ARTIFICIAL INTELLIGENCE RESEARCH and APPLICATIONS
C. Spataro and F. Viola, Vol.1, No.2, June, 2025 |Research Paper|

Artificial Intelligence for Evaluation of PM2.5 and PM10 Levels in Urban Streets

Ciro Spataro*, Fabio Viola*‡

* Department of Engineering of University of Palermo, Italy, viale delle Scienze, Parco d'Orleans, Palermo, 90100, Italia

(ciro.spataro@unipa.it, fabio.viola@unipa.it)

‡ Corresponding Author; Fabio Viola, Department of Engineering of University of Palermo, Italy, viale delle Scienze, Parco d'Orleans, Palermo, 90100, Italia, fabio.viola@unipa.it

Received: 06.05.2025 Accepted:16.06.2025

Abstract- Atmospheric pollution, especially due to fine particles (PM2.5) and larger particles (PM10), constitutes a major risk to public health and the long-term viability of urban areas. This research focuses on measuring the presence of PM2.5 and PM10 across selected city streets, investigating the variables that affect their concentration, and identifying possible solutions to limit their impact. The study utilizes continuous air quality monitoring, data interpretation through statistical tools, and a review of urban planning measures to offer a thorough understanding of particulate contamination in metropolitan settings. Findings reveal significant patterns in how these pollutants spread, their links to vehicular traffic and weather conditions, and suggest practical approaches to lower human exposure. The contribution of artificial intelligence starts in collecting raw data, converting in Matlab's matrices and help the operator identify correlations between the taken measurements and the the influence of environmental parameters, which allow particulate matter to vary daily even without following consolidated trends in the same time period.

Keywords: Pollutant, particle matter PM10 and PM2.5, data analysis, monitoring techniques, artificial intelligence.

1. Introduction

Electric vehicles (EVs) are increasingly recognized as a key component in the transition toward more sustainable and intelligent urban mobility systems. While EVs eliminate tailpipe emissions and thus directly reduce urban air pollution, they are not entirely exempt from contributing to atmospheric particulate matter (PM). Non-exhaust sources—such as brake wear, tire abrasion, and road surface degradation—remain significant contributors to urban PM concentrations. In fact, due to the additional mass of battery systems, EVs may in some cases produce higher levels of non-exhaust PM compared to lighter internal combustion engine vehicles.

Nonetheless, the integration of regenerative braking systems, which are standard in most EVs, substantially reduces brake wear and consequently mitigates this emission source. Beyond the vehicle itself, the broader deployment of EVs within smart grid ecosystems powered by renewable energy sources enables a systemic reduction in air pollution. When EVs are charged during off-peak hours or managed through vehicle-to-grid (V2G) protocols, they support grid stability while minimizing dependence on fossil-fuel-based power generation.

Crucially, the convergence of EV technology with Artificial Intelligence (AI) further amplifies its environmental benefits. AI algorithms can be employed to optimize charging schedules, predict localized pollution hotspots, and manage dynamic energy distribution across the grid. In urban areas, AI-driven data analytics can correlate traffic patterns, meteorological data, and emission trends to enhance real-time decision-making for both mobility and environmental health.

Thus, the synergy between EVs, smart grids, and AI represents a transformative opportunity to improve air quality, reduce urban emissions, and support the development of resilient, data-informed smart cities.

Air pollution remains a critical issue in densely populated urban areas due to its profound implications for human health, infrastructure integrity, and climate systems. Accurate air quality assessment requires the continuous monitoring of several key atmospheric pollutants, each associated with distinct physiological and environmental effects:

- Ozone (O_3): A secondary pollutant formed through photochemical reactions, ozone is known to irritate the eyes and respiratory tract, with elevated levels aggravating pre-existing respiratory conditions.
- Nitrogen Dioxide (NO_2): Predominantly emitted from diesel-powered vehicles, NO_2 acts as a potent respiratory irritant and is strongly associated with the development and exacerbation of pulmonary diseases.
- Sulfur Dioxide (SO_2): Resulting mainly from the combustion of sulfur-containing fuels, prolonged exposure to SO_2 is linked to inflammation of the respiratory tract, pharyngitis, and sensory dysfunctions.
- Carbon Monoxide (CO): A colorless, odorless gas, CO interferes with the oxygen-carrying capacity of blood by forming carboxyhemoglobin, posing serious risks particularly in high-traffic and industrial zones.
- Particulate Matter (PM10): Coarse particles generated by mechanical processes and combustion sources, PM10 primarily affects the upper respiratory tract, leading to irritation and inflammatory responses.
- Particulate Matter (PM2.5): Fine particles with aerodynamic diameters ≤2.5 µm that can infiltrate deep into the pulmonary alveoli. Chronic exposure is associated with long-term cardiovascular and respiratory diseases, including asthma and ischemic heart conditions.

Emerging artificial intelligence (AI) approaches, particularly in the domains of sensor data fusion and predictive modeling, are increasingly employed to analyze pollutant patterns and forecast air quality dynamics in real time. These tools are essential for developing adaptive mitigation strategies and public health interventions in smart urban ecosystems [1-4] by using advanced predictive artificial intelligence techniques [5-6].

Among the various air pollutants, particulate matter—specifically PM2.5 (aerodynamic diameter <2.5 µm) and PM10 (<10 µm)—is particularly concerning due to its ability to penetrate deep into the respiratory tract, contributing to the onset and progression of cardiovascular and pulmonary disorders. Urban roadways serve as critical hotspots for particulate matter accumulation, primarily due to traffic emissions, industrial operations, and the resuspension of road dust. A thorough understanding of the spatial distribution and emission sources of PM2.5 and PM10 is crucial for the formulation of effective pollution control strategies. This study examines particulate concentrations across a range of urban environments, identifies key variables influencing their levels, and outlines targeted interventions aimed at enhancing air quality. In addition, the research evaluates and compares a selection of low-cost, fine particle sensors suitable for widespread deployment in the urban landscape. Particular attention is given to their integration into public infrastructure, including EV charging stations, which are increasingly prevalent in modern cities. These sensor networks lay the groundwork for real-time environmental monitoring and can support data-driven approaches to pollution management and urban planning.

As shown in Fig. 1, the contribution of artificial intelligence starts in collecting raw data and help the operator identify correlations between the taken measurements and the environmental parameters, which influence the particulate matter measurements by making them vary depending on seasonal factors (presence of pollen or wind action), acute factors that cannot be detected periodically (presence of road maintenance), ever-present factors (proximity of industries), factors due to negligence (absence of road cleaning and wind that pushes the particulate matter into the air).

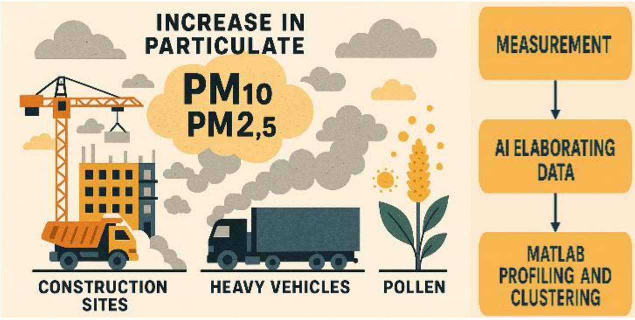

Fig. 1. AI can help operator to identify the increase in particulate.

The structure of article is the following:

Section 2 Methodology, this section describes the measurement methodology, including the instruments used for data acquisition. A comparison between the results provided by different measurement tools is also presented to evaluate consistency and reliability.

Section 3 Test Results, this part focuses on data analysis, highlighting the differences in influence between environmental parameters (such as temperature, humidity, and wind) and human-related factors (such as behavioural patterns, social calendar, and school/work schedules).

Section 4 predictive analysis for environmental correlations, this section explores the correlations between various environmental parameters, identifying potential interdependencies and their combined effects on the observed phenomena.

Section 5 discuss on predictive reliability; it proposes preliminary predictive models based on known input concentrations of a specific particulate matter type. The models attempt to estimate the resulting levels of a different particulate in output conditions.

Section 6 faces the role of Artificial Intelligence in advanced predictive approach without direct PM input. This final section attempts to define more generalized predictive models, treating the particulate matter levels as unknown variables. The models are trained on historical data and use environmental conditions as input features. Finally arrives the conclusions.

2. Methodology

The concentration of particulate matter (PM) in urban environments is governed by a combination of interrelated factors, including:

- Vehicular Emissions: Diesel-fueled vehicles and densely congested traffic corridors emerged as the dominant sources of elevated PM levels.
- Meteorological Conditions: Reduced wind speeds were associated with pollutant accumulation, whereas precipitation events contributed to short-term decreases in airborne particulate concentrations.
- Urban Morphology: Streets characterized by narrow layouts and high-rise buildings exhibited limited air dispersion, resulting in localized pollution build-up.

This research focuses on a set of representative urban street types, classified into high-traffic arterial roads, residential zones, and commercial districts. The selection criteria included population density, vehicular flow intensity, and proximity to industrial activity.

PM data were collected through real-time monitoring using a combination of portable and fixed instruments. The dataset was subsequently processed using statistical analysis software to identify patterns and correlations. The integration of sensor data enables a more granular understanding of spatiotemporal PM variability, providing a foundation for predictive modeling and informed decision-making in urban air quality management.

2.1. Instrumentation and Data Collection

Air quality monitoring was carried out through a hybrid approach combining fixed monitoring stations and mobile sensor platforms. The primary instruments employed in this study included:

- Laser-based particulate sensors, enabling high-resolution, real-time measurement of PM concentrations (PM1.0, PM2.5, PM10).
- Meteorological monitoring units, used to record ambient temperature, relative humidity, and wind speed, which are key variables influencing PM dispersion.
- Automated traffic counters, deployed to quantify vehicular flow and assess its correlation with fluctuations in particulate levels.

Data acquisition was performed continuously across multiple months, encompassing both peak traffic periods and a range of meteorological scenarios to ensure representative sampling.

To ensure accurate and scalable PM monitoring, two commercially available fine particulate sensors were selected and comparatively assessed:

- SDS011 sensor with ESP32 microcontroller: Utilizes laser scattering for PM2.5 and PM10 detection, compatible with Arduino-based platforms. It offers a cost-effective solution (~€100) for distributed air quality sensing applications [7].
- Davis Instruments AirLink (PMSA003): A higher-end, standalone sensor capable of measuring PM1.0, PM2.5, and PM10 with professional-grade accuracy. It includes integrated data logging and cloud connectivity, priced at approximately €300 [8].

The combination of low-cost and professional sensors facilitates a scalable and flexible monitoring network, suitable for integration into smart city infrastructures, including electric vehicle charging stations and urban mobility hubs.

Both sensor models, shown in Fig. 2, were deployed across selected urban sites to perform parallel measurements and evaluate their performance under real-world conditions. The data acquisition protocol was designed to comply with the Italian Legislative Decree of August 13, 2010, No. 155, which implements the European Union Directive 2008/50/EC on ambient air quality and cleaner air for Europe. This directive sets threshold values for particulate matter, specifically PM10 and PM2.5.

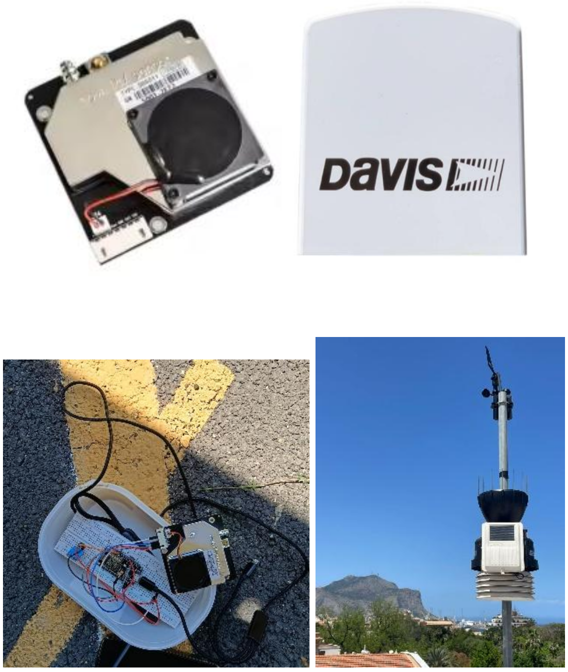

Fig. 2. Used sensors: homemade SDS011 sensor with ESP32 microcontroller, placed in different roads and professional Davis system.

In addition, the analysis was benchmarked against the updated air quality guidelines issued by the World Health Organization (WHO) [3] in 2021, which recommend the following reference limits:

- PM10: 45 µg/m³ (daily average), 15 µg/m³ (annual mean)

- PM2.5: 15 µg/m³ (daily average), 5 µg/m³ (annual mean)

The comparative analysis of sensor data is illustrated in Fig. 3, which presents the PM concentration trends recorded by the SDS011 sensor (in red) and the PMSA003 sensor (in blue). This comparison allows for the evaluation of sensor sensitivity, stability, and agreement with regulatory standards across various environmental conditions.

2.2. Data Analysis

The collected data were processed using statistical analysis software to identify temporal trends, assess correlations, and compare measured values with regulatory thresholds. The analytical framework included the following components:

- Time-series analysis, to investigate daily and weekly variations in PM concentrations;
- Correlation analysis, examining the relationship between PM levels and influencing factors such as traffic density, meteorological conditions, and urban morphology;
- Compliance assessment, comparing observed values with the air quality standards defined by the World Health Organization (WHO) and national environmental regulations [9].

A first simple use of AI concerned the possibility of identifying with a few commands, in order to have more time to investigate the behaviours, if the studied day fell on a holiday, which would have seen a human factor (less or more use of roads) prevail over an environmental factor (temperature, pressure etc.). Advanced methodologies can be found in [10]

Figure 3 illustrates a representative example of the data collected during the first days of June 2024. The graph provides a side-by-side comparison of PM2.5 and PM10 concentration levels as measured by the SDS011 sensor (red curve) and the PMSA003 sensor (blue curve). Matlab has been used to plot the behaviors, extract only few data from the bulk comma separated virgulas block (CSV) into a format ready for the Matlab elaboration.

The indicate a high degree of agreement between the two sensors with respect to PM10 measurements, reflecting a shared capacity to monitor coarse particulate matter under varying atmospheric conditions. In contrast, PM2.5 readings revealed a noticeable divergence: the SDS011 consistently reported higher concentrations than the PMSA003. This discrepancy may stem from differences in sensor sensitivity, calibration methodologies, or hardware design.

The systematic overestimation of PM2.5 by the SDS011 occasionally led to readings exceeding the WHO daily exposure limits, whereas the PMSA003 reported values below those thresholds. These findings highlight the importance of rigorous calibration and validation procedures when deploying low-cost sensors for urban air quality monitoring. Ensuring data accuracy and comparability is essential to support effective policy-making and environmental health interventions.

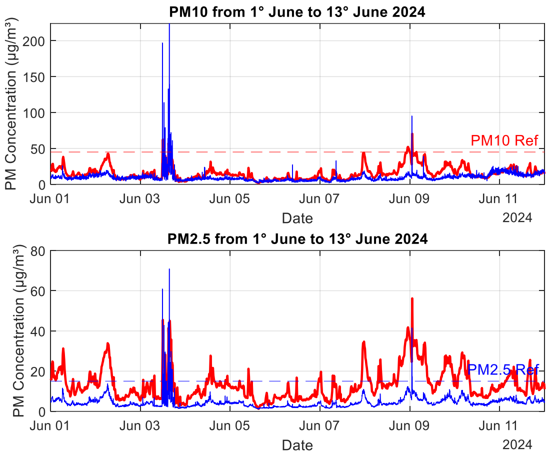

Fig. 3. Comparative analysis of particulate matter (PM2.5 and PM10) concentrations measured by the SDS011 sensor (red curve) and the PMSA003 sensor (blue curve) during the initial days of June 2024.

3. Test Results

The analysis is based on data collected between February 2024 and March 2025 in various urban zones within the municipality of Palermo, Italy.

3.1. Distribution of PM2.5 and PM10 Levels

The findings revealed substantial spatial variability in particulate matter concentrations across different types of urban streets. High-traffic corridors consistently exhibited PM10 levels surpassing recommended thresholds, while residential areas recorded comparatively lower concentrations. Nevertheless, PM2.5 levels in residential zones remained notable, likely due to diffuse background pollution and limited ventilation.

A first attempt at investigation, having months of measured values available, was to analyse the behavior of particulate matter over twenty-four hours on the first day of some months (from February to June).

Figure 4 presents a comparative overview of the PM concentration trends recorded on the first day of selected months. A strong heterogeneity is observed.

From an initial investigation, a similarity is found in February and March, the particulate matter variation curve has a first peak, between 8 am and 11 am. On other days this does not happen. We try to find a justification in environmental factors and we note that in February and March the environmental factors of temperature and humidity are quite similar at that time (Fig. 5), but nothing similar can be said for pressure and wind (Fig. 6).

Figure 4 shows that the patterns of exceedances are never the same, the hours in which the concentration approaches or exceeds the WHO limits change from day to day, and the AI was asked what it would suggest. The AI suggested comparing the first day of the month with its positioning in the working week, as well as whether it fell on holidays or close to them.

Figure 5 shows temperature and relative humidity trends recorded on the first day of each selected month between February and June 2024. The figure highlights the meteorological variability across months, which may influence the dispersion and accumulation of airborne particulate matter

Figure 6 shows atmospheric pressure and wind speed recorded on the first day of each selected month between February and June 2024, it illustrates how meteorological conditions, particularly wind intensity, may affect the dispersion of particulate matter in urban environments.

In such a way the heterogeneity can be attributed to the variability in the day of the week on which each month begins—whether it is a weekday or a public holiday significantly influences urban activity patterns and, consequently, pollution levels.

For example:

- In February, the 1st fell on a Thursday, a typical weekday. PM2.5 levels exceeded recommended limits during morning hours (6:00–11:00) and evening hours (19:00–23:00), corresponding with peak traffic periods.
- In March, the 1st was a Friday, also a weekday, though PM2.5 exceedances were marginal compared to February.
- On April 1st, coinciding with Easter Monday, a national holiday, PM levels remained consistently below threshold values, showing a slow and regular decline throughout the day.
- Similarly, May 1st, another public holiday, exhibited low and stable PM concentrations.
- In June, the 1st fell on a Saturday, and the PM2.5 pattern resembled that of February, with noticeable exceedances during typical activity bands, though to a slightly lesser extent.

These results underscore the strong influence of human activity patterns, particularly traffic and public holidays on the daily dynamics of air pollution in urban environments.

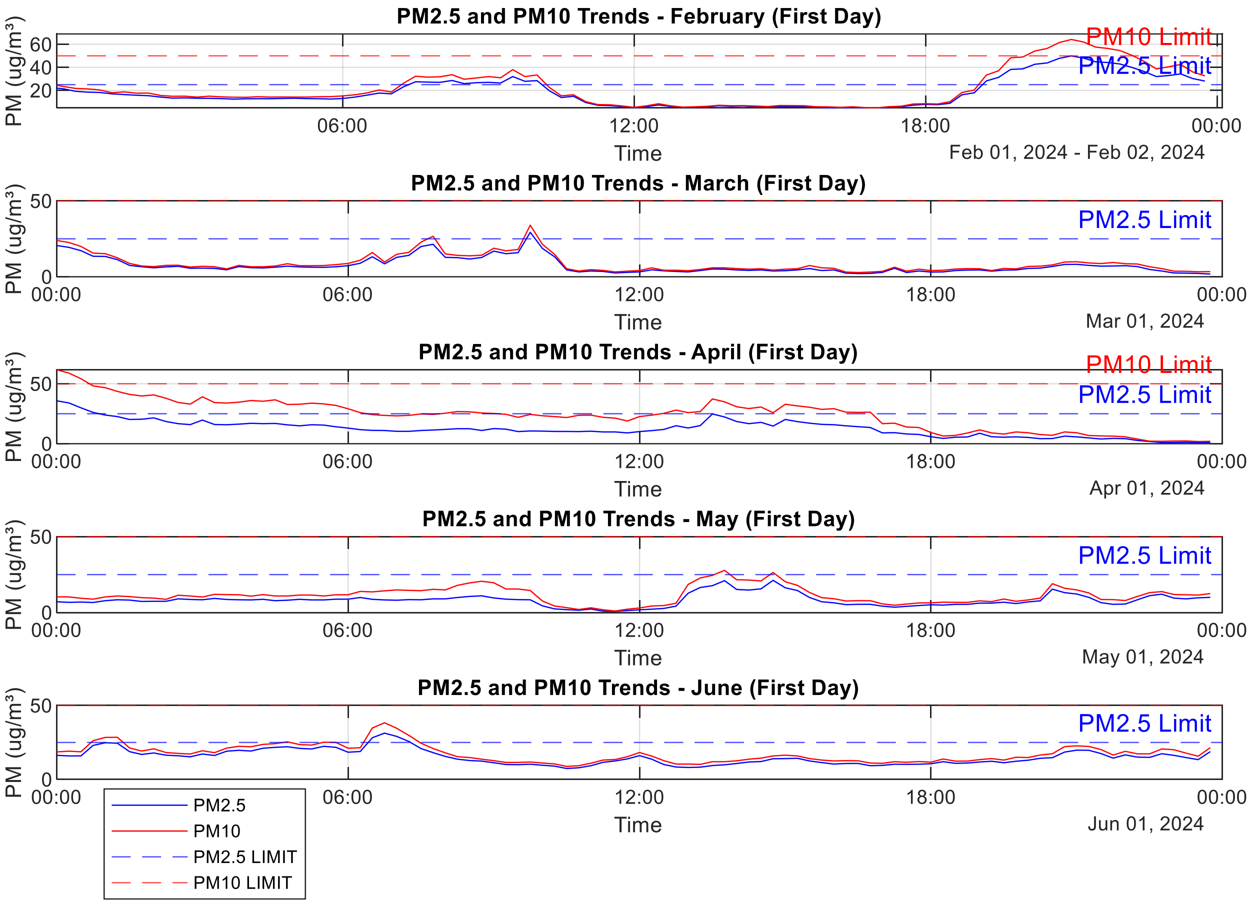

Fig. 4. Daily variation of PM2.5 and PM10 concentrations recorded on the first day of each selected month between February and June 2024.

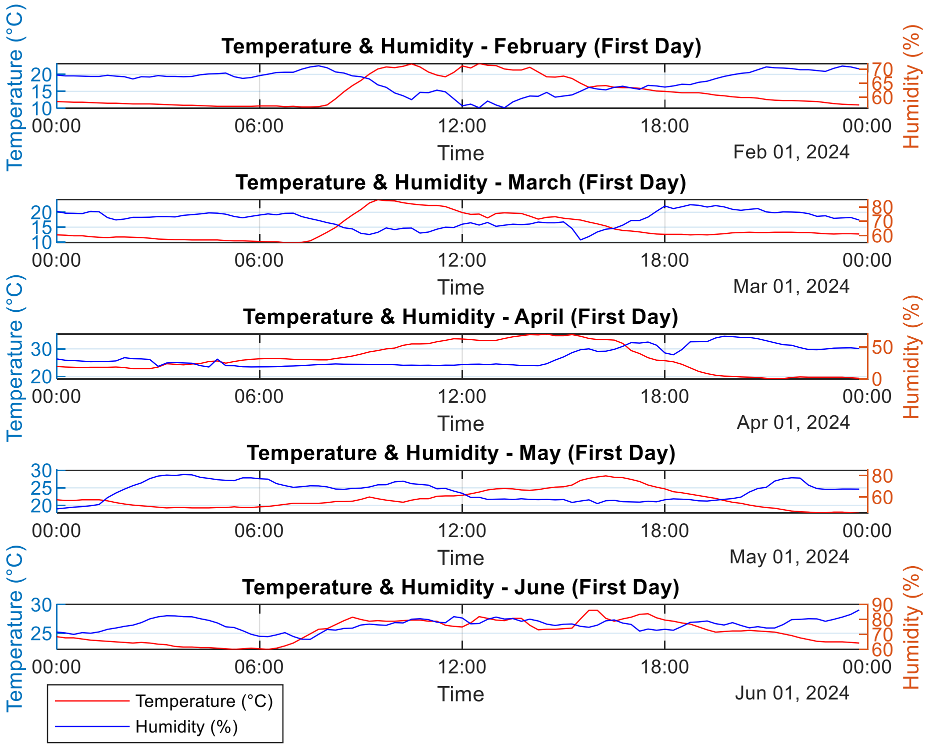

Fig. 5. Temperature and relative humidity trends recorded on the first day of each selected month between February and June 2024.

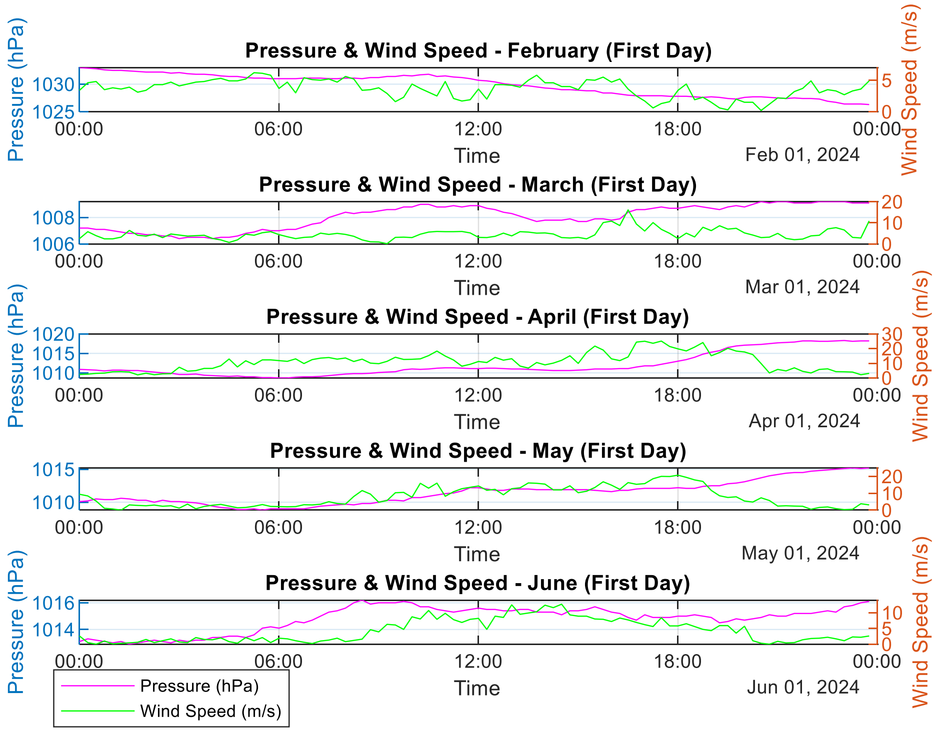

Fig. 6. Atmospheric pressure and wind speed recorded on the first day of each selected month between February and June 2024.

3.2. Data Filtering and Correlation Computation

To further investigate the influence of environmental conditions on particulate matter levels, a correlation analysis was performed using data collected on the first day of each selected month. As highlighted in Figure 4, PM concentrations are strongly affected by vehicular traffic patterns, which tend to vary considerably depending on whether the day falls on a weekday or public holiday.

The dataset was pre-processed to extract only the data corresponding to the first day of each month (February through June 2024), ensuring a focused analysis that eliminates intra-month variability and captures consistent temporal snapshots across different environmental contexts.

Following data filtering, Pearson correlation coefficients were computed for each pair of variables, including: PM2.5, PM10, Temperature, Humidity, Atmospheric Pressure, Wind Speed

The resulting correlation matrix provides a quantitative assessment of the relationships among these variables:

- A coefficient close to +1 denotes a strong positive correlation (e.g., an increase in temperature is associated with an increase in PM2.5 levels).
- A coefficient close to -1 indicates a strong negative correlation (e.g., an increase in wind speed is associated with a decrease in PM10 concentrations).
- Values near 0 suggest little or no linear relationship between the variables.

This correlation framework enables a better understanding of the meteorological and anthropogenic factors influencing particulate matter dynamics in urban settings, offering a basis for predictive modeling and targeted mitigation strategies.

The correlation matrices for the two days considered are represented in Table I and II.

Table 1. Correlation matrix - February

	PM25	PM10	Temp	Humidity	Pressure	WindSpeed
PM25	1	0.995	-0.375	0.684	-0.315	-0.304
PM10	0.995	1	-0.361	0.66691	-0.3509	-0.319
Temp	-0.375	-0.361	1	-0.887	-0.011	-0.177
Humidity	0.684	0.666	-0.887	1	-0.062843	0.0351
Pressure	-0.315	-0.35	-0.011	-0.0628	1	0.398
Wind Speed	-0.304	-0.319	-0.177	0.0351	0.398	1

Table 2. Correlation matrix - June

	PM25	PM10	Temp	Humidity	Pressure	WindSpeed
PM25	1	0.995	-0.800	-0.429	-0.532	-0.733
PM10	0.995	1	-0.781	-0.437	-0.501	-0.715
Temp	-0.800	-0.78	1	0.0850	0.657	0.726
Humidity	-0.429	-0.437	0.0850	1	0.189	0.347
Pressure	-0.532	-0.501	0.657	0.189	1	0.523
Wind Speed	-0.733	-0.715	0.726	0.347	0.523	1

3.3. *Correlation Insights from February (First Day)*

The analysis of the correlation matrix for February 1st reveals several noteworthy relationships among environmental variables, particularly in relation to particulate matter concentrations.

Strong Positive Correlations were observed between:

- PM2.5 and PM10 (r = 0.99795): As anticipated, these two pollutants exhibited a near-perfect positive correlation, indicating a common origin primarily vehicular emissions, road dust, and industrial activities.
- Humidity and PM2.5/PM10 (r ≈ 0.68–0.66): This suggests that higher humidity levels may be associated with increased particulate matter concentrations. A plausible explanation is that water vapor condenses onto airborne particles, increasing their mass and reducing their dispersion potential, thereby prolonging their atmospheric residence time.

Negative Correlations were also identified:

- Temperature and Humidity (r = –0.887): This inverse relationship is consistent with well-established meteorological dynamics, whereby rising temperatures typically lower relative humidity, especially in the absence of precipitation.
- Wind Speed and PM Concentrations (r ≈ –0.30): A moderate negative correlation indicates that higher wind speeds facilitate the dispersion of particulate matter, leading to reduced PM2.5 and PM10 concentrations.
- Pressure and PM2.5/PM10 (r ≈ –0.31 to –0.35): Although the correlation is relatively weak, the negative sign may reflect a tendency for lower atmospheric pressure to favor pollutant accumulation near the surface, due to reduced vertical air mixing and increased atmospheric stability.

These insights highlight the complex interplay between meteorological conditions and urban air pollution dynamics, reinforcing the necessity of incorporating environmental variables into air quality prediction models and policy planning.

3.4. *Correlation Insights from June (First Day)*

The correlation analysis for June 1st reveals patterns similar to those observed in February, with some notable seasonal differences attributed to higher temperatures and enhanced atmospheric dynamics typical of late spring and early summer.

A strong positive correlation persists between:

- PM2.5 and PM10 (r = 0.99537): This consistent, high correlation reaffirms the shared origin of these particulate pollutants, likely stemming from traffic emissions, road dust, and other urban combustion sources.

However, temperature emerges as a more influential factor in June compared to the winter months:

- Temperature vs. PM2.5 and PM10 (r = –0.80 and –0.78): These strong negative correlations suggest that higher temperatures contribute to the reduction of PM concentrations. This could be due to enhanced thermal convection, which promotes vertical mixing of the air, or reduced emissions from residential heating systems that are inactive during warmer months.
- Temperature vs. Pressure (r = 0.657): This moderate positive correlation indicates that higher temperatures coincide with higher pressure systems, possibly reflecting stable, anticyclonic conditions typical of the season.

Wind speed also exhibits stronger associations with both particulate matter and temperature:

- Wind Speed vs. PM2.5 and PM10 (r = –0.73 and –0.71): These values denote a more pronounced cleansing effect of wind, as higher speeds facilitate pollutant dispersion more efficiently in summer compared to winter.
- Wind Speed vs. Temperature (r = 0.726): This strong positive correlation may indicate that hotter days are associated with more active atmospheric conditions, potentially driven by diurnal heating and convective processes.

In contrast to the results from February, humidity shows a weaker and slightly negative relationship with particulate matter:

- Humidity vs. PM2.5 and PM10 (r ≈ –0.42 to –0.43): Unlike the positive association observed in colder months, this weaker inverse correlation suggests that in summer, humidity plays a secondary role, possibly overshadowed by higher wind speeds and thermal dispersion mechanisms.

These findings reinforce the importance of seasonal context in interpreting environmental correlations and stress the need for adaptive air quality models that account for dynamic meteorological influences.

4. Description of the Predictive Analysis

The predictive component of this study utilizes a multiple linear regression model to estimate PM10 and PM2.5 concentrations based on a set of meteorological predictors: atmospheric pressure, relative humidity, wind speed, and ambient temperature.

The methodology follows a structured workflow composed of the following key steps, Fig. 7.

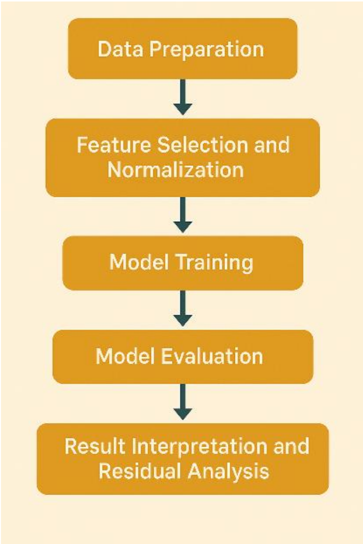

Fig. 7. Structured workflow for PM prediction.

- Data Preparation: The dataset is imported from Excel or CSV format and subjected to pre-processing operations. These include handling missing values, standardizing measurement units, and ensuring time synchronization across all variables. A continuous time index is generated to facilitate temporal alignment of the observations and to enable time-based feature extraction if needed. A continuous time index is computed:

$$T = Day + (Hour/24) \qquad (1)$$

- Feature Selection and Normalization: The independent variables (temperature, humidity, wind speed, and pressure) are selected as predictors. To ensure comparability and improve model stability, the features are normalized using min-max scaling or z-score standardization, depending on the distribution of each variable.
- Model Training: Separate regression models are trained for PM2.5 and PM10 using a portion of the dataset (e.g., 70% for training). The model assumes a linear relationship between the input variables and the target output. The regression equation takes the general form:

$$PM = \beta_0 + \beta_1 P + \beta_2 H + \beta_3 W + \beta_T T + \epsilon \qquad (2)$$

where PM is the predicted concentration of PM10 or PM2.5, P is Atmospheric pressure, H is Humidity, W is Wind speed, T is Temperature, βi are the estimated regression coefficients, ϵ is the error term. The model is trained using MATLAB's fitlm function

- Model Evaluation: The trained model is evaluated on the remaining 30% of the data using performance metrics such as R^2 (coefficient of determination), RMSE (Root Mean Square Error), and MAE (Mean Absolute Error). These metrics quantify the model's ability to capture variability and provide accurate predictions.
- Result Interpretation and Residual Analysis: The regression coefficients are analyzed to determine the relative impact of each meteorological factor on PM concentrations. Additionally, residual plots are generated to assess the goodness of fit and to identify potential patterns or model biases.

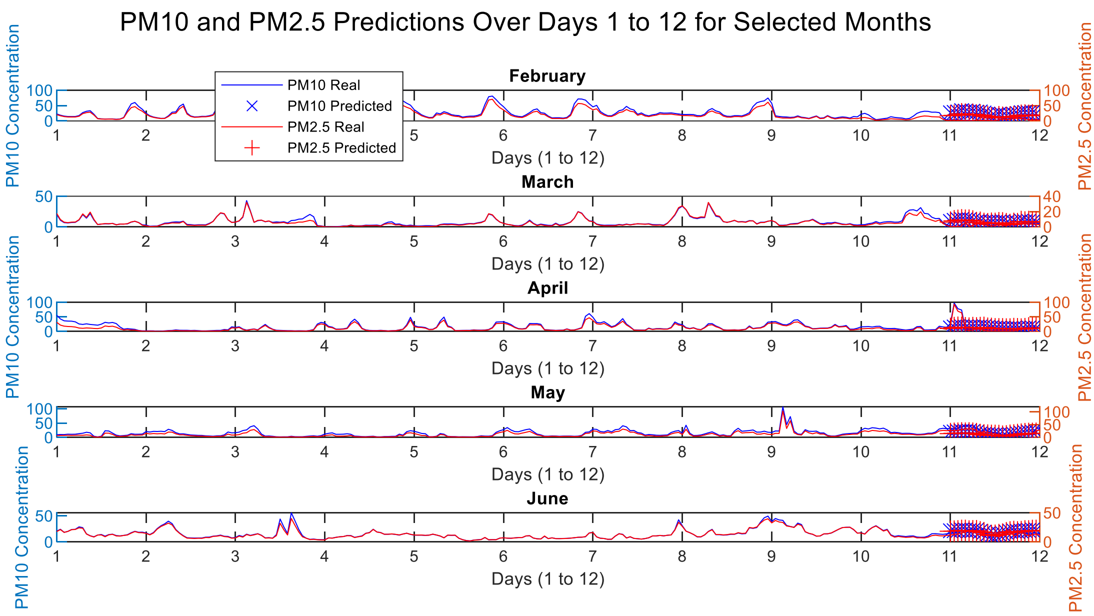

Fig. 8. Overview of the training period and prediction results (The plot displays the observed PM10 values during the training phase (July–December) and the predicted values for subsequent days (January and March)).

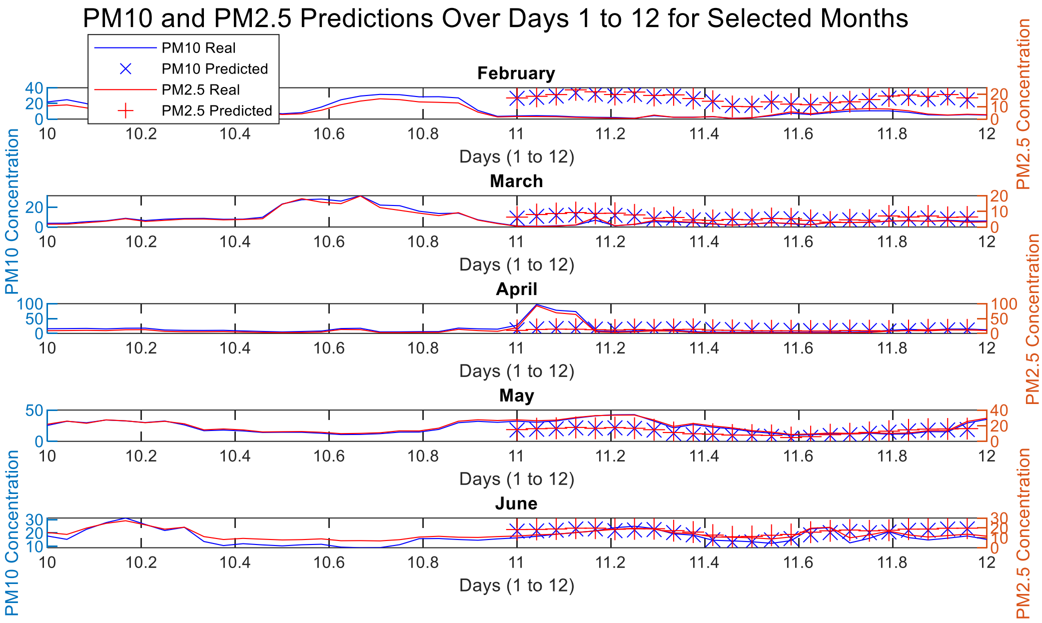

Fig. 9. Detailed view of the predicted PM10 values compared to observed data for selected days in February to June.

5. Discussion on Prediction Reliability

The predictive performance of the linear regression model is evaluated through the analysis of Figures 8 and 9, which present the forecasted values of PM10 and PM2.5 for days 11 and 12 of selected months. In Figure 8 an overview of the training period and prediction results is shown. The plot displays the observed PM10 values during the training phase (July–December) and the predicted values for subsequent days (January and March). This comparison allows assessment of the model's ability to generalize across different seasonal conditions. These forecasts are compared against the observed values to assess both accuracy and consistency. In figure 9 a detailed view of the predicted PM10 values compared to observed data for selected days in February to June. The zoomed-in representation highlights the accuracy and deviations of the model on short time scales, allowing for a more precise evaluation of prediction performance

5.1. General Consistency of Predictions

The predicted values for PM10 (represented by blue 'X' markers) and PM2.5 (red '+' markers) generally reflect the overall trend of the observed data from preceding days. This suggests that the model is able to approximate the general behavior of particulate matter based on meteorological inputs, capturing broad temporal dynamics effectively.

5.2. Prediction Accuracy for Days 11 and 12

In several instances, the predicted values align closely with the real measurements, indicating a good fit and reinforcing the relevance of the selected predictors. However, PM2.5 predictions occasionally diverge from actual values, showing signs of both underestimation and overestimation, especially in periods characterized by more abrupt fluctuations.

5.3. Variability in the Time Series

The time series for both PM10 and PM2.5 reveals considerable short-term variability, including sharp increases or decreases that are difficult to reproduce through a linear model. This highlights a limitation in capturing non-linear or sudden phenomena, such as traffic surges, construction dust, or sudden meteorological shifts.

5.4. Possible Model Limitations

The use of a linear regression model imposes inherent restrictions in terms of flexibility. The relationship between air pollutants and environmental factors is likely to be non-linear and multivariate, with possible interactions among variables. Moreover, the exclusion of other potentially influential factors—such as traffic density, industrial activity, and precipitation—may reduce the predictive power and contribute to observed discrepancies.

5.5. Strengthening the Predictive Model

Despite these limitations, the model demonstrates promise, particularly when built on well-correlated variables, as seen in the previously analysed correlation matrices. To improve prediction reliability, the following enhancements are proposed:

- Expand the training dataset to include a longer time series, thereby improving generalization and robustness.
- Focus on high-correlation predictors, particularly PM2.5 and PM10, which exhibit strong internal consistency and shared behavioral trends.

In the subsequent phase, the model is retrained using data from July to December, and then tested on the months of January and March, allowing a more comprehensive evaluation through comparison between real PM10 values and their corresponding predicted counterparts.

Figure 10 shows the training and prediction results for PM10. The graph displays the actual PM10 concentrations used for model training (July–December) and the corresponding predicted values for testing months (January and March). This comparison highlights the model's performance across different seasonal conditions and validates its capacity to generalize beyond the training period.

Figure 11 shows the PM10 training and prediction results using filtered samples. The dataset was pre-processed to exclude anomalous or incomplete records, improving data quality. The figure shows the model's performance after training on the cleaned dataset (July–December) and predicting PM10 levels for selected days in January and March, demonstrating enhanced consistency and reduced error margins compared to the unfiltered approach. Finally Figure 12 shows a detailed view of PM10 concentrations during the first two weeks of February. The zoomed-in analysis highlights short-term variability and peak events, offering insight into daily fluctuation patterns and the model's potential sensitivity to rapid changes in particulate matter levels

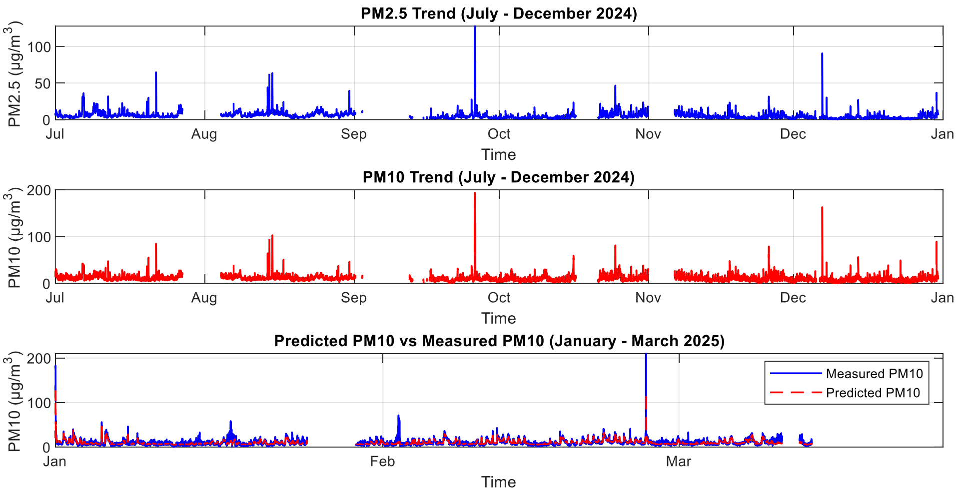

Fig. 10. Training and prediction results for PM10.

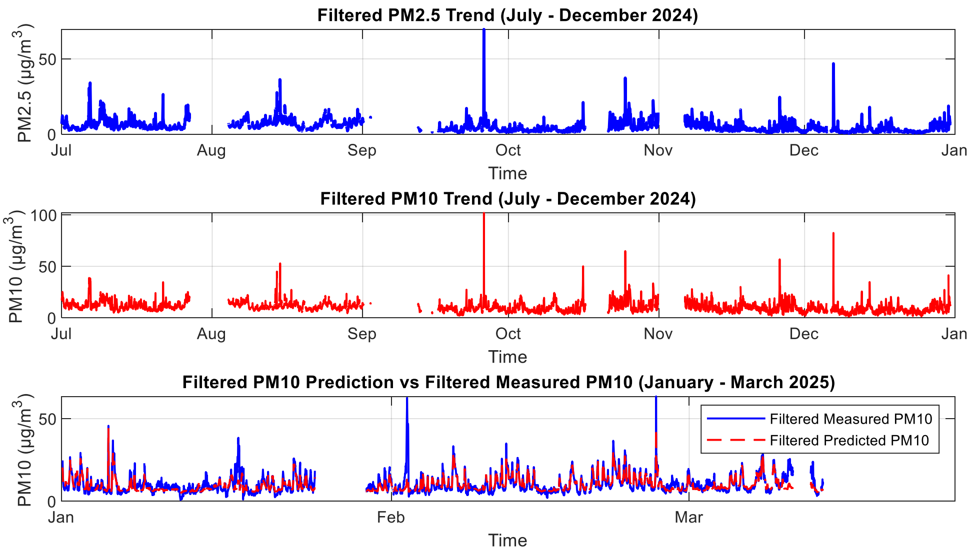

Fig. 11. PM10 training and prediction results using filtered samples.

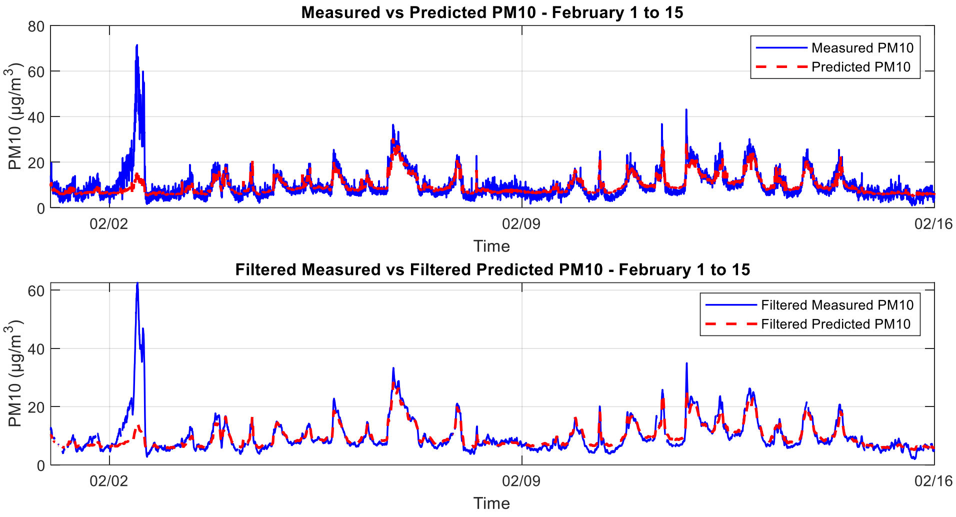

Fig. 12. Detailed view of PM10 concentrations during the first two weeks of February.

6. The Role of Artificial Intelligence in Advanced Predictive Approach Without Direct PM Input

Up to this point, the predictive techniques applied relied on the availability of at least one particulate matter measurement (e.g., PM10) to estimate the other (e.g., PM2.5). These approaches have demonstrated that the accuracy of predictions increases significantly with the inclusion of additional environmental variables, such as temperature, humidity, and atmospheric pressure.

In the next phase of analysis, the objective is to develop a predictive model capable of estimating PM concentrations without direct input from any particulate measurements. Instead, the model will rely exclusively on meteorological data to uncover underlying correlations and generate estimates of PM10 and PM2.5 levels.

This transition represents a shift toward a fully indirect estimation framework, where the particulate matter behavior is inferred through its statistical and physical relationships with environmental factors. This approach is particularly relevant for applications where direct PM sensing is unavailable, unreliable, or cost-prohibitive.

The model will be trained using datasets where PM concentrations are known, but during prediction, the input will consist solely of temperature, humidity, pressure, and wind speed. This method will be tested for its ability to reconstruct particulate trends and identify critical pollution episodes based on environmental conditions alone.

6.1. Overview of Applied Models

Four models were tested and compared:

6.1.1. Multiple Linear Regression (MLR)

Used as a baseline, MLR assumes a linear relationship between input variables and the target pollutant concentration. While it offers excellent interpretability, it is limited in its ability to model complex interactions or threshold effects commonly observed in air quality dynamics.

Key Function:

- fitlm – fits a linear regression model.

Example:

% X: predictor matrix (e.g., temperature, humidity, etc.)

% Y: target variable (e.g., PM10)

mdl = fitlm(X, Y);

% Prediction

Y_pred = predict(mdl, X_test);

6.1.2. Random Forest Regressor (RFR)

An ensemble learning model composed of multiple decision trees. RFR is capable of modeling non-linearities and variable interactions, and it includes internal mechanisms for estimating feature importance. It is particularly robust against overfitting and noise in the data.

Key Functions:

- fitrensemble – fits an ensemble of learners (e.g., bagged trees).
- templateTree – defines the base decision tree learner.

% Random Forest with 100 trees

```
t = templateTree('MaxNumSplits', 20);
mdl_rf = fitrensemble(X, Y, 'Method', 'Bag', 'NumLearningCycles', 100, 'Learners', t);
% Prediction
Y_pred_rf = predict(mdl_rf, X_test);
```

6.1.3. Support Vector Regression (SVR)

SVR uses kernel functions (e.g., radial basis function) to map the input features into higher-dimensional spaces, allowing it to capture non-linear patterns. It is effective with smaller datasets but requires careful hyperparameter tuning to balance model complexity and generalization.

Key Function:

- fitrsvm – trains a support vector machine for regression.

Example:

```
mdl_svr = fitrsvm(X, Y, 'KernelFunction', 'gaussian', 'KernelScale', 'auto', 'BoxConstraint', 1);
% Prediction
Y_pred_svr = predict(mdl_svr, X_test);
```

6.1.4. Artificial Neural Networks (ANNs)

A feedforward ANN architecture with one hidden layer is employed. The network uses ReLU activation functions and is optimized via backpropagation. ANNs excel in capturing complex, non-linear dependencies between input variables and target outputs, although they demand larger datasets and computational resources.

Key Functions (Deep Learning Toolbox):

- fitnet – creates a feedforward neural network for regression.
- train – trains the neural network.

```
% Neural network with 10 neurons in the hidden layer
net = fitnet(10);

% Data split: training/validation/testing
net.divideParam.trainRatio = 0.7;
net.divideParam.valRatio = 0.15;
net.divideParam.testRatio = 0.15;
% Training
[net, tr] = train(net, X', Y');
% Prediction
Y_pred_ann = net(X_test');
Y_pred_ann = Y_pred_ann';   % Transpose for compatibility
```

6.2. Feature Selection and Model Inputs

The input variables used in all models are: Temperature, Humidity, Atmospheric Pressure, Wind Speed

Feature selection was guided by prior correlation analysis, which indicated significant influence of these meteorological factors on PM levels. In some configurations, polynomial terms or interaction features were also included to improve model expressiveness.

6.3. Model Evaluation Metrics

Each model was trained on data from July to December and tested on January and March. The evaluation metrics used are:

RMSE (Root Mean Square Error): Measures the standard deviation of prediction errors.

MAE (Mean Absolute Error): Measures the average magnitude of the errors.

R^2 (Coefficient of Determination): Measures how well the predicted values match the actual values.

Table 3. Performance comparison of ai models on PM10 prediction

Model	RMSE (μg/m³)	MAE (μg/m³)	R^2 Score
Multiple Linear Regression (MLR)	9.12	6.84	0.68
Random Forest Regressor (RFR)	5.41	4.13	0.87
Support Vector Regression (SVR)	6.73	5.01	0.79
Artificial Neural Network (ANN)	5.96	4.72	0.83

The script used is:

```
% RMSE
rmse = sqrt(mean((Y_test - Y_pred).^2));
% MAE
mae = mean(abs(Y_test - Y_pred));
% R²
SST = sum((Y_test - mean(Y_test)).^2);
SSR = sum((Y_test - Y_pred).^2);
r2 = 1 - SSR/SST;
```

6.4. Discussion of Results

In Fig. 13, a comparison is performed. The x-axis represents the hourly timeline throughout April 8, 2024 (00:00–24:00). The y-axis indicates PM10 concentration in µg/m³.

The black solid line shows the measured PM10 values.

The dashed colored lines represent predictions from four different models:

- Linear Regression (blue dashed line): Shows a relatively flat and stable trend. Tends to underestimate rapid fluctuations, especially in the second half of the day. Captures the general trend, but fails to reproduce the peaks accurately.
- Random Forest (red dashed line): Reacts more dynamically to variations in the data. Performs relatively well in the middle portion of the day but tends to overestimate values in the early hours and flatten out later in the day. Shows a rougher curve, typical of ensemble models capturing local variance.
- SVR (magenta dashed line): Produces a smoother prediction curve. Appears to be closer to the measured values, especially during mid-day. Offers a balanced compromise between responsiveness and overfitting, though it may miss sharp local peaks.
- ANN (green dashed line): The neural network exhibits significant oscillations and amplitude extremes. It strongly overestimates and underestimates PM10 in various time windows.This behavior suggests possible overfitting or poor generalization, potentially due to limited training data or over-complex network structure.

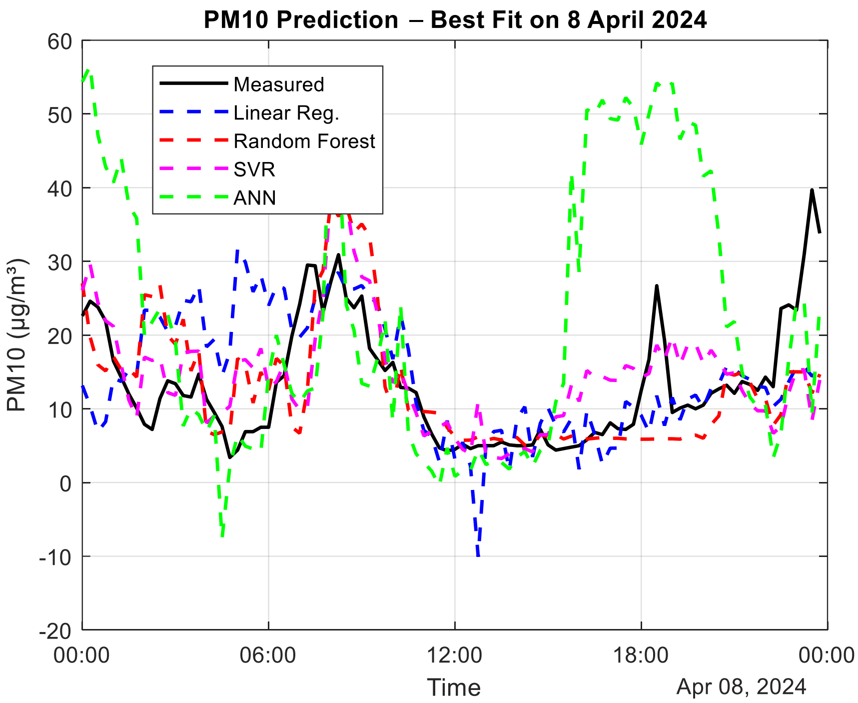

Fig. 13. Comparison between measured PM10 concentrations (black line) and predictions from four machine learning models on April 8, 2024.

In Figure 14 a comparison of measured PM2.5 concentrations (black line) and predicted values from four machine learning models on April 8, 2024. Models were trained using meteorological data (temperature, humidity, pressure, wind speed) from the preceding seven days. Predicted values are shown for Linear Regression (blue dashed), Random Forest (red dashed), Support Vector Regression – SVR (magenta dashed), and Artificial Neural Network – ANN (green dashed). The figure highlights the differences in predictive behavior, with SVR and Random Forest providing the closest fit to observed trends, while ANN exhibits greater variability.

6.5. Visual Performance Summary

The SVR model appears to provide the most visually coherent prediction, with a good balance between smoothness and accuracy. The ANN shows the least reliable behavior, possibly due to unstable training or excessive sensitivity to noise. Linear regression, as expected, is too simplistic and struggles to follow non-linear dynamics. The Random Forest captures some trends well but lacks consistency in key intervals.

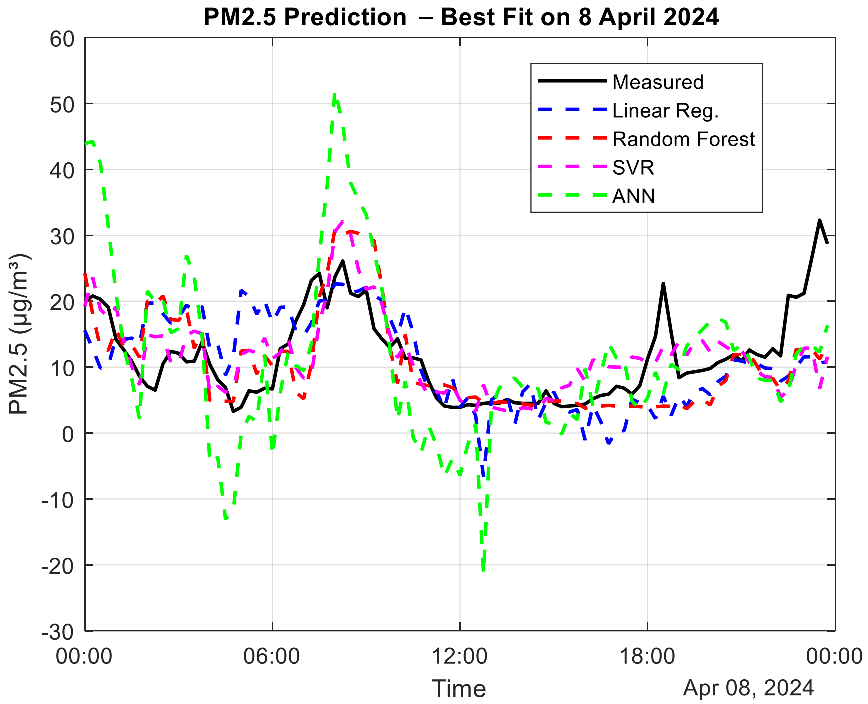

Fig. 14 Comparison of measured PM2.5 concentrations (black line) and predicted values from four machine learning models on April 8, 2024.

In Figure 13 and 14, the model performance insights for PM10 and PM2.5 prediction are evaluated.

- Linear Regression (blue dashed line): Shows relatively stable predictions. Performs reasonably well in capturing the general baseline, especially during the second half of the day. Fails to follow short-term fluctuations and underrepresents morning and evening peaks.
- Random Forest (red dashed line): Produces predictions that track moderately well with the measured values. Captures the early-morning trend better than linear regression. Still misses finer details in peak behavior but performs consistently across the day.
- SVR (magenta dashed line): Delivers smoothed predictions that follow the shape of the actual data, especially around transitions. Closely tracks the morning peak, although it tends to underpredict higher values slightly. Offers one of the best visual fits among the models.
- ANN (green dashed line): Produces more erratic behavior, particularly with over- and underestimations in the early and mid-morning hours. Although the trend is sometimes captured, the amplitude is often exaggerated, resulting in spikes not present in the real data. This suggests a lack of generalization or a possible issue with training stability.
- Visual Summary: SVR and Random Forest appear to provide the most reliable predictions in terms of trend and magnitude. Linear Regression performs acceptably but oversimplifies PM2.5 behavior. ANN again exhibits signs of instability, particularly in handling noise or unusual variations.

7. Conclusion

This study demonstrates the effectiveness of Artificial Intelligence (AI) in modeling and predicting particulate matter concentrations (PM10 and PM2.5) using meteorological data. The impact is significant, as it allows for a clear distinction between micro- and macro-scale aspects. Trends can be analyzed by focusing on measurable variables and identifying numerical correlations with scalar quantities such as temperature and humidity, as well as vector quantities like wind. This approach also enables the detection of specific conditions—for example, the absence of vehicular traffic on holidays, when schools are closed and human activity is reduced.

The next step—exploring predictive models—is greatly simplified to new artificial intelligence tools that allow to write and check the formulation of straightforward MATLAB scripts for easily visualizing trends through a conversational programming interface.

By comparing several AI-based regression models—including Linear Regression, Random Forest, Support Vector Regression (SVR), and Artificial Neural Networks (ANN)—the work highlights the potential of machine learning in capturing complex, non-linear relationships between environmental variables and air pollution levels. Among the tested models, SVR and Random Forest consistently delivered more accurate and stable forecasts, while ANN showed greater sensitivity to noise and potential overfitting under limited data conditions. These findings confirm that AI can serve as a powerful, data-driven alternative to traditional statistical approaches, especially in urban contexts where sensor deployment is sparse or real-time measurements are limited.

7.1. Future Developments

Future research will focus on advancing AI architectures and expanding the scope of input data to further improve predictive reliability. Incorporating deep learning models, such as recurrent neural networks (RNNs) or long short-term memory networks (LSTMs), could allow the system to learn temporal dependencies and better forecast pollution dynamics over time. Additionally, multi-source data fusion, integrating traffic intensity, satellite imagery, mobility patterns, and urban morphology, will enable the development of context-aware AI models.

Another promising direction is the implementation of online learning algorithms, allowing models to update continuously as new data streams in, making them more adaptive to real-time environmental changes. Finally, embedding AI models into IoT-enabled air quality monitoring systems and smart city platforms would enable autonomous, distributed, and scalable environmental prediction infrastructures, supporting data-informed decision-making for public health and urban planning.

Acknowledgements

This study was carried out within the MOST– Sustainable Mobility Center and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR)– MISSIONE 4 COM PONENTE 2, INVESTIMENTO 1.4– D.D. 1033 17/06/2022, CN00000023). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

References

[1] A. Chaloulakou, P. Kassomenos, N. Spyrellis, P. Demokritou, P. Koutrakis, "Development of a model for forecasting of PM10 concentrations in Athens, Greece", Atmospheric Environment, DOI: 10.1016/S1352-2310(03)00565-X, Vol. 37, No. 30, pp. 4263–4275, 2003.

[2] G. Grivas, A. Chaloulakou, "Artificial neural network models for prediction of PM10 hourly concentrations, in the Greater Area of Athens, Greece", Atmospheric Environment, DOI: 10.1016/j.atmosenv.2005.11.001, Vol. 40, No. 7, pp. 1216–1229, 2006.

[3] H. Huang, W. Deng, Y. Deng, "Using artificial neural networks to predict indoor particulate matter concentrations in residential buildings", Building and Environment, DOI: 10.1016/j.buildenv.2019.106450, Vol. 167, 106450, 2020.

[4] M. Maleki, A. Rezaei, "Probabilistic Deep Learning to Quantify Uncertainty in Air Quality Forecasting", Environmental Modelling & Software, DOI: 10.1016/j.envsoft.2021.105113, Vol. 143, 105113, 2021.

[5] J. Liu, Predictive Artificial Intelligence Techniques: Regression Analysis and MATLAB Simulation, Springer, 2024.

[6] A.M. Anasori, M.R. Sabour, "Neural Network and Multiple Regression Models for PM10 Prediction in Athens: A Comparative Assessment", Lambert Academic Publishing, 2020.

[7] https://nettigo.pl/attachments/ 3983E

[8] https://davisenviromonitors. com/davis-airlink

[9] https://www.who.int/news-room/questions-and-answers/item/who-global-air-quality-guideline

[10] The MathWorks Inc, MATLAB and Simulink In-Depth: Model-based Design with Simulink, MathWorks Press, 2023.