months: A case study in the Tricity agglomeration, Poland. *Atmosphere*, *9*(6), 203. https://doi.org/10.3390/atmos9060203

Penkała, M., Rogula-Kozłowska, W., Ogrodnik, P., Bihałowicz, J.S., & Iwanicka, N. (2023). Exploring the relationship between particulate matter emission and the construction material of road surface: Case study of highways and motorways in Poland. *Materials*, *16*, 1200. https://doi.org/10.3390/ma16031200

Polichetti, G., Cocco, S., Spinali, A., Trimarco, V., & Nunziata, A. (2009). Effects of particulate matter (PM_{10}, $PM_{2.5}$ and PM_{1}) on the cardiovascular system. *Toxicology*, *261*, 1–8. https://doi.org/10.1016/j.tox.2009.04.035

Połednik, B. (2022). Emissions of air pollution in industrial and rural region in Poland and health impacts. *Journal of Ecological Engineering*, *23*, 250258. https://doi.org/10.12911/22998993/151986

Ramentol, E., Grimm, S., Stinzendorfer, M., & Wagner, A. (2023). Short-term air pollution forecasting using embeddings in neural networks. *Atmosphere*, *14*, 298. https://doi.org/10.3390/atmos14020298

Rogula-Kozłowska, W., Klejnowski, K., Rogula-Kopiec, P., Ośródka, L., Krajny, E., Błaszczak, B., & Mathews, B. (2014). Spatial and seasonal variability of the mass concentration and chemical composition of $PM_{2.5}$ in Poland. *Air Quality, Atmosphere & Earth*, *7*, 41–58. https://doi.org/10.1007/s11869-013-0222-y

Sharma, E., Deo, R.C., Prasad, R., Parisi, A.V., & Raj, N. (2020). Deep air quality forecasts: Suspended particulate matter modeling with convolutional neural and long short-term memory networks. *IEEE Access*, *8*, 209503–209516. https://doi.org/10.1109/ACCESS.2020.3039002

Sowka, I., Chlebowska-Styś, A., Pachurka, Ł., Rogula-Kozłowska, W., & Mathews, B. (2019). Analysis of particulate matter concentration variability and origin in selected urban areas in Poland. *Sustainability*, *11*, 5735. https://doi.org/10.3390/su11205735

Swetha, G., Datla, R., Vishnu, C., & Mohan, C.K. (2024). M2-APNet: A multimodal deep learning network to predict major air pollutants from temporal satellite images. *Journal of Applied Remote Sensing*, *18*, 012005–012005. https://doi.org/10.1117/1.JRS.18.012005

Tariq, S., Loy-Benitez, J., & Yoo, C. (2023). Enhancing the sustainable management of fine particulate matter-related health risks at subway stations through sequential forecast and gated probabilistic transformer. *Building and Environment*, *244*, 110780. https://doi.org/10.1016/j.buildenv.2023.110780

Tong, W., Limperis, J., Hamza-Lup, F., Hu, Y., & Li, L. (2024). Robust transformer-based model for spatiotemporal $PM_{2.5}$ prediction in California. *Earth Science Informatics*, *17*, 315–328. https://doi.org/10.1007/s12145-023-01138-w

Tran, D., Nguyen, H., Tran, B., La Vecchia, C., Luu, H.N., & Nguyen, T. (2021). Fast and precise single-cell data analysis using a hierarchical autoencoder. *Nature Communications*, *12*, 1029. https://doi.org/10.1038/s41467-021-21312-2

Qin, S., Liu, F., Wang, J., & Sun, B. (2014). Analysis and forecasting of the particulate matter (PM) concentration levels over four major cities of China using hybrid models. *Atmospheric Environment*, *98*, 665–675. https://doi.org/10.1016/j.atmosenv.2014.09.046

Vovk, T., Kryza, M., & Werner, M. (2024). Using random forest to improve EMEP4PL model estimates of daily $PM_{2.5}$ in Poland. *Atmospheric Environment*, *332*, 120615. https://doi.org/10.1016/j.atmosenv.2024.120615

Won, W.-S., Oh, R., Lee, W., Ku, S., Su, P.-C., & Yoon, Y.-J. (2021). Hygroscopic properties of particulate matter and effects of their interactions with weather on visibility. *Scientific Reports*, *11*, 16401. https://doi.org/10.1038/s41598-021-95834-6

Zeng, Q., Wang, L., Zhu, S., Gao, Y., Qiu, X., & Chen, L. (2023). Long-term $PM_{2.5}$ concentrations forecasting using CEEMDAN and deep transformer neural network. *Atmospheric Pollution Research*, *14*(9), 101839. https://doi.org/10.1016/j.apr.2023.101839

Zheng, Z., & Zhang, Z. (2023). A temporal convolutional recurrent autoencoder based framework for compressing time series data. *Applied Soft Computing*, *147*, 110797. https://doi.org/10.1016/j.asoc.2023.110797

https://meteostat.net/en/

https://www.gov.pl/web/gios-en

PUBLICATION INFO		
SUBMITTED: 2025.07.23	ACCEPTED: 2025.10.20	PUBLISHED ONLINE: 2025.12.5