

Analysis of the Impact of Weather on PM_{2.5} Variability Using Machine Learning-Based Predictive Method

Agnieszka Głowacka¹, Aleksander Bies², Natalia Malcherczyk-Bies³, Paulina Trybek⁴

¹ Faculty of Science and Technology, University of Silesia, 75. Pułku Piechoty 1A, 41-500 Chorzów, Poland

² Institute of Computer Science, Faculty of Science and Technology, University of Silesia in Katowice, Będzińska 39, 41-200, Sosnowiec, Poland

³ Institute of Biology, Biotechnology and Environmental Protection, Faculty of Natural Sciences, University of Silesia in Katowice, Jagiellońska 28, 40-035, Katowice, Poland

⁴ Institute of Physics, University of Silesia in Katowice, 75. Pułku Piechoty 1A, 41-500 Chorzów, Poland

a.bies@gsuite.edu.pl, aleksander.bies@us.edu.pl, natalia.malcherczyk@us.edu.pl,
paulina.trybek@us.edu.pl

Abstract

Fine particulate matter smaller than 2.5 micrometers significantly impacts human health, yet its concentrations are governed by complex interactions between emissions and meteorological conditions. This study investigates the seasonal drivers of PM_{2.5} in the Silesian Voivodeship and assesses the predictive performance of machine learning models, including Light Gradient Boosting Machine, Random Forest, and Support Vector Regression. Meteorological variables, temporal factors, and gaseous pollutants were analyzed to identify key determinants of pollution episodes. During the heating season, elevated residential heating emissions combined with stagnant atmospheric conditions dominate PM_{2.5} variability, while in the non-heating season, atmospheric dynamics and seasonal timing play a stronger role. Light Gradient Boosting Machine consistently achieved the highest predictive accuracy, capturing nonlinear interactions and providing interpretable insights into feature importance. These results highlight the potential of machine learning approaches for accurate air quality modeling and inform strategies for pollution mitigation and public health protection.

Analiza vpliva vremena na variabilnost PM_{2.5} z uporabo napovedne metode, ki temelji na strojnem učenju

Povzetek

Drobni delci s premerom, manjšim od 2,5 mikrometra, pomembno vplivajo na zdravje ljudi, vendar njihove koncentracije določajo kompleksne interakcije med emisijami in

meteorološkimi razmerami. Ta študija preučuje sezonske dejavnike, ki vplivajo na koncentracije PM_{2.5} v Šlezijskem vojvodstvu, ter ocenjuje napovedno uspešnost modelov strojnega učenja, vključno z algoritmi Light Gradient Boosting Machine, Random Forest in Support Vector Regression. Za opredelitev ključnih dejavnikov epizod onesnaženja so bile analizirane meteorološke spremenljivke, časovni dejavniki in plinasta onesnaževala. V kurilni sezoni na spremenljivost PM_{2.5} pretežno vplivajo povečane emisije iz ogrevanja gospodinjstev v kombinaciji s stagnirajočimi atmosferskimi razmerami, medtem ko imajo zunaj kurilne sezone večjo vlogo atmosferska dinamika in sezonski časovni dejavniki. Model Light Gradient Boosting Machine je dosledno dosegal največjo napovedno natančnost, saj je učinkovito zajemal nelinearne interakcije ter omogočal razložljiv vpogled v pomembnost posameznih značilk. Ti rezultati poudarjajo potencial pristopov strojnega učenja za natančno modeliranje kakovosti zraka ter prispevajo k oblikovanju strategij za zmanjševanje onesnaženosti zraka in varovanje javnega zdravja.

Keywords: air quality modeling, fine particulate matter, machine learning, meteorological drivers, PM_{2.5}

1 INTRODUCTION

Air pollution remains one of the most pressing environmental and public health challenges of the 21st century. Among various pollutants, fine particulate matter with an aerodynamic diameter smaller than 2.5 micrometers (PM_{2.5}) is of particular concern due to its ability to penetrate deeply into the respiratory and cardiovascular systems, leading to numerous adverse health outcomes [1,2]. According to the World Health Organization (WHO), long-term exposure to high PM_{2.5} concentrations significantly increases morbidity and premature mortality [3]. The WHO estimates that ambient air pollution contributes to approximately 4.2 million premature deaths annually, primarily due to cardiovascular and respiratory diseases [3]. Exposure to PM_{2.5} has been consistently linked to a wide range of adverse health outcomes. Due to their small size and large surface area, PM_{2.5} particles can bypass the upper respiratory defense mechanisms, reaching the alveolar region and even entering the bloodstream. Once inhaled, they induce oxidative stress, systemic inflammation, and endothelial dysfunction, which can impair multiple organ systems [12,13]. Numerous epidemiological and toxicological studies have demonstrated that chronic exposure to elevated PM_{2.5} levels increases the risk of cardiovascular diseases, including ischemic heart disease, myocardial infarction, arrhythmia, and stroke [13,14]. For instance, a large cohort study reported that every 10 µg/m³ increase in annual mean PM_{2.5} concentration was associated with approximately 12% higher incidence of myocardial infarction and a 21% increase in ischemic heart disease mortality [14].

Respiratory effects are equally significant. Long-term exposure to PM_{2.5} has been associated with the development and progression of Chronic Obstructive Pulmonary Disease (COPD),

asthma, and lung cancer [15,16]. In children, exposure to fine particulate matter has been linked to impaired lung function development, increased respiratory infections, and higher rates of asthma exacerbations [17]. Additionally, short-term increases in $PM_{2.5}$ concentrations have been correlated with elevated hospital admissions for respiratory and cardiovascular conditions, highlighting both the acute and chronic effects of exposure to $PM_{2.5}$ [18]. Emerging evidence also indicates that $PM_{2.5}$ exposure affects other physiological systems. Studies suggest associations between fine particulate matter and cognitive decline, dementia, adverse pregnancy outcomes, and metabolic disorders such as diabetes [19,20]. These findings emphasize that $PM_{2.5}$ poses a systemic health burden, with no safe threshold for exposure identified [12]. Understanding the mechanisms and environmental determinants that control $PM_{2.5}$ levels is therefore essential not only for air quality management but also for protecting human health.

Meteorological conditions play a fundamental role in shaping air quality. Variables such as temperature, relative humidity, wind speed and direction, atmospheric pressure, or precipitation influence both the dispersion and dilution of pollutants, as well as their chemical transformations and removal processes [4,5]. This is particularly relevant in the context of secondary particulate matter formation. Precursor gases, such as sulfur dioxide (SO_2), nitrogen dioxide (NO_2), and carbon monoxide (CO), emitted from anthropogenic sources, undergo chemical reactions that are closely dependent on weather conditions. Meteorological factors, such as solar radiation, temperature, and humidity, catalyze oxidation processes and gas-to-particle conversions, leading to the transformation of these gases into secondary $PM_{2.5}$ particles. For instance, stagnant weather conditions with low wind speeds and temperature inversions favor the accumulation of pollutants near the ground surface, whereas rainfall and strong winds contribute to their removal. Despite the well-documented role of weather in air quality, the relationships between meteorological parameters and $PM_{2.5}$ concentrations are often complex, non-linear, and dependent on local geographic conditions and emission characteristics [6].

Traditional statistical methods, such as linear regression or time series models, have been widely applied to analyze air pollution dynamics, but they often fail to capture the complex and nonlinear dependencies between meteorological variables and $PM_{2.5}$ concentrations [7]. Machine learning (ML) techniques, including Random Forest, Gradient Boosting, Support Vector Regression, and Artificial Neural Networks, provide powerful alternatives capable of modeling nonlinear relationships and uncovering hidden patterns in multidimensional datasets. Recent studies have demonstrated that ML-based models often outperform conventional statistical approaches in predicting $PM_{2.5}$ levels, particularly under highly variable environmental conditions [8, 9, 10, 11].

The aim of this study is to analyze the impact of meteorological factors on $PM_{2.5}$ variability and to develop predictive models using machine learning approaches. By integrating meteorological and air quality data, this study seeks to (i) identify key meteorological drivers of $PM_{2.5}$ concentration variability, (ii) evaluate the predictive performance of several machine learning algorithms, and (iii) provide interpretable insights into the relative importance of different

meteorological parameters. The proposed data-driven framework contributes to a deeper understanding of atmospheric processes influencing air quality and supports the development of more accurate PM_{2.5} forecasting systems.

Despite the growing body of literature on PM_{2.5} prediction using machine learning methods, many studies focus primarily on predictive accuracy, often without explicitly addressing the role of meteorological drivers in shaping pollution dynamics in specific regions. In particular, relatively few studies combine predictive modeling with an in-depth analysis of weather–pollution interactions at a regional scale characterized by complex emission sources and challenging atmospheric conditions.

Our study addresses this gap by focusing on the Upper Silesian region (Poland), an area known for its high levels of industrial and residential emissions and unfavorable meteorological conditions that promote pollutant accumulation. Unlike many previous approaches, the proposed framework not only applies machine learning models for PM_{2.5} prediction but also explicitly integrates meteorological variables to investigate their influence on pollution variability. This dual perspective combining predictive performance with interpretability of weather–pollution relationships—constitutes the main contribution of this work.

2 MATERIALS AND METHODS

2.1 Data Sources and Description

This study focuses on the Silesian Voivodeship (southern Poland), one of the most industrialized and urbanized regions in the country, characterized by a high population density, intensive traffic, and numerous industrial emission sources. These factors make the region particularly prone to air quality deterioration and elevated PM_{2.5} concentrations, especially during the heating season. The Silesian Voivodeship covers an area of approximately 12,300 km² and has a population exceeding 4.4 million inhabitants. The topography and the region's location relative to surrounding mountain ranges, particularly the Carpathians to the south, may limit efficient air mass exchange, thereby promoting the accumulation of pollutants under unfavorable meteorological conditions. The regional climate is temperate continental, with cold winters, moderate precipitation, and frequent temperature inversions that exacerbate smog episodes.

The air quality data used in this study were obtained from the public monitoring system maintained by the Chief Inspectorate of Environmental Protection [21]. The dataset contains hourly PM_{2.5} measurements for the year 2023 from automatic monitoring stations located within the Silesian Voivodeship. The data include the station identifier, date and hour of measurement, and PM_{2.5} concentration expressed in micrograms per cubic meter (µg/m³).

The geographical coordinates (latitude and longitude) of each GIOŚ monitoring station were not included directly in the PM_{2.5} measurement files but were retrieved from separate GIOŚ metadata resources containing detailed information about station locations. These coordinates enabled spatial referencing and matching of air quality data with corresponding meteorological observations.

Meteorological data were obtained from the Meteostat service [22], which provides harmonized and quality-controlled weather data from a global network of meteorological stations. The following parameters were used in the analysis:

- air temperature (°C),
- relative humidity (%),
- wind speed (m/s),
- atmospheric pressure (hPa),
- precipitation (mm).

Both datasets were then merged by timestamp (date and hour) to form a unified time series linking PM_{2.5} concentrations with concurrent meteorological conditions at each location.

The final dataset has an hourly temporal resolution and covers the period from January 1 to December 31, 2023. These integrated data form the basis for subsequent modeling and analysis of the impact of meteorological factors on PM_{2.5} variability in the Silesian region.

2.2 Data Preprocessing

Before applying machine learning algorithms, several preprocessing steps were performed to ensure data quality, temporal consistency, and compatibility between datasets. The workflow included handling missing and inconsistent values, detecting outliers, transforming variables, and generating additional features relevant to PM_{2.5} prediction.

2.2.1 Data Cleaning

The analysis of data revealed a high degree of measurement completeness. For the primary target variable, PM_{2.5}, missing values accounted for only 0.79% of the dataset, while for gaseous precursors (NO₂ and CO), the missingness rate did not exceed 0.73%. Meteorological parameters, including air temperature, wind speed, and atmospheric pressure, were fully recorded with no missing entries.

The raw datasets occasionally contained incomplete or corrupted records due to instrument errors or data transmission failures. In particular, sporadic gaps were observed in air quality and meteorological variables. To address these issues, missing numerical values were filled using linear interpolation, which estimates unknown values based on their immediate neighbors in the time series.

This approach is widely applied in environmental and air quality research, as atmospheric variables and pollutant concentrations typically exhibit continuous and gradual temporal variation rather than abrupt changes [23]. Linear interpolation therefore preserves the natural dynamics of the data while avoiding artificial discontinuities. Short gaps (up to 2–3 hours) were interpolated, while longer missing sequences were excluded from the analysis to prevent bias.

The interpolation process can be expressed as:

$$x_t = x_{t_1} + \frac{(x_{t_2} - x_{t_1}) \cdot (t - t_1)}{(t_2 - t_1)}$$

where x_t is the interpolated value at time t and x_{t_1} , x_{t_2} are the nearest known observations before and after the missing timestamp. Additionally, any records without station identifiers or city names were removed, as these could not be spatially matched with meteorological data.

2.2.2 Exploratory Data Analysis

Figure 1 illustrates the average hourly PM_{2.5} concentrations during the heating and non-heating seasons. The results show a clear diurnal pattern, with substantially higher PM_{2.5} levels observed during the heating period, particularly in the late evening and early morning hours. This trend reflects increased residential heating activity and limited atmospheric dispersion due to temperature inversions and weak winds. During the non-heating season, PM_{2.5} concentrations are notably lower and more stable throughout the day. This can be attributed to the absence of heating emissions and improved atmospheric mixing associated with higher temperatures and stronger wind speeds.

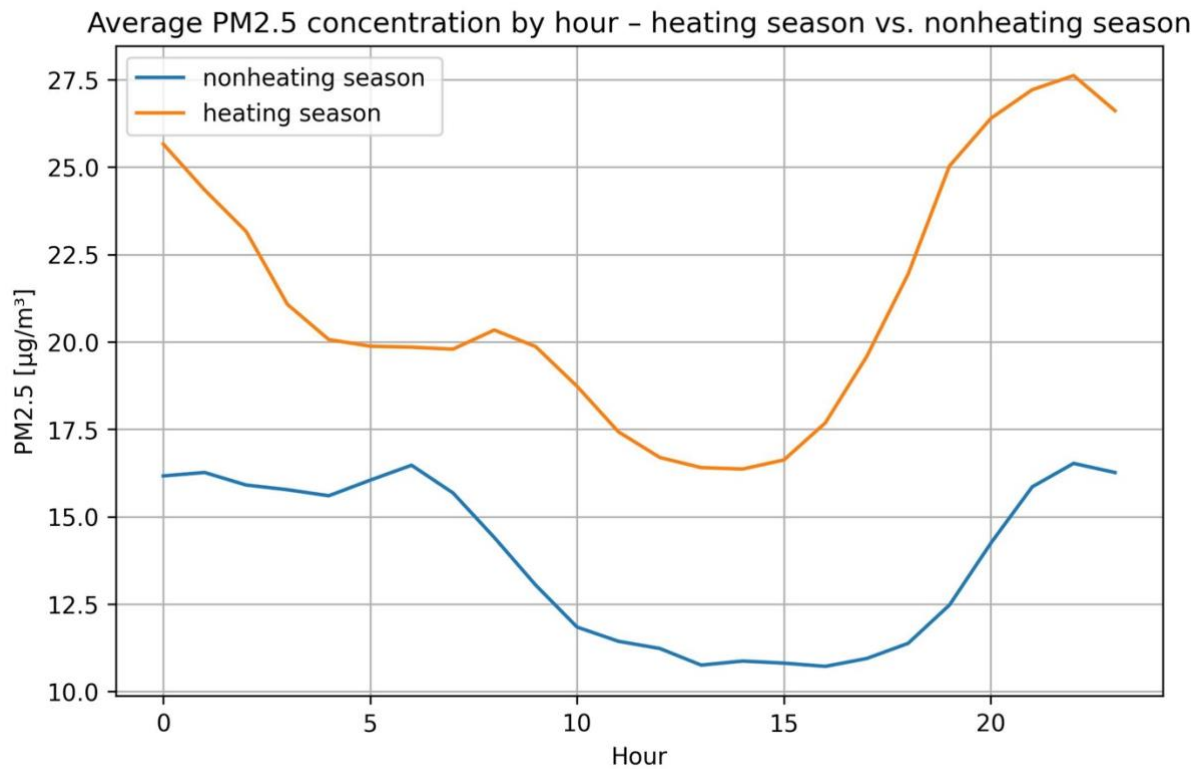


Figure 1 Average PM_{2.5} concentration by hour – heating season vs. non-heating season.

Correlation analysis was performed separately for the heating and non-heating seasons to identify the strength and direction of relationships between PM_{2.5} concentrations, other pollutants, and meteorological parameters. The results are illustrated in Figures 2 and 3.

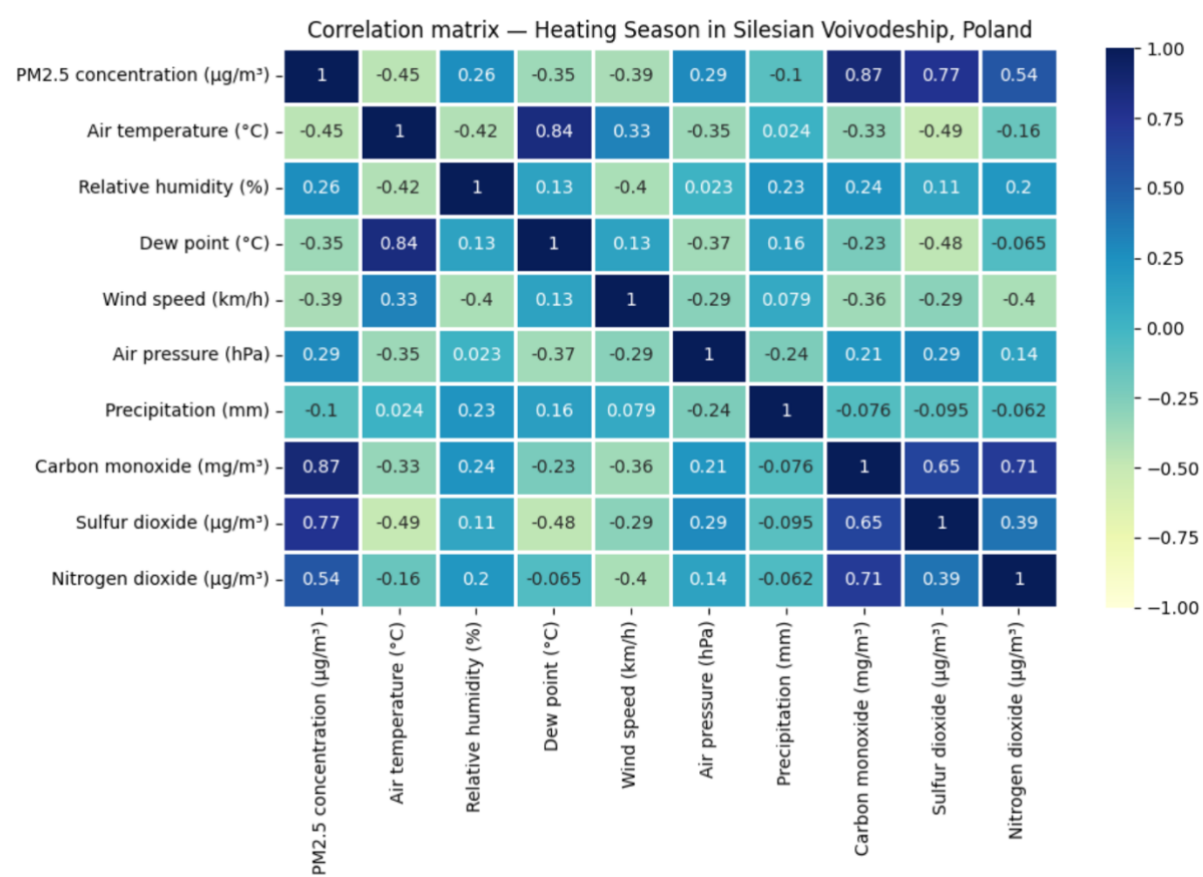


Figure 2 Correlation matrix of meteorological and pollutant variables during the heating season.

During the heating season, PM_{2.5} concentrations exhibited strong positive correlations with carbon monoxide (CO) ($r = 0.87$), nitrogen dioxide (NO₂) ($r = 0.54$), and sulfur dioxide (SO₂) ($r = 0.77$). These relationships indicate that particulate matter and gaseous pollutants share similar anthropogenic sources, primarily fossil fuel combustion in residential heating and vehicular traffic.

A moderately negative correlation was observed between PM_{2.5} and air temperature ($r = -0.45$), suggesting that lower temperatures favor pollutant accumulation, likely due to both increased emissions from heating and limited vertical mixing in cold, stable atmospheric conditions.

Wind-related parameters - wind speed ($r = -0.39$) also showed negative correlations with PM_{2.5}, confirming that stronger winds enhance pollutant dispersion. In contrast, relative humidity displayed a weak positive correlation ($r = 0.26$), implying that moist air may facilitate the secondary formation or agglomeration of fine particles.

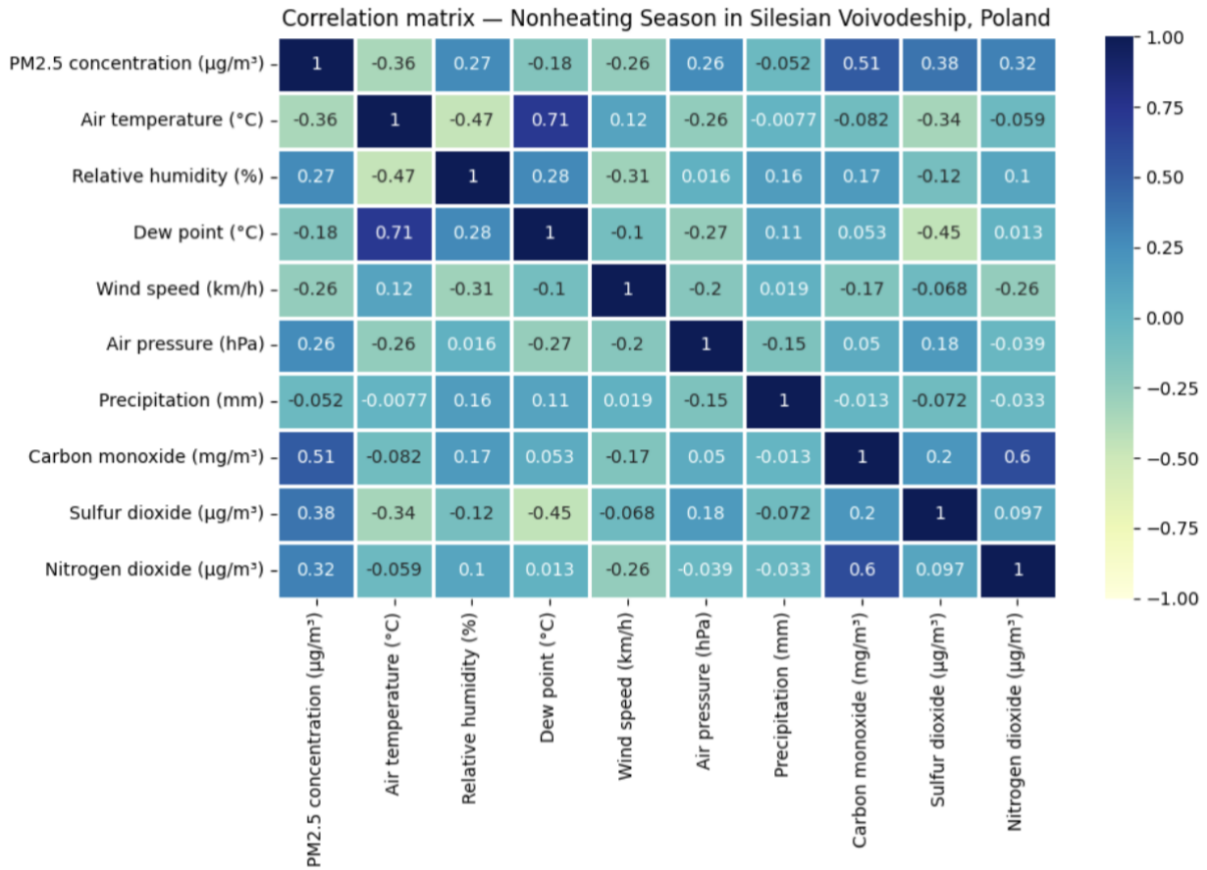


Figure 3 Correlation matrix of meteorological and pollutant variables during the non-heating season.

During the non-heating season (Figure 3), the relationships between PM_{2.5} and other variables were generally weaker. The correlations between PM_{2.5} and gaseous pollutants decreased (e.g., CO: $r = 0.51$; NO₂: $r = 0.32$; SO₂: $r = 0.38$), indicating a reduced influence of common emission sources and combustion-related processes during this period.

Meteorological influences became more pronounced: temperature ($r = -0.36$) and wind speed ($r = -0.26$) both showed negative correlations, emphasizing the role of atmospheric mixing and ventilation in pollutant removal during warmer months.

Overall, these results indicate that PM_{2.5} variability is governed by a combination of emission intensity and meteorological dispersion processes. In winter, emissions dominate due to heating activities and stagnant weather conditions, whereas in summer, atmospheric dynamics play a stronger role in determining pollutant concentrations. These insights are crucial for the subsequent feature selection and modeling stages, as they highlight which meteorological parameters most strongly influence PM_{2.5} levels in different seasons.

2.2.3 Feature engineering

To enhance the predictive performance and interpretability of the machine learning models, a structured feature engineering strategy was applied, focusing on temporal dynamics, meteorological conditions, and interactions between air quality and weather variables.

Temporal characteristics were represented using calendar-based features, including month, hour, and dayofweek, allowing the models to capture seasonal, diurnal, and weekly variability in PM_{2.5} concentrations. A binary indicator, `heating_season` (defined as November–March), was introduced to distinguish periods with elevated emissions related to residential heating from the non-heating season.

To account for short-term persistence and autocorrelation in meteorological processes, lagged variables were constructed for air temperature, relative humidity, wind speed, air pressure, and precipitation using 1-hour, 6-hour, and 12-hour delays. These lagged features reflect delayed atmospheric responses and are commonly applied in air pollution time-series modeling [8, 9]. In addition, 6-hour and 12-hour moving averages were calculated for each meteorological variable to smooth short-term fluctuations and capture broader mesoscale trends.

Several derived meteorological indicators were included to better represent atmospheric dispersion conditions. The `ventilation_index`, defined as a function of wind speed and atmospheric mixing potential, and the `stagnation_index` were used to characterize favorable and unfavorable dispersion regimes, respectively. [4, 5]. A binary precipitation indicator (`precip_binary`) was added to distinguish wet and dry conditions, while the `cold_and_calm` variable was designed to capture situations of low temperature and weak wind, which are conducive to pollutant accumulation.

To model nonlinear and synergistic effects between emissions and meteorological drivers, interaction terms were introduced between key gaseous pollutants (NO₂, SO₂, CO) and meteorological variables, including air temperature, relative humidity, wind speed, air pressure, precipitation, stagnation index, and ventilation index. These interaction features enable the models to capture compound effects that influence PM_{2.5} formation, accumulation, and removal under varying atmospheric conditions [6, 10, 11]. Overall, the selected feature set integrates temporal patterns, delayed meteorological responses, dispersion-related indicators, and pollutant–meteorology interactions, providing a comprehensive representation of the processes governing PM_{2.5} variability across different seasons and weather regimes.

A 5-fold cross-validation with random shuffling was employed to ensure a robust and unbiased evaluation of model performance. This strategy minimizes the effect of a specific data partitioning and yields a more reliable estimate of the models' generalization capability on unseen data. To ensure reproducibility, a fixed random state was maintained throughout the process. This approach does not introduce temporal data leakage because the models rely only on external factors (weather and gases) rather than past PM_{2.5} values (lags). Each observation is treated as an independent physical state of the atmosphere. This allows the models to learn

the non-linear relationships between weather stability and air pollution across different seasons without the risk of information leakage from chronological ordering.

2.3 Selected Algorithms

The selected models represent different families of machine learning approaches, including kernel-based (SVR) and ensemble tree-based methods (Random Forest, LightGBM, XGBoost), which are widely applied in air quality prediction studies. Their ability to capture nonlinear dependencies and complex interactions makes them particularly suitable for modeling PM_{2.5} variability driven by meteorological conditions. Given the inherent nonlinearity and region-specific characteristics of PM_{2.5} dynamics, the use of such advanced models enables a more realistic and robust representation of atmospheric processes.

2.3.1 Support Vector Regression (SVR)

Support Vector Regression (SVR) is a regression extension of Support Vector Machines (SVM) designed to model nonlinear relationships. The objective of SVR is to determine a function

$$f(x) = w^T \varphi(x) + b$$

that approximates the observed data while minimizing the prediction error within a predefined tolerance margin ε . Errors smaller than ε are not penalized, whereas deviations exceeding this margin are handled using slack variables. By applying a kernel function $K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$, most commonly the Radial Basis Function (RBF),

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right)$$

SVR is able to capture complex, nonlinear relationships between the input features (e.g., meteorological variables) and the target variable (PM_{2.5} concentration). The regularization parameter C controls the trade-off between model complexity and tolerance for prediction errors, γ determines the influence range of individual training samples, and ε defines the width of the insensitive loss region. While SVR generally provides stable and accurate predictions, its performance strongly depends on careful hyperparameter tuning.

The SVR model was implemented using a computational pipeline. First, a standard scaling procedure (StandardScaler) was used to normalize the features, as SVR is sensitive to the scale of the data. We chose the Radial Basis Function (RBF kernel) to better capture the non-linear patterns of air pollutants. Following the optimization, the regularization parameter was set to 10 to balance model accuracy and simplicity, while the epsilon was set to 0.1 to define a tolerance margin for prediction errors, ensuring the model's stability.

2.3.2 Random Forest (RF)

Random Forest is a robust ensemble learning method built from a collection of decision trees. Each tree is trained on a bootstrap sample of the data and a random subset of features, which imparts diversity and reduces overfitting. In regression tasks, the final prediction is the average

output of all trees. RF is well suited for air quality modeling due to its resilience to noise, capacity to capture nonlinear interactions, and built-in feature importance measures (helpful for interpretability). Studies often show random forest as a baseline strong performer in environmental prediction tasks [26].

The Random Forest model was configured with 300 trees to ensure stable predictions. To prevent overfitting, the maximum depth of each tree was limited to 15, and the minimum number of samples per leaf was set to 2. These settings allow the model to learn complex patterns without becoming too specialized to the training data.

2.3.3 LightGBM (Light Gradient Boosting Machine)

LightGBM is an efficient boosting framework developed by Microsoft, designed for large datasets and high speed. Unlike level-wise growth in traditional boosting, LightGBM uses leaf-wise tree growth, which expands the leaf with the largest information gain. This strategy often leads to better accuracy with fewer trees. It also supports categorical features directly and allows parallel training. In the environmental modeling domain, LightGBM is favored when dataset size and computational efficiency are key constraints, while still delivering competitive accuracy [27].

In this study, the LightGBM model was tuned using 730 estimators and a low learning rate (0.02) to ensure stable and gradual convergence. To reduce the risk of overfitting, model complexity was constrained by limiting the maximum tree depth to 5 and the number of leaves to 75.

2.3.4 XGBoost

XGBoost (eXtreme Gradient Boosting) is a powerful algorithm that builds decision trees sequentially rather than independently. Each new tree focuses on correcting the errors made by the previous ones. This process helps the model gradually reduce mistakes and accurately predict complex data, such as PM2.5 air quality levels.

The algorithm is highly accurate, fast, and handles missing data well. It uses regularization to prevent overfitting, which happens when a model learns the training data too specifically and fails on new information. However, XGBoost is harder to interpret than a standard tree models, and requires careful tuning of settings (hyperparameters). If these are set incorrectly, the model may perform poorly or use a large amount of RAM on big datasets.

The XGBoost model was configured with 570 trees (`n_estimators`) and a low learning rate of 0.03 to ensure precise and stable predictions. To prevent overfitting, the maximum depth of each tree was limited to 3, keeping the individual models simple. The training process used regularization techniques (L1 and L2) and only a subsample of the data (50% of rows and 90% of columns) for each tree to help the model generalize better to new data.

2.4 Results

2.4.1 Model performance during heating and non-heating seasons

The predictive performance of the evaluated machine learning models was analyzed separately for the heating and non-heating periods to account for seasonal variability in emission intensity and meteorological conditions. The results of the evaluation metrics—Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2)—are summarized in table below.

Model	Season	RMSE ($\mu\text{g}/\text{m}^3$)	MAE ($\mu\text{g}/\text{m}^3$)	R^2
SVR	Heating	6.19	3.74	0.90
SVR	Non-heating	3.71	2.66	0.73
Random Forest	Heating	5.24	3.57	0.93
Random Forest	Non-heating	3.32	2.50	0.78
LightGBM	Heating	4.58	3.10	0.95
LightGBM	Non-heating	3.00	2.26	0.84
XGBoost	Heating	5.51	3.69	0.92
XGBoost	Non-heating	3.49	2.67	0.75

Overall, all models achieved better predictive performance during the heating season, when $\text{PM}_{2.5}$ concentrations are higher and more stable. This allowed for clearer relationships between emissions, meteorological conditions, and pollutant levels to be captured. Among the tested approaches, the LightGBM model exhibited the best performance during this period, achieving the lowest RMSE ($4.58 \mu\text{g}/\text{m}^3$) and MAE ($3.10 \mu\text{g}/\text{m}^3$), as well as the highest coefficient of determination ($R^2 = 0.95$). Random Forest ($R^2 = 0.93$) and XGBoost ($R^2 = 0.92$) followed closely, though XGBoost showed slightly higher error values (RMSE = $5.51 \mu\text{g}/\text{m}^3$), while Support Vector Regression (SVR) showed comparatively lower accuracy ($R^2 = 0.90$).

During the non-heating season, prediction accuracy decreased for all models, reflecting lower $\text{PM}_{2.5}$ concentrations and increased temporal variability.

Nevertheless, LightGBM again outperformed the other methods, yielding the lowest RMSE ($3.00 \mu\text{g}/\text{m}^3$) and MAE ($2.26 \mu\text{g}/\text{m}^3$), and the highest value ($R^2 = 0.84$). In this season, Random Forest ranked second ($R^2 = 0.78$), whereas XGBoost achieved an R^2 of 0.75 of with an RMSE of $3.49 \mu\text{g}/\text{m}^3$. SVR demonstrated the lowest explanatory power during this period ($R^2 = 0.73$). While both gradient boosting models performed well, LightGBM proved to be more effective than XGBoost in capturing the specific dynamics of air quality.

2.4.2 Seasonal prediction patterns

As shown in Figure 4, the LightGBM model accurately reproduces the temporal evolution of PM_{2.5} concentrations during the heating season, including major pollution episodes. Predicted peaks closely follow the observed values, confirming the model's ability to capture high-emission conditions associated with residential heating and limited atmospheric dispersion.

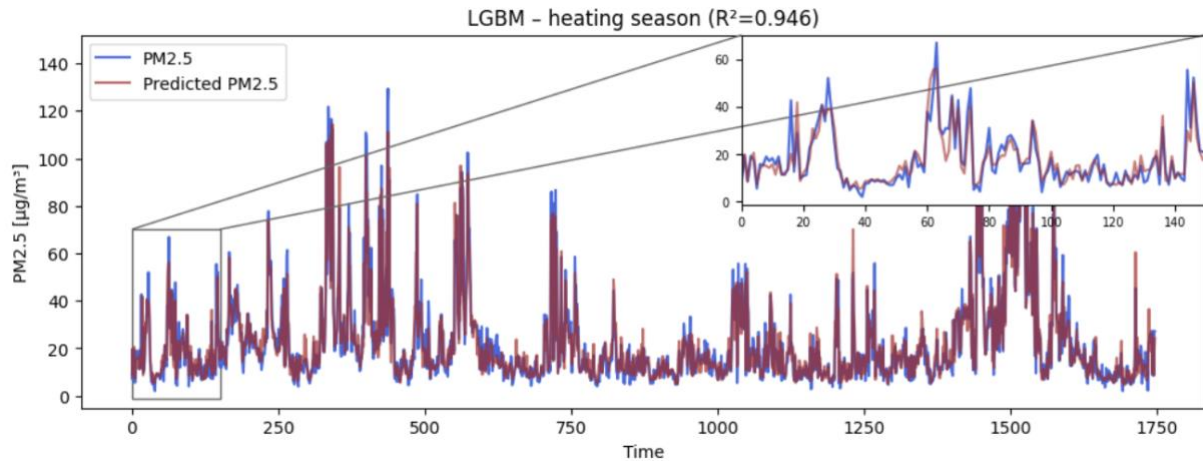


Figure 4 Observed and LightGBM-predicted PM_{2.5} concentrations during the heating season, showing strong agreement between measured and modeled values ($R^2 = 0.946$).

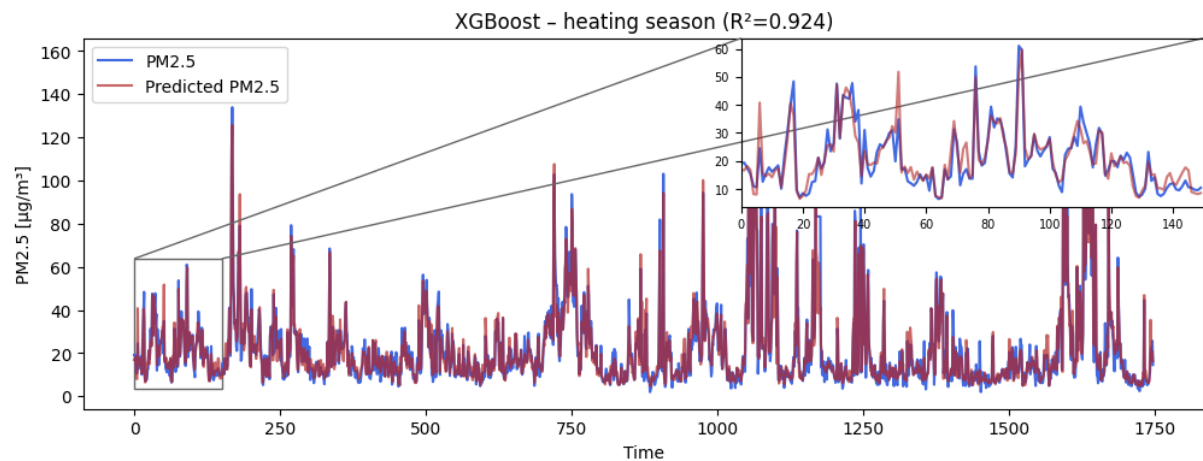


Figure 5 Observed and XGBoost-predicted PM_{2.5} concentrations during the heating season, showing strong agreement between measured and modeled values ($R^2 = 0.924$).

Looking at the visualizations (Figures 4 and 5), both models (LightGBM and XGBoost) demonstrate an impressive ability to follow the real-time fluctuations of PM_{2.5} concentrations. In both cases, the predicted values (red line) closely overlap with the actual measurements (blue line), even during sudden, sharp spikes in pollution. However, LightGBM shows a slight statistical advantage with an value R^2 of 0.946, compared R^2 to 0.924 for XGBoost.

When examining the zoomed-in sections of the graphs, it becomes clear that LightGBM is slightly more precise at capturing the exact peaks and valleys of the data. While XGBoost follows the general trend very well, it occasionally underestimates the highest spikes or smooths out some of the smaller variations more than LightGBM does. This higher sensitivity allows LightGBM to explain nearly 95% of the variance in air quality during this period, making it the most reliable tool for forecasting high-pollution events during the heating season.

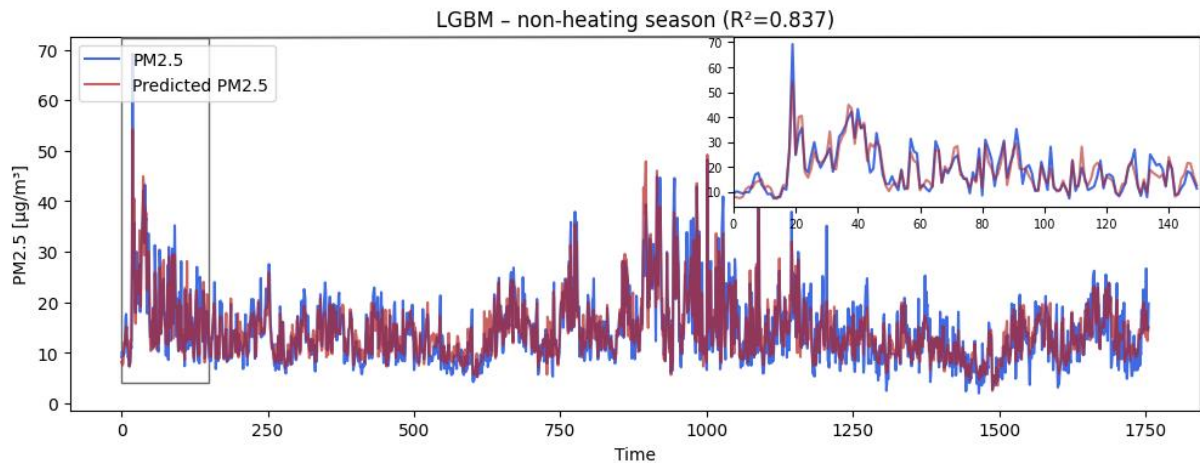


Figure 6 Observed and LightGBM-predicted $PM_{2.5}$ concentrations during the nonheating season, showing strong agreement between measured and modeled values ($R^2 = 0.837$).

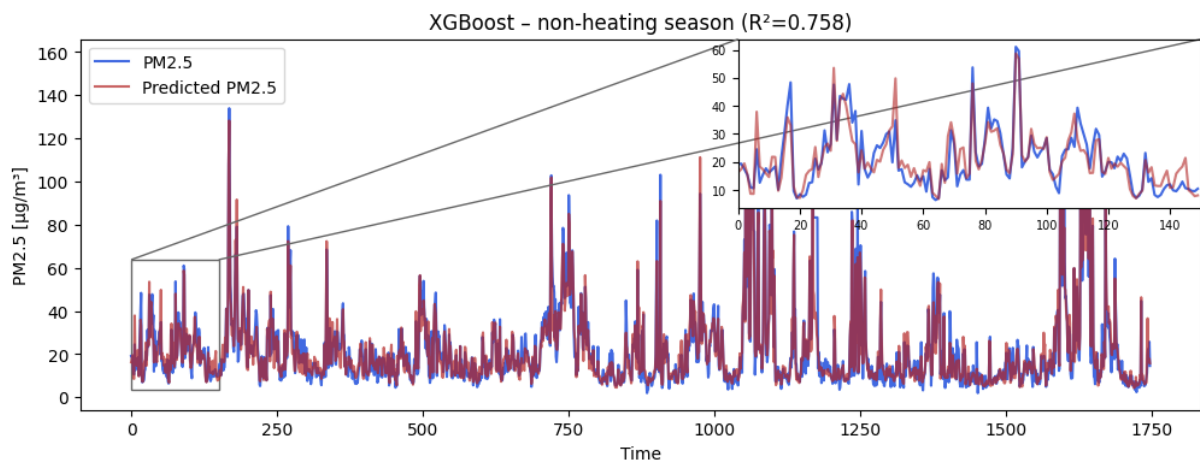


Figure 6 Observed and XGBoost-predicted $PM_{2.5}$ concentrations during the nonheating season, showing strong agreement between measured and modeled values ($R^2 = 0.758$).

When looking at the non-heating season charts (Figures 6 and 7), we can see that both models find the task more challenging due to lower $PM_{2.5}$ concentrations and more "noisy" or unstable data. However, the performance gap between the two becomes even more noticeable here.

LightGBM maintains a much stronger grip on the data with an R^2 of 0.837. The graph shows that while the concentrations are generally lower, LightGBM still manages to track the sudden fluctuations quite accurately. The zoomed-in window reveals that the red prediction line stays very close to the blue actual values, successfully capturing the small, frequent changes in air quality.

In contrast, XGBoost sees its performance drop significantly to an R^2 of 0.758. Looking at the visualization, the XGBoost model appears slightly unstable or less precise in this season. It struggles more with the increased variability, and the gap between the actual spikes and the predicted peaks is wider than in the LightGBM model. This suggests that for the more unpredictable air quality patterns found outside the heating season, LightGBM is a much more robust and reliable choice.

2.4.3 Feature importance

The Light Gradient Boosting Machine (LightGBM) model was selected for feature importance analysis due to its superior predictive performance in both the heating and non-heating seasons and its ability to provide interpretable estimates of predictor relevance.

During the heating season (Figure 8), the most influential predictors of $PM_{2.5}$ concentrations are primarily associated with combustion-related emissions and atmospheric stagnation. The highest importance is observed for interaction terms involving air pressure and carbon monoxide ($air_pressure \times CO$), indicating that elevated CO concentrations combined with stable, high-pressure conditions strongly favor $PM_{2.5}$ accumulation.

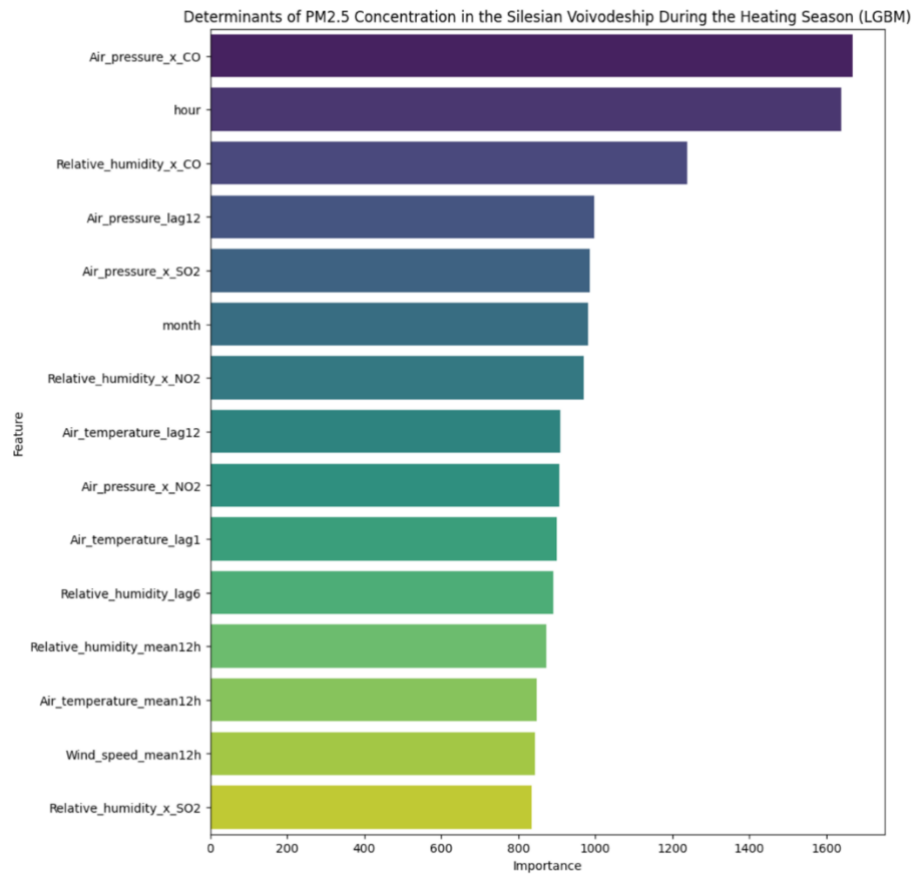


Figure 8 Relative importance of the 15 most influential predictors of PM_{2.5} concentrations during the heating season, as identified by the LightGBM model.

Temporal factors, particularly the hour of the day, also play a key role, reflecting pronounced diurnal emission patterns related to residential heating activity.

Several interaction terms linking relative humidity and air pressure with gaseous pollutants (CO, SO₂, NO₂) rank among the top predictors, highlighting the importance of meteorological modulation of pollutant chemistry and dispersion. Lagged air pressure and air temperature variables (especially 12-hour lags) further emphasize the role of persistent synoptic-scale conditions. Among purely meteorological variables, relative humidity, lagged air temperature, and mean wind speed indicate that short-term thermal conditions and reduced ventilation significantly influence PM_{2.5} levels during the heating period. The prominence of lagged predictors confirms the temporal persistence of pollution episodes under wintertime stagnation, consistent with previous studies [4, 5, 6, 10].

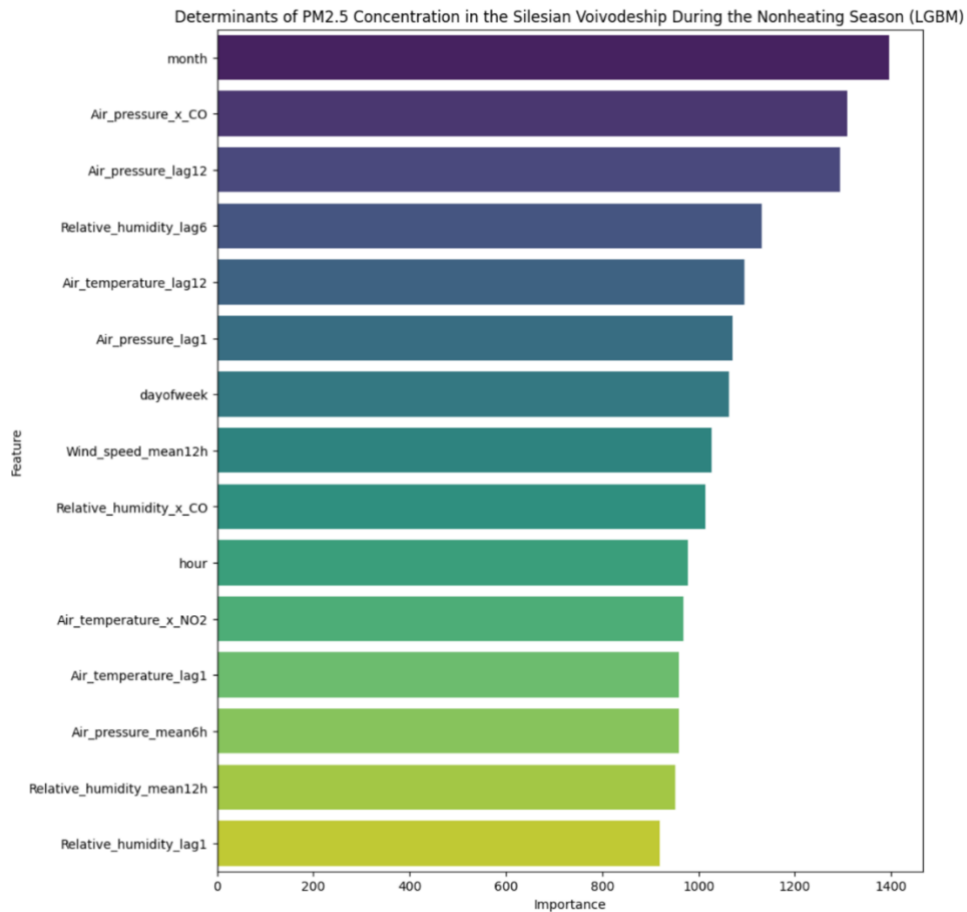


Figure 9 Relative importance of the 15 most influential predictors of PM_{2.5} concentrations during the non-heating season, as identified by the LightGBM model.

In the non-heating season (Figure 9), the structure of feature importance changes noticeably, reflecting a shift from emission-dominated to meteorology-driven controls of PM_{2.5} variability. The month emerges as the single most important predictor, underscoring the strong seasonal modulation of background PM_{2.5} levels and atmospheric processes such as photochemistry and boundary layer development. Interaction terms involving air pressure and CO, as well as lagged air pressure, remain important, indicating that synoptic-scale circulation still influences pollutant accumulation even under generally cleaner conditions.

Relative humidity (both lagged and averaged) and air temperature lagged by 12 hours rank highly, suggesting that moisture-related processes and delayed thermal effects play a greater role in secondary aerosol formation during the non-heating period. Temporal variables such as day of week and hour gain relative importance, reflecting the influence of traffic-related emissions and daily activity cycles. Wind speed (12-hour mean) also appears among the key predictors, highlighting the enhanced role of atmospheric ventilation in pollutant removal outside the heating season.

Overall, the feature importance analysis demonstrates that while PM_{2.5} concentrations during the heating season are primarily controlled by combustion-related emissions amplified by stagnant meteorological conditions, the non-heating season is characterized by a more complex interplay of seasonal timing, atmospheric dynamics, and secondary formation processes. These

results confirm that the LightGBM model effectively captures season-dependent drivers of PM_{2.5} variability and provides a physically meaningful interpretation of air quality dynamics across contrasting emission regimes.

2.5 Discussion

The obtained results confirm the significant influence of both meteorological conditions and emission-related factors on the variability of PM_{2.5} concentrations in the Silesian Voivodeship, with clear seasonal differences in the strength and structure of these relationships. During the heating season, PM_{2.5} concentrations are higher and more persistent, reflecting increased emissions from residential heating combined with unfavorable dispersion conditions, such as low air temperatures, weak winds, and frequent atmospheric stagnation. These conditions enhance pollutant accumulation and lead to stronger and more predictable relationships between PM_{2.5}, gaseous pollutants, and meteorological variables, in agreement with previous studies conducted in highly urbanized and industrial regions [4, 5, 10].

Among the evaluated machine learning models, the Light Gradient Boosting Machine (LightGBM) consistently achieved the highest predictive performance in both the heating and non-heating seasons. In the heating period, LightGBM explained nearly all variability in PM_{2.5} concentrations ($R^2 = 0.95$), outperforming Random Forest and Support Vector Regression (SVR). This demonstrates its strong capability to capture complex, nonlinear interactions between emissions and meteorological drivers under high-pollution conditions. Although model performance decreased during the non-heating season, LightGBM remained the most accurate approach ($R^2 = 0.8$), indicating that PM_{2.5} variability, while reduced, is still governed by systematic meteorological and temporal processes rather than random fluctuations.

Feature importance analysis further supports these findings by revealing distinct seasonal drivers of PM_{2.5} concentrations. During the heating season, combustion-related pollutants (CO, SO₂, and NO₂) and their interactions with atmospheric pressure and humidity dominate the model structure, highlighting the central role of residential heating emissions and atmospheric stagnation in PM_{2.5} formation and accumulation. The prominence of lagged meteorological variables (6–12 h) confirms the temporal persistence of pollution episodes and the cumulative nature of smog events under limited ventilation conditions.

In contrast, during the non-heating season, feature importance shifts toward seasonal timing (month), meteorological conditions, and ventilation-related parameters, such as wind speed and air pressure. This indicates that under lower emission intensity, PM_{2.5} concentrations are more strongly influenced by atmospheric dynamics, moisture-related processes, and secondary aerosol formation rather than direct primary emissions. The increased importance of temporal variables (hour and day of week) suggests a relatively greater influence of traffic-related and diurnal activity patterns during this period.

Overall, the comparison of model performance confirms the suitability of machine learning approaches for air quality modeling, particularly for capturing nonlinear dependencies and interactions between multiple environmental drivers. Gradient boosting-based models,

especially LightGBM, provide both high predictive accuracy and physically meaningful interpretability across contrasting seasonal regimes. These findings are consistent with previous studies highlighting the advantages of ensemble and boosting techniques over traditional linear models in PM_{2.5} prediction [7, 8, 10].

An important aspect to consider is the scope of the dataset and its implications for the generalizability of the results. The analysis is based on data from the Silesian Voivodeship, a region characterized by a unique combination of high industrial activity, dense urbanization, and significant residential heating emissions. Consequently, the identified relationships—particularly the strong interactions between meteorological conditions and emission-related variables—may be influenced by region-specific emission patterns and atmospheric dynamics. While the proposed modeling framework is transferable to other locations, the predictive performance and the relative importance of features may vary depending on local environmental conditions, emission structures, and climate characteristics. Therefore, caution should be exercised when generalizing these findings to regions with different pollution sources or meteorological regimes.

A limitation of the present study is the lack of explicit information on local traffic intensity, industrial activity, and spatial emission heterogeneity, which may further influence PM_{2.5} concentrations, particularly during the non-heating season. Future research should incorporate additional data sources, such as satellite-derived aerosol optical depth (AOD), traffic flow indicators, and terrain-related parameters, to improve the spatial representation of emissions and enhance the robustness of predictive models.

2.6 Conclusion

Meteorological conditions play a crucial role in shaping the variability of PM_{2.5} concentrations in the Silesian Voivodeship, with these relationships exhibiting pronounced seasonal dependence. The influence of both meteorological factors and emission-related pollutants is significantly stronger during the heating season, when increased residential heating emissions coincide with unfavorable dispersion conditions, such as low air temperatures, weak winds, and frequent atmospheric stagnation.

Among the applied machine learning approaches, the Light Gradient Boosting Machine (LightGBM) consistently achieved the highest predictive performance in both heating and non-heating seasons, most accurately reproducing observed PM_{2.5} concentrations and temporal pollution patterns. The superior performance of LightGBM highlights its ability to capture complex, nonlinear interactions between meteorological variables, temporal factors, and gaseous pollutants under varying emission regimes.

Feature importance analysis revealed that combustion-related pollutants (CO, SO₂, and NO₂), together with meteorological parameters associated with limited atmospheric ventilation—such as air pressure, wind speed, relative humidity, and lagged air temperature—are the dominant drivers of PM_{2.5} concentrations during the heating season. In contrast, during the non-heating

period, PM_{2.5} variability is more strongly influenced by seasonal timing and atmospheric dynamics, including ventilation efficiency and moisture-related processes.

Overall, the results demonstrate that machine learning models, particularly gradient boosting-based methods, provide a robust and interpretable framework for air quality modeling. The insights gained from this study can support the development of early warning systems for smog episodes and inform environmental management and public health policies aimed at mitigating fine particulate matter pollution.

Acknowledgement

This work was supported by the AI2SEP project (No. 2023-1-PL01-KA220-HED 000166765). We gratefully acknowledge the project's support and funding, which enabled this research

References

- [1] Pope, C. A., & Dockery, D. W. (2006). Health effects of fine particulate air pollution: Lines that connect. *Journal of the Air & Waste Management Association*, 56(6), 709–742.
- [2] Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D., Pozzer, A. (2015). The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature*, 525(7569), 367–371.
- [3] World Health Organization (2023). Ambient (outdoor) air pollution. WHO Fact Sheet.
- [4] Tai, A. P. K., Mickley, L. J., & Jacob, D. J. (2010). Correlations between fine particulate matter (PM_{2.5}) and meteorological variables in the United States: Implications for long-term PM_{2.5} prediction. *Atmospheric Environment*, 44(32), 3976–3984.
- [5] Zhang, Y., et al. (2015). Impact of meteorological conditions on PM_{2.5} concentrations across China: A review of methodology and mechanism. *Environment International*, 86, 150–162.
- [6] Papanastasiou, D. K., Melas, D., Kioutsioukis, I. (2017). Meteorological and photochemical controls on PM_{2.5} and O₃ over Europe: Analysis of 2001–2014 observations. *Atmospheric Chemistry and Physics*, 17(2), 923–940.
- [7] Guttikunda, S. K., Kopakka, R. V. (2014). Source emissions and health impacts of urban air pollution in Hyderabad, India. *Air Quality, Atmosphere & Health*, 7(2), 195–207.
- [8] Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C. (2020). Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. *Environmental Pollution*, 263, 114625.
- [9] Chen, L., Guo, B., Huang, J., Xu, J., & Su, Q. (2019). Machine learning methods for air quality forecasting: A review. *Applied Sciences*, 9(16), 3280.
- [10] Yu, R., Yang, Y., Yang, L., Han, G., Chen, Y., Zhao, J. (2021). Comparative analysis of machine learning models for predicting PM_{2.5} concentrations. *Atmosphere*, 12(1), 45.

- [11] Grange, S. K., Carslaw, D. C. (2019). Using meteorological normalisation to detect interventions in air quality time series. *Science of the Total Environment*, 653, 578–588.
- [12] Brook, R. D., Rajagopalan, S., Pope, C. A. III, et al. (2010). Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the American Heart Association. *Circulation*, 121(21), 2331–2378.
- [13] Di, Q., et al. (2017). Air pollution and mortality in the Medicare population. *New England Journal of Medicine*, 376(26), 2513–2522.
- [14] Bai, L., et al. (2023). Long-term exposure to fine particulate matter and risk of myocardial infarction: A nationwide cohort study. *JAMA Network Open*, 6(2), e2254013.
- [15] Cohen, A. J., et al. (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: An analysis of data from the Global Burden of Disease Study 2015. *The Lancet*, 389(10082), 1907–1918.
- [16] Turner, M. C., et al. (2011). Long-term ambient fine particulate matter exposure and mortality in the American Cancer Society cohort. *Environmental Health Perspectives*, 119(7), 1130–1135.
- [17] Guarnieri, M., & Balmes, J. R. (2014). Outdoor air pollution and asthma. *The Lancet*, 383(9928), 1581–1592.
- [18] Dominici, F., Peng, R. D., Bell, M. L., et al. (2006). Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *JAMA*, 295(10), 1127–1134.
- [19] Chen, H., et al. (2022). Air pollution and risk of dementia: A systematic review and meta-analysis. *Environmental Research*, 203, 111821.
- [20] Zhang, J., et al. (2021). Ambient PM_{2.5} exposure and adverse pregnancy outcomes: A meta-analysis. *Science of the Total Environment*, 755, 142556.
- [21] Chief Inspectorate of Environmental Protection (GIOŚ). (2023). Air Quality Monitoring System in Poland.
- [22] Meteostat. (2023). Historical Weather Data API.
- [23] Spatio-temporal analysis of air quality and its interpolation methods (Li, Y.; inny – 2020)
- [24] Fan, G. F., Liang, M. Q., Li, J. R., & Ma, W. L. (2017). Applications of SVR-PSO model and multivariate linear regression model in PM_{2.5} concentration forecasting. *International Journal of Applied Evolutionary Computation (IJAECE)*, 8(4), 53-69.
- [25] Ghafariasl, P., Zeinalnezhad, M., & Ahmadishokoo, A. (2024). Optimizing pm_{2.5} forecasting accuracy with hybrid meta-heuristic and machine learning models. *arXiv preprint arXiv:2407.01647*.
- [26] Özüpak, Y., Alpsalaz, F., & Aslan, E. (2025). Air Quality Forecasting Using Machine Learning: Comparative Analysis and Ensemble Strategies for Enhanced Prediction. *Water, Air, & Soil Pollution*, 236(7), 464.
- [27] Tırınk, S. (2025). Machine learning-based forecasting of air quality index under long-term environmental patterns: A comparative approach with XGBoost, LightGBM, and SVM. *PloS one*, 20(10), e0334252.

Agnieszka Głowacka, Faculty of Science and Technology, University of Silesia

Agnieszka Głowacka is a second-year Master's student of Micro and Nanotechnology at the University of Silesia in Katowice. Her interests focus on the application of modern technologies in the analysis and processing of scientific data. She is particularly interested in interdisciplinary approaches combining elements of physics, chemistry, and informatics, as well as practical applications of advanced technologies in science and industry.

Aleksander Bies, Institute of Computer Science, Faculty of Science and Technology, University of Silesia

Aleksander Bies is a third-year PhD student in Computer Science at the University of Silesia in Katowice. His research interests focus on machine learning and the application of artificial intelligence methods to time-series analysis. He is particularly interested in developing predictive models, identifying temporal patterns, and applying advanced computational techniques to complex sequential data.

Natalia Malcherczyk-Bies, Institute of Biology, Biotechnology and Environmental Protection, University of Silesia

Natalia Malcherczyk-Bies is a PhD student at the Institute of Biology, University of Silesia in Katowice. Her research interests focus on plant biology, the effects of environmental factors on plant development, and the use of plants from novel ecosystems as resources for creating green areas to support urban adaptation to climate change.

Paulina Trybek, Institute of Physics, University of Silesia

Paulina Trybek is an assistant professor at the University of Silesia in Katowice and works at the Institute of Physics, where she specializes in complex systems and biomedical time-series analysis. She is actively involved in numerous student projects, supporting the development of data analysis competencies. She is also the coordinator of the project "Developing Talents in Artificial Intelligence to Solve Disruptive Environmental Problems".