In-Vehicle $PM_{2.5}$ Prediction during Commuting Periods Based on Fuzzy Clustering and Dynamically Weighted Neural Networks

Xiangyu Li[a], Xi Cheng[b], Gabriel Gomes[c], Yizheng Wu[d]

[a]*Department of Civil and Environmental Engineering, Northwestern University, 633 Clark St, Evanston, 60208, IL, USA*
[b]*Department of Civil and Environmental Engineering, Cornell University, 726 University Ave, Ithaca, 14850, NY, USA*
[c]*Department of Mechanical Engineering, University of California, Berkeley, 6141 Etcheverry Hall, Berkeley, 94720, CA, USA*
[d]*Key Laboratory of Transport Industry of Big Data Application Technologies for Comprehensive Transport, Beijing Jiaotong University, 3 Shangyuancun, Haidian District, Beijing, 100044, PR China*

Abstract

Urban traffic growth has intensified air pollution, raising public health concerns, particularly for commuters exposed to localized high concentrations of vehicle emissions. This study presents an innovative framework for accurately predicting in-vehicle $PM_{2.5}$ concentrations, integrating data-driven insights with advanced machine learning techniques. A dynamic ensemble learning model was developed, combining General Regression Neural Networks (GRNN), Convolutional Neural Networks (CNN), and Attention Regressors. Multi-dimensional data, including meteorological parameters, traffic conditions, and in-vehicle sensor readings, were collected and preprocessed to ensure robust model performance. Using Fuzzy C-Means (FCM) clustering, input samples were partitioned into subspaces, enabling dynamic weighting of model predictions based on subspace-specific errors. Experimental results demonstrated that the proposed framework significantly outperformed individual models and traditional static ensemble methods, achieving lower Mean Squared Error (MSE) and Mean Absolute Error (MAE), as well as a higher coefficient of determination (R^2). Cyclists were identified as the most vulnerable commuters, experiencing the highest $PM_{2.5}$ exposure concentrations (104 $\mu g/m^3$) and inhaled doses (13 $\mu g/km$). The dynamic ensemble framework effectively captured spatial-temporal variations, providing actionable insights for improving air quality, optimizing vehicle ventilation strategies, and mitigating health risks. These findings contribute to urban sustainability by supporting data-driven decision-making in air quality management and transportation planning.

Keywords: Urban air pollution, Commuting exposure, PM2.5 concentration, Ensemble learning, Dynamic Ensemble

1. Introduction

Urbanization and the rapid growth of transportation systems have significantly increased air pollution, particularly in densely populated metropolitan areas such as Beijing. Traffic emissions are now among the leading contributors to ambient air pollution, with motor vehicles responsible for nearly half of the PM2.5 concentration levels in urban settings. These emissions exacerbate public health concerns, necessitating comprehensive strategies to mitigate pollution exposure. The Chinese government has recognized the urgency of addressing this issue, as evidenced by initiatives such as the “blue sky defense battle,” the Green Travel Action Plan, and broader commitments to carbon neutrality. However, effectively addressing pollution exposure requires nuanced insights into its spatial and temporal dynamics, particularly within traffic microenvironments where pollutant concentrations are highly localized.

PM2.5, fine particulate matter with a diameter of less than 2.5 µm, is especially problematic due to its ability to carry harmful substances, persist in the atmosphere, and penetrate deep into the human respiratory system. Exposure to PM2.5 has been linked to severe health outcomes, including cardiovascular and respiratory diseases, and premature mortality. Urban commuters, often exposed to high concentrations of PM2.5 during daily travel, represent a vulnerable population. Studies have shown that prolonged commuting accounts for a substantial portion of daily pollutant inhalation, with exposure levels varying significantly by transportation mode and route.

Existing research has primarily focused on pollutant concentrations in specific microenvironments without adequately capturing the dynamic interplay between commuting behaviors and exposure patterns. Moreover, the health impacts of commuting-related PM2.5 exposure remain understudied, particularly concerning demographic differences such as age and gender. To address these gaps, this study investigates the PM2.5 exposure levels of commuters in Beijing's Haidian District across four transportation modes—taxi, bus, bicycle, and electric motorcycle. By examining the spatial and temporal distribution of PM2.5 concentrations and their correlation with meteorological conditions and background pollutant levels, this study provides a comprehensive assessment of commuting-related exposure.

Leveraging a hybrid ensemble learning framework, we integrate static and dynamic methodologies to enhance the accuracy and reliability of PM2.5 predictions. This approach captures both the stability of long-term exposure trends and the adaptability needed for real-time predictions in complex urban environments. Through this framework, we analyze the exposure dose of PM2.5 across demographic groups and commuting scenarios, offering data-driven strategies to improve commuter health and inform urban pollution management policies.

This study makes the following contributions: 1. It develops and applies a hybrid static-dynamic ensemble learning framework that integrates models like GRNN, CNN, and AttentionNet for accurate PM2.5 predictions. 2. It offers a quantitative analysis of inhalation doses for different commuting modes, highlighting the most vulnerable groups and scenarios. 3. It provides actionable recommendations for mitigating health risks, including optimizing commuting routes and leveraging real-time air quality monitoring systems.

The remainder of this paper is organized as follows: Section 2 reviews the existing literature on traffic-related air pollution and predictive modeling. Section 3 describes the hybrid ensemble learning framework and data acquisition methods used in this study. Section 4 presents experimental results and evaluates the effectiveness of the proposed methodology. Finally, Section 5 concludes with key findings, implications, and directions for future research.

2. Literature Review

Traffic-related air pollution has emerged as a critical source of urban air quality degradation, with significant implications for human health. Pollutants such as particulate matter (PM), polycyclic aromatic hydrocarbons, and other vehicular emissions are well-documented for their adverse effects, including increased risks of lung cancer and cardiovascular diseases [4, 1]. Among these, PM2.5—fine particulate matter with a diameter of less than 2.5 μm—presents a particular threat due to its high surface area, capacity to adsorb toxic substances like heavy metals, and prolonged atmospheric persistence, which exacerbates its impacts on air quality and public health [12].

2.1. Pollutant-Centered Studies

Numerous studies have demonstrated that urban areas with dense traffic, such as Beijing, experience significantly elevated PM2.5 pollution levels. For example, [21] reported that road-mobile sources contributed 23.49% to PM10 levels and 30.40% to PM2.5 concentrations in Xi'an. Similarly, [2] attributed 17% of PM2.5 in Beijing's urban environment to vehicular emissions. These findings underscore the necessity of investigating traffic-related PM2.5 exposure, particularly given the increasing number of motor vehicles in cities like Beijing, which now exceeds 6.57 million.

The exposure patterns of urban populations to PM2.5 vary based on commuting mode, route, and time. [9] highlighted that cyclists traveling through high-traffic routes are particularly vulnerable to ultrafine particles (UFPs), while [7] observed significantly elevated PM10 exposure levels near bus stops in Guangzhou. [10] further reported that open-air transport modes, such as rickshaws, exhibited approximately 30% higher PM2.5 exposure compared to enclosed air-conditioned vehicles. These studies emphasize the importance of understanding the spatial and temporal distribution of particulate matter exposure in urban transport microenvironments to mitigate associated health risks.

2.2. Meteorological Influences and Modeling Approaches

Meteorological factors such as wind speed, humidity, and temperature significantly influence pollutant dispersion and accumulation. [20] demonstrated a positive correlation between indoor and outdoor PM2.5 concentrations with temperature and relative humidity, while [6] observed that weaker wind speeds led to higher pollutant concentrations. These findings highlight the dynamic nature of PM2.5 distribution and its dependence on environmental conditions, necessitating advanced modeling techniques to account for such variability.

2.3. Ensemble Learning and Predictive Modeling Techniques

Recent advancements in ensemble learning techniques [16] have shown great promise in improving the accuracy and robustness of PM2.5 exposure prediction models. Ensemble learning combines multiple base models to leverage their strengths while mitigating individual weaknesses, resulting in enhanced predictive performance. This study employs both static and dynamic ensemble methods:

- **Static Ensemble Integration:** This approach aggregates predictions from base models using fixed weights, determined by prior performance metrics such as Mean Squared Error (MSE). In this study, models like General Regression Neural Network (GRNN) [15], Convolutional Neural Network (CNN) [14], and AttentionNet [19] are used to capture spatial, temporal, and meteorological features of PM2.5 data.
- **Dynamic Ensemble Integration:** Dynamic ensembles adapt model weights based on input-specific characteristics. By employing Fuzzy C-Means (FCM) clustering [3], the framework evaluates model performance within localized clusters to assign weights dynamically, ensuring greater context sensitivity.

2.3.1. Additional Advances in Ensemble Learning

Recent studies have further highlighted the potential of ensemble learning in air pollution modeling. [5] demonstrated that ensemble methods reduce variance and improve model robustness by aggregating diverse base learners. [23] explored the utility of ensemble methods such as bagging, boosting, and stacking in predictive modeling, emphasizing their applicability to complex environmental datasets. For PM2.5 exposure, [13] used deep ensemble learning to integrate convolutional and recurrent networks, achieving significant improvements in spatiotemporal pollutant prediction. Similarly, [11] proposed a hybrid ensemble integrating machine learning and physical models to address atmospheric pollutant variability. These advancements underscore the versatility of ensemble techniques in capturing the multifaceted nature of air pollution data.

2.3.2. Specific Models and Applications

The General Regression Neural Network (GRNN) effectively captures non-linear relationships in PM2.5 concentrations, making it suitable for air quality forecasting. For instance, a hybrid model integrating Ensemble Empirical Mode Decomposition (EEMD) [17] with GRNN has demonstrated improved accuracy in predicting PM2.5 concentrations by addressing nonlinear data patterns [18].

Convolutional Neural Networks (CNNs) are adept at extracting spatial features from geographic data, which is essential for modeling the spatial distribution of PM2.5. A recent study introduced a hybrid model combining CNN with a Gated Recurrent Unit (GRU) to predict PM2.5 concentrations, effectively capturing both spatial and temporal dependencies [8]. Attention mechanisms prioritize significant temporal variations, enhancing the prediction of short-term exposure trends. For example, an attention-based CNN–LSTM model has been proposed for urban PM2.5 concentration prediction, effectively learning spatiotemporal correlations and improving forecasting accuracy [22].

2.4. Significance of This Study

By integrating these models - GRNN for non-linear relationships, CNN for spatial feature extraction, and attention mechanisms for temporal prioritization - the ensemble framework balances stability and adaptability, addressing both long-term exposure trends and localized variations in urban environments.

The hybrid ensemble approach used in this study provides a comprehensive framework for modeling PM2.5 exposure with high accuracy, enabling data-driven recommendations to mitigate health risks associated with urban commuting.

3. Methodology

To accurately predict PM2.5 exposure levels within vehicle environments, this study develops an innovative hybrid ensemble framework. The methodology leverages fuzzy clustering combined with dynamic weighting across three deep learning models: General Regression Neural Network (GRNN), Convolutional Neural Network (CNN), and Attention Regressor. The following sections elaborate on the methodology's components and their contributions to enhancing prediction accuracy and robustness.

3.1. Fuzzy Clustering-Based Dynamic Weighting Framework

This study introduces a dynamic integration mechanism, employing Fuzzy C-Means (FCM) clustering to partition the feature space into subspaces. Each training sample is assigned a degree of membership to multiple clusters, facilitating nuanced weight adjustments for predictive models. The framework follows these steps:

3.1.1. Fuzzy C-Means Clustering

Given a dataset with N samples $\{x_i \in \mathbb{R}^d\}_{i=1}^N$, the FCM algorithm minimizes the objective function:

$$J_m = \sum_{i=1}^{N} \sum_{k=1}^{C} u_{ik}^m \|x_i - c_k\|^2, \tag{1}$$

where u_{ik} denotes the membership degree of sample x_i to cluster k, c_k is the cluster center, and m is the fuzzification coefficient (typically $m = 2$). Membership degrees satisfy $0 \leq u_{ik} \leq 1$ and $\sum_{k=1}^{C} u_{ik} = 1$. These degrees are used to weight model predictions dynamically based on cluster proximity.

3.2. Model Descriptions

Three complementary deep learning models are integrated within the ensemble framework to capture distinct aspects of PM2.5 concentration dynamics:

3.2.1. General Regression Neural Network (GRNN)

GRNN, a non-parametric regression model, uses kernel functions to estimate the target variable. For an input x, its output is given by:

$$\hat{y}(x) = \frac{\sum_{i=1}^{N} \exp\left(-\frac{\|x-x_i\|^2}{2\sigma^2}\right) y_i}{\sum_{i=1}^{N} \exp\left(-\frac{\|x-x_i\|^2}{2\sigma^2}\right)}, \tag{2}$$

where σ is the smoothing factor. GRNN is effective in capturing localized variations, making it suitable for environments with high temporal fluctuations.

3.2.2. Convolutional Neural Network (CNN)

To extract spatial patterns, CNN processes input features as sequences or channels. For a convolutional layer with weights $W^{(1)}$ and bias $b^{(1)}$, the output is:

$$Z^{(1)} = \sigma\big(\text{Conv}(X, W^{(1)}) + b^{(1)}\big), \tag{3}$$

where $\sigma(\cdot)$ is an activation function such as ReLU. CNN effectively identifies relationships among sensor readings, such as correlations between temperature and PM2.5 levels.

3.2.3. Attention Regressor

Attention mechanisms enhance the model's ability to focus on critical time intervals or features. For an input $X \in \mathbb{R}^{L \times d}$, the hidden representation is computed as:

$$H = \mathrm{ReLU}(XW_1 + b_1), \tag{4}$$

where $W_1 \in \mathbb{R}^{d \times h}$ and $b_1 \in \mathbb{R}^h$. Attention scores α_i are calculated as:

$$\alpha_i = \frac{\exp(s_i)}{\sum_{j=1}^{L} \exp(s_j)}, \quad s_i = H_i w_{att}, \tag{5}$$

where $w_{att} \in \mathbb{R}^h$ is a trainable parameter. Attention enhances the model's focus on key attributes, boosting performance in heterogeneous environments.

3.3. Dynamic Weighting and Ensemble Prediction

The predictions from GRNN, CNN, and Attention Regressor are combined using dynamic weights determined by their performance in different clusters:

3.3.1. Cluster-Specific Weighted MSE

For cluster k, the Weighted Mean Squared Error (WMSE) of model m is:

$$\mathrm{WMSE}_k^m = \frac{\sum_{i=1}^{N} u_{ik}^{\alpha} (\hat{y}_{i,m} - y_i)^2}{\sum_{i=1}^{N} u_{ik}^{\alpha}}, \tag{6}$$

where α is a weighting exponent.

3.3.2. Dynamic Model Weights

Model weights within each cluster are computed as:

$$\omega_k^m = \frac{1/(\mathrm{WMSE}_k^m + \epsilon)}{\sum_{m'} 1/(\mathrm{WMSE}_k^{m'} + \epsilon)}, \tag{7}$$

where ϵ prevents division by zero.

3.3.3. Final Prediction

The ensemble prediction for input x is:

$$\hat{y}_{\text{ensemble}}(x) = \sum_{k=1}^{C} u_k(x) \sum_{m=1}^{M} \omega_k^m \hat{y}_m(x). \tag{8}$$

Road Name	Serial Number	Road Grade	Length (km)	Lanes	Bus Stops	Intersections
Beijing Jiaotong University East Road	L1	Branch Road	0.44	2	2	2
College South Road	L2	Collector Road	2.00	4	4	4
Zhongguancun South Street	L3	Arterial Road	1.20	8	2	2

Table 1: Road Section Information

3.4. Advantages of the Framework

This methodology combines the strengths of individual models, ensuring robustness and adaptability across varying environments. By leveraging FCM clustering and dynamic weighting, the ensemble framework outperforms traditional static integration methods in predictive accuracy and interpretability.

4. Data Acquisition

4.1. Routes of Sampling

To investigate the PM2.5 exposure of people using different commuting modes, the study selected the commercial core area, residential area, and university urban area of the Haidian District in Beijing. This area experiences high commuting volumes with significant motor vehicle and non-motor vehicle traffic. The region was chosen due to its heavy traffic congestion, absence of non-traffic source emissions from heavy industries, and diverse commuting modes, including taxi, bus, bicycle, and electric motorcycle.

A typical commuting route starts at point A (the east gate of Beijing Jiaotong University) and ends at point B (the digital building in the Zhongguancun area), as shown in Figure 1. The route passes through three sections: Beijing Jiaotong University East Road (L1), College South Road (L2), and Zhongguancun South Street (L3), before returning to the starting point. The one-way route length is approximately 3.64 km, with a trip duration of 10-30 minutes, depending on the commuting mode and traffic conditions. Table 1 summarizes the characteristics of each road section.

4.2. Sampling Schedule

Data collection was conducted over one consecutive week from June 12 to June 18, 2022. Sampling occurred three times daily during morning and

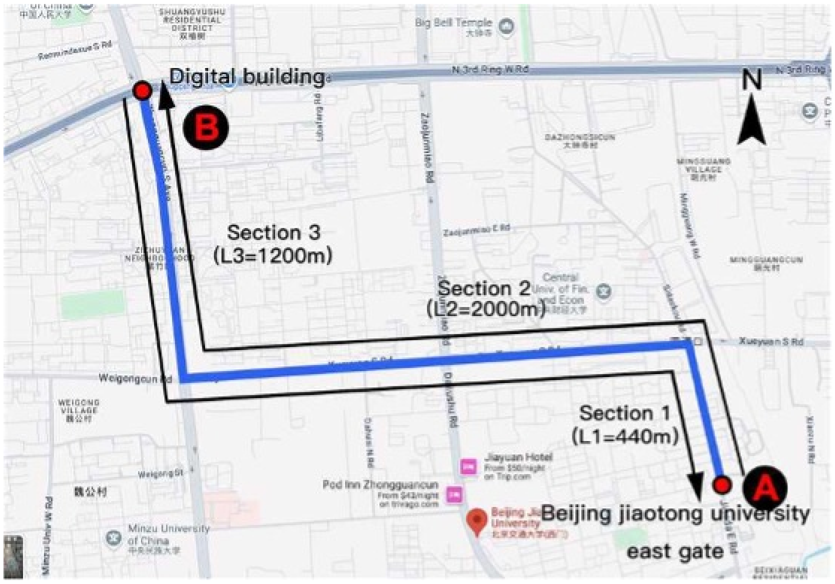


Figure 1: Study Area and Selected Routes

evening peak periods (7:30–9:30 AM, 5:00–7:00 PM) and non-peak periods (11:30 AM–2:00 PM). The schedule included both weekdays and weekends to capture temporal variations in PM2.5 exposure concentrations.

4.3. Sampling Devices

The PM2.5 concentrations were monitored using a SidePak™ AM520 Personal Aerosol Monitor from TSI, USA. This instrument uses light scattering to measure particle mass concentrations (0.1–10 μm) in real-time at 1-second intervals. The device is portable and positioned near the user's breathing zone to record exposure concentrations accurately.

Geographic location data were recorded using an Explorer V-900 navigation recorder, capturing latitude and longitude information simultaneously with PM2.5 concentrations. During data collection, trained personnel wore both devices while commuting via different transportation modes. Additionally, air quality and meteorological data were retrieved from the Wanliu air monitoring station (2.8 km from the study area) and the China Meteorological Administration.

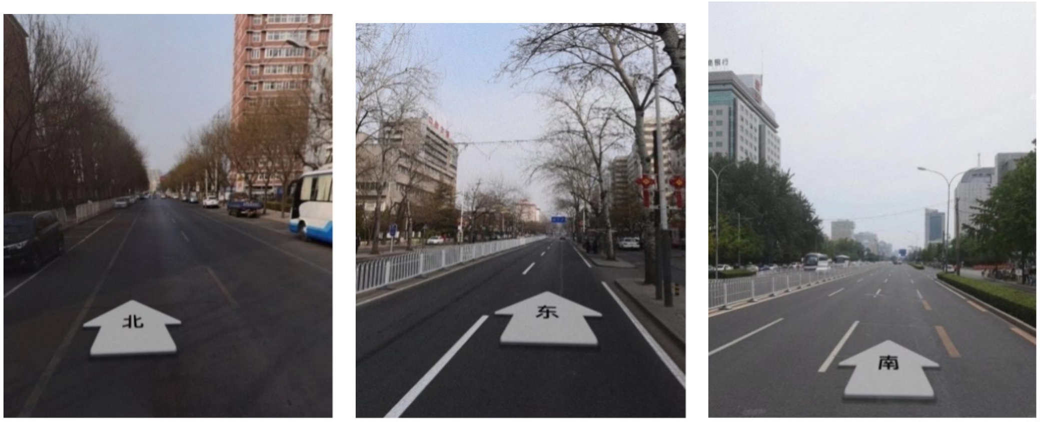


Figure 2: Realistic Images of Different Road Sections: Beijing Jiaotong University East Road (L1), College South Road (L2), and Zhongguancun South Street (L3)

4.4. Calculation of Exposure Dose

The exposure dose for each commuter was calculated using the following equation:

$$D = \int \frac{1}{60} IR \cdot C \, dT \tag{9}$$

where:

- D — Total inhalation dose to humans (µg),
- IR — Breathing volume (L/min),
- C — PM2.5 concentration in the breathing zone ($\mu g/m^3$),
- T — Commuting duration (s).

The raw data were imported into a computer for preprocessing. Outliers were identified and removed if the difference between two consecutive readings exceeded tenfold. The average daily exposure concentration was calculated as the arithmetic mean of all data points for each commuting mode. This method ensures accurate assessments of inhalation particle concentrations based on individual characteristics and activity levels.

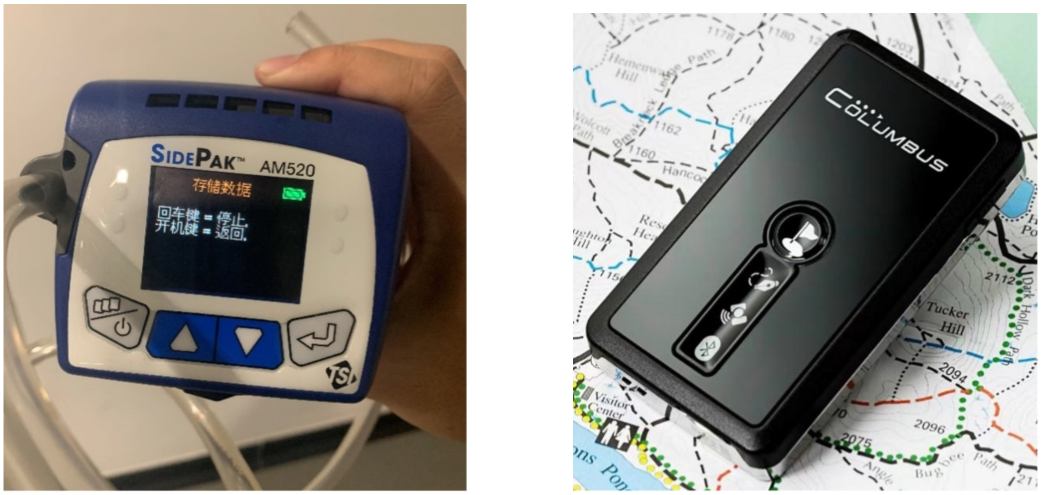


Figure 3: Experimental Instruments: (Left) SidePak™ AM520 Personal Aerosol Monitor, (Right) Explorer V-900 Navigation Recorder

5. Experimental Results

This section presents a comprehensive evaluation of the factors influencing in-vehicle $PM_{2.5}$ concentrations, detailed statistical analyses, and the performance of various predictive models. The discussion is structured into two primary components: the analysis of factors influencing in-vehicle $PM_{2.5}$ and the evaluation of predictive model performance.

5.1. Analysis of Factors Influencing In-Vehicle $PM_{2.5}$

5.1.1. Data Distribution Characteristics

The distribution characteristics of in-vehicle $PM_{2.5}$ concentrations were analyzed using histograms and box plots (Figure 4). The histogram revealed a right-skewed distribution with most $PM_{2.5}$ values ranging between 15 and 60 μ g/m^3, indicating generally low concentrations but occasional high pollution levels. The box plot further indicated a median concentration of approximately 25 μ g/m^3 with several outliers, suggesting elevated pollution under specific conditions.

5.1.2. Correlation Analysis

Correlation analysis identified significant linear relationships between in-vehicle $PM_{2.5}$ and selected variables (Figure 5). External $PM_{2.5}$ exhibited the strongest positive correlation (r=0.65), highlighting its influence on in-vehicle

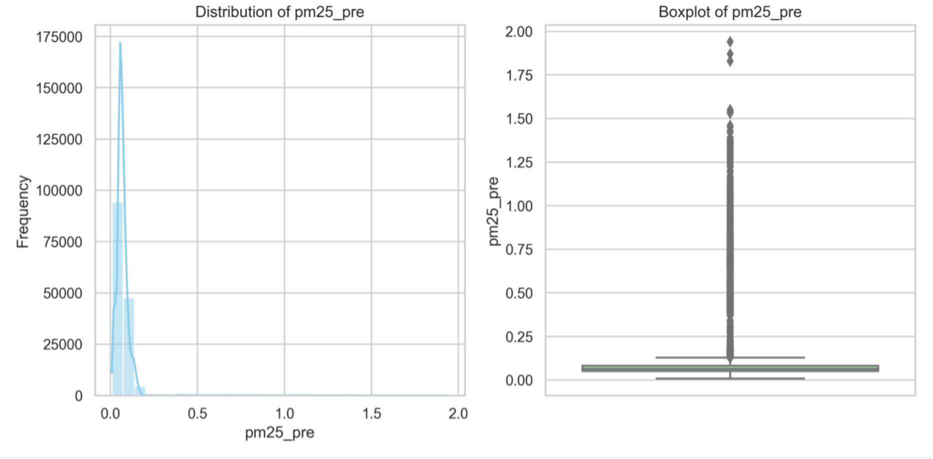


Figure 4: Histogram and box plot of in-vehicle $PM_{2.5}$ concentrations.

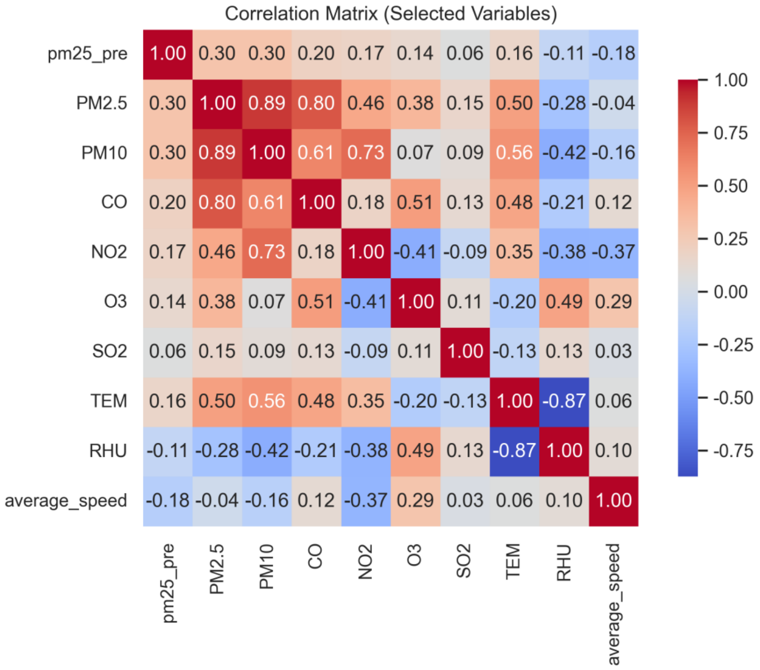
concentrations. AQI also showed a strong positive correlation (r=0.60), while relative humidity (RHU) demonstrated a negative correlation (r=-0.30), suggesting a potential mitigating effect on $PM_{2.5}$ levels. Traffic variables, such as average speed (r=-0.40), indicated that higher vehicle speeds reduced in-vehicle concentrations, likely due to improved ventilation.

5.1.3. Variable Relationships Analysis

Scatter plots with regression lines were employed to visualize relationships between in-vehicle $PM_{2.5}$ and key variables (Figure 6). External $PM_{2.5}$ demonstrated a steep linear trend, emphasizing its dominant influence. Average speed exhibited a negative relationship, while RHU showed a trend of decreasing $PM_{2.5}$ with increasing humidity. These results validated the correlation analysis findings and provided insights into how environmental and traffic factors interact with in-vehicle air quality.

Figure 7 illustrates the scatterplot regression relationships between in-vehicle $PM_{2.5}$ concentration and four key factors (external $PM_{2.5}$, CO, average speed, and temperature). Each subplot includes a scatterplot and a regression line:

- **$PM_{2.5}$ vs. External $PM_{2.5}$**: The scatterplot demonstrates a significant positive correlation, with a steep regression line slope indicating that an increase in external $PM_{2.5}$ concentration leads to a substantial rise in in-vehicle $PM_{2.5}$ levels.



	pm25_pre	PM2.5	PM10	CO	NO2	O3	SO2	TEM	RHU	average_speed
pm25_pre	1.00	0.30	0.30	0.20	0.17	0.14	0.06	0.16	-0.11	-0.18
PM2.5	0.30	1.00	0.89	0.80	0.46	0.38	0.15	0.50	-0.28	-0.04
PM10	0.30	0.89	1.00	0.61	0.73	0.07	0.09	0.56	-0.42	-0.16
CO	0.20	0.80	0.61	1.00	0.18	0.51	0.13	0.48	-0.21	0.12
NO2	0.17	0.46	0.73	0.18	1.00	-0.41	-0.09	0.35	-0.38	-0.37
O3	0.14	0.38	0.07	0.51	-0.41	1.00	0.11	-0.20	0.49	0.29
SO2	0.06	0.15	0.09	0.13	-0.09	0.11	1.00	-0.13	0.13	0.03
TEM	0.16	0.50	0.56	0.48	0.35	-0.20	-0.13	1.00	-0.87	0.06
RHU	-0.11	-0.28	-0.42	-0.21	-0.38	0.49	0.13	-0.87	1.00	0.10
average_speed	-0.18	-0.04	-0.16	0.12	-0.37	0.29	0.03	0.06	0.10	1.00

Figure 5: Correlation matrix of selected variables with in-vehicle $PM_{2.5}$.

- **$PM_{2.5}$ vs. CO**: The scatterplot reveals a clear positive correlation. The regression line slope suggests that higher CO concentrations are accompanied by increased $PM_{2.5}$ levels, implying that both pollutants may originate from the same source.

- **$PM_{2.5}$ vs. Average Speed**: The scatterplot shows a noticeable negative correlation, with the regression line slope indicating that as vehicle speed increases, $PM_{2.5}$ concentration decreases significantly. This trend may be attributed to enhanced air circulation within the vehicle at higher speeds.

- **$PM_{2.5}$ vs. Temperature**: The scatterplot exhibits a weak positive correlation. The regression line suggests that temperature has a relatively small impact on $PM_{2.5}$ concentration, although a slight positive relationship is still observed.

These scatterplot regression analyses not only validate the findings from the correlation analysis but also provide a more detailed understanding of

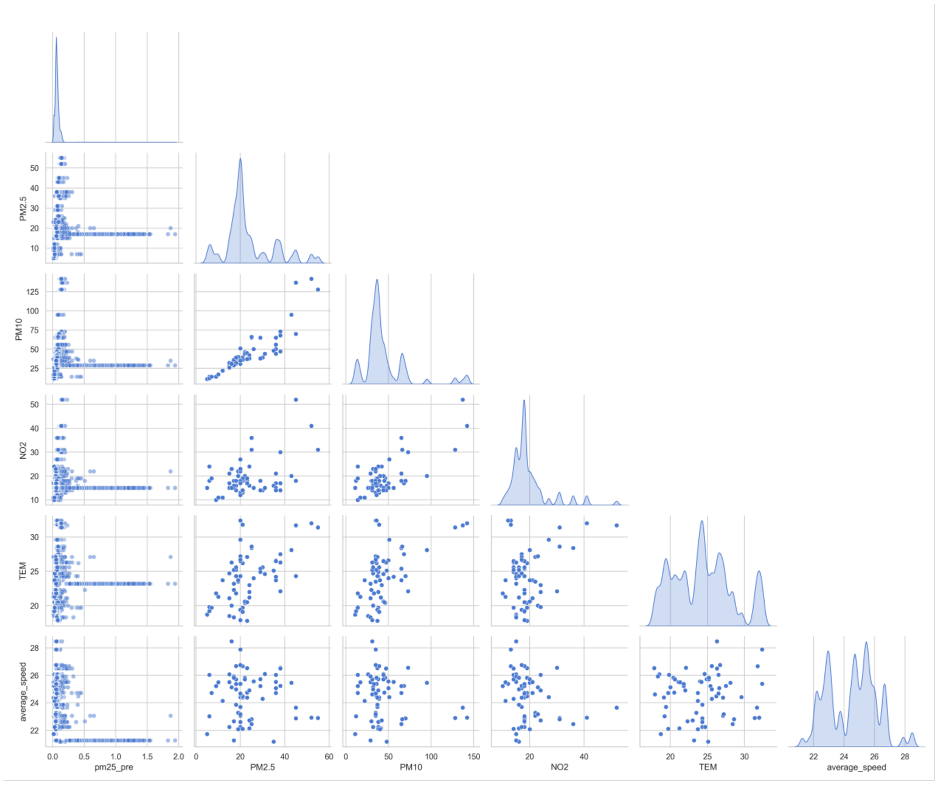


Figure 6: Scatter plots and regression lines for in-vehicle $PM_{2.5}$ vs. key variables.

the relationship patterns between the variables. This contributes to a deeper comprehension of the dynamic mechanisms governing in-vehicle $PM_{2.5}$ concentration.

5.1.4. Comparative Analysis by Category

Box plots (Figure 8) revealed that in-vehicle $PM_{2.5}$ concentrations were significantly lower during nighttime compared to daytime, attributable to reduced traffic and pollution sources. Open-window driving resulted in higher $PM_{2.5}$ levels than closed-window conditions, emphasizing the impact of ventilation modes. Statistical tests confirmed significant differences (p¡0.05) among time periods and driving modes.

5.1.5. Temporal and Spatial Analysis

Time series analysis (Figure 9) illustrated diurnal variations in in-vehicle $PM_{2.5}$, with peaks during morning and evening rush hours, corresponding to

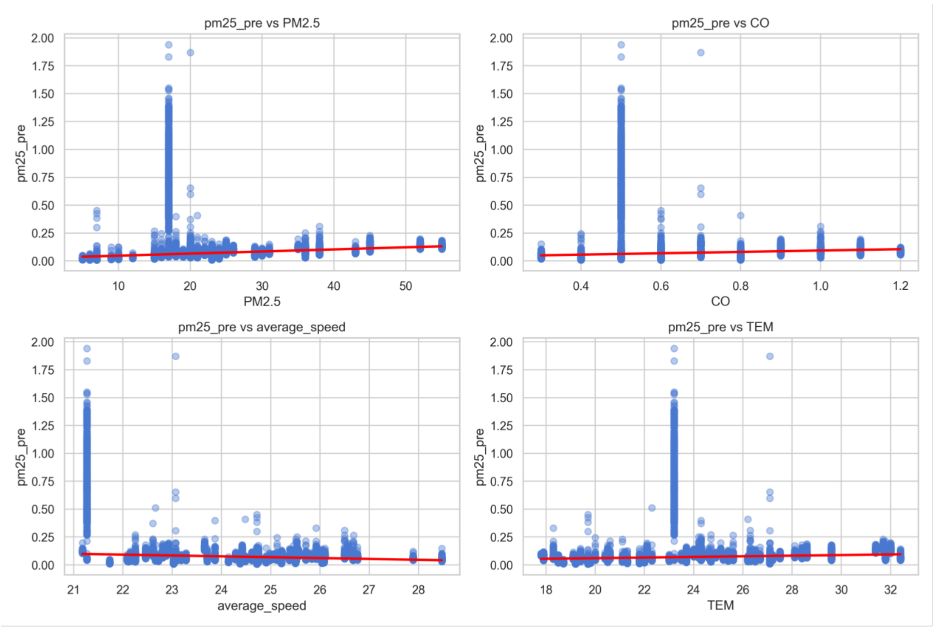


Figure 7: Scatter plots and regression lines for in-vehicle $PM_{2.5}$ vs. key variables.

high traffic volumes and external pollution. Geographic scatter plots identified hotspots with elevated in-vehicle concentrations, aligning with areas of high congestion and proximity to pollution sources, as displayed in Figure 10.

5.2. Model Prediction Performance

5.2.1. Meta-Learning Network Selection

To construct an efficient and generalizable predictive model for in-vehicle $PM_{2.5}$ concentrations, a systematic evaluation of several deep learning models was conducted. The models evaluated included Recurrent Neural Network (RNN), Long Short-Term Memory Network (LSTM), Convolutional Neural Network (CNN), General Regression Neural Network (GRNN), Attention-based Regression Model (Attention Regressor), and Deep Belief Network (DBN). Each model demonstrated unique structural advantages, contributing to their ability to capture different aspects of the prediction task.

The evaluation employed four key performance metrics to comprehensively assess the models' predictive capabilities:

- **Mean Squared Error (MSE):** MSE measures the average squared

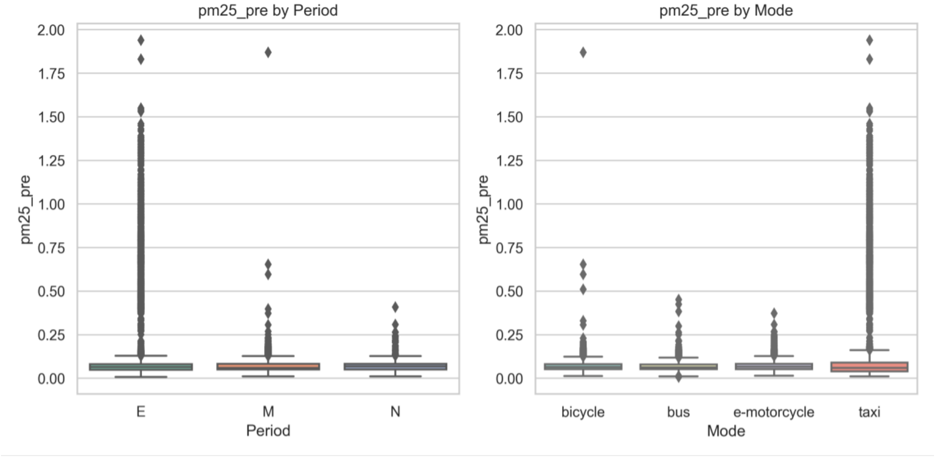


Figure 8: Box plots comparing in-vehicle $PM_{2.5}$ across time periods and driving modes.

difference between predicted and actual values. It is calculated as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where y_i and $\hat{y}_i$ are the actual and predicted values, respectively, and n is the number of samples. Lower MSE indicates higher prediction accuracy.

- **Mean Absolute Error (MAE):** MAE quantifies the average absolute error between predictions and actual values. It is expressed as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} \mid y_i - \hat{y}_i \mid$$

MAE provides an intuitive measure of the prediction error magnitude, with smaller values indicating better performance.

- **Root Mean Squared Error (RMSE):** RMSE is the square root of MSE and shares the same units as the original data. It is calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

RMSE emphasizes larger errors, making it sensitive to outliers.

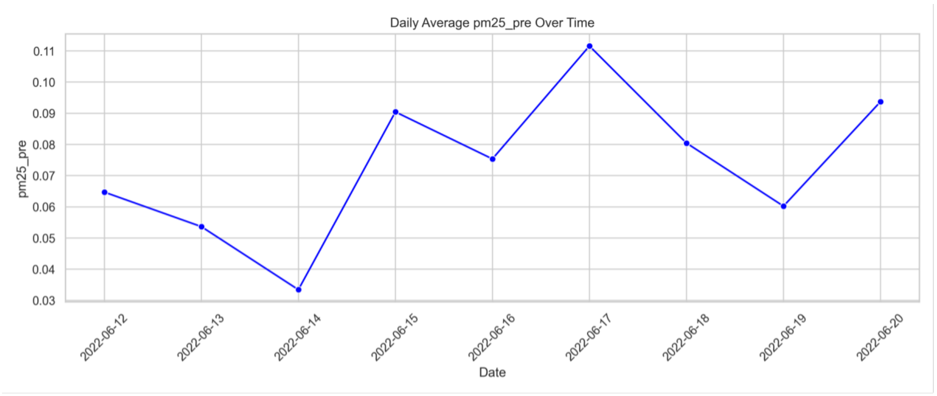


Figure 9: Time series of daily average in-vehicle $PM_{2.5}$ concentrations.

- **Coefficient of Determination (R^2):** R^2 represents the proportion of variance in the dependent variable explained by the model. It is given by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

  where $\bar{y}$ is the mean of actual values. Higher R^2 values, closer to 1, indicate stronger explanatory power.

5.2.2. Model Performance Evaluation

All models were trained and tested on identical datasets to ensure fair comparisons. While the models achieved similar performance in terms of MSE, significant differences were observed in MAE, RMSE, and R^2 scores. GRNN outperformed other models, demonstrating superior performance across MAE, RMSE, and R^2, indicating its strong capability to capture non-linear variations in in-vehicle $PM_{2.5}$ concentrations. CNN and Attention Regressor also exhibited high prediction accuracy, particularly in R^2, achieving 0.9002 and 0.9046, respectively, reflecting their effectiveness in modeling $PM_{2.5}$ variations.

In contrast, traditional RNN and LSTM models exhibited moderate performance, with average MAE and RMSE scores, and relatively lower R^2 values. This indicates limitations in capturing the complexity of the dataset, which may be attributed to the high dimensionality and diversity of in-vehicle environmental features. DBN, despite its competitive MAE, underperformed

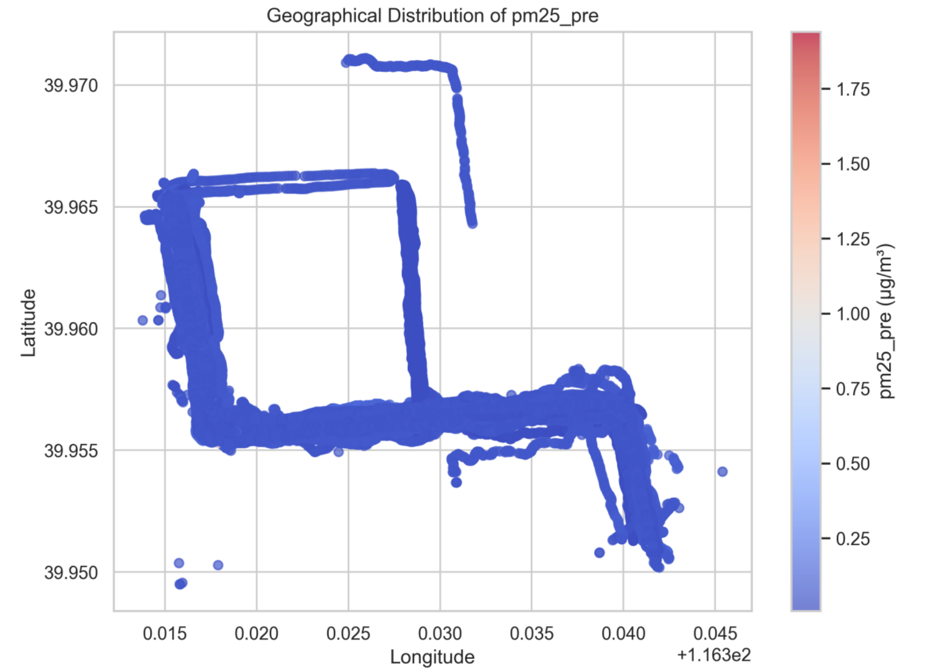


Figure 10: Time series of daily average in-vehicle $PM_{2.5}$ concentrations.

in RMSE and R^2 metrics, highlighting its challenges in handling heterogeneous data.

5.2.3. Base Learner Selection for Ensemble Learning

Based on these evaluations, GRNN, CNN, and Attention Regressor were selected as the base learners for the ensemble learning framework. These models demonstrated complementary strengths:

- **GRNN:** Leveraging its non-parametric characteristics, GRNN captures subtle local variations, making it suitable for modeling fine-grained fluctuations in $PM_{2.5}$.

- **CNN:** Excelling in spatial feature extraction, CNN effectively identifies patterns among multi-sensor data, such as synchronized changes between temperature and $PM_{2.5}$.

- **Attention Regressor:** By incorporating attention mechanisms, this

Model	MSE	MAE	RMSE	$\mathbf{R^2}$
RNN	0.0001	0.0081	0.0115	0.8482
LSTM	0.0001	0.0081	0.0115	0.8482
CNN	0.0001	0.0070	0.0093	0.9002
GRNN	0.0001	0.0057	0.0076	0.9343
Attention Regressor	0.0001	0.0067	0.0091	0.9046
DBN	0.0001	0.0067	0.0097	0.8914
Ensemble Model	0.0002	0.0052	0.0130	0.9622

Table 2: Performance comparison of individual and ensemble models.

Metric	Value
Mean Squared Error (MSE)	0.0002
Mean Absolute Error (MAE)	0.0052
Root Mean Squared Error (RMSE)	0.0130
R^2 Score	0.9622

Table 3: Performance metrics of the ensemble model.

model dynamically adjusts its focus on key features, enhancing prediction accuracy under complex conditions.

5.2.4. Dynamic Ensemble Strategy

The ensemble strategy utilizes fuzzy clustering to partition input data into subspaces, dynamically adjusting model weights based on their performance within each cluster. Initially, Fuzzy C-Means (FCM) clustering assigns membership degrees to each sample for multiple clusters. Weighted Mean Squared Error (WMSE) is then used to calculate model weights for each subspace. The final ensemble prediction is obtained by combining predictions from all models, weighted by both the subspace membership degrees and the model-specific weights.

This dynamic strategy effectively leverages the soft clustering capabilities of FCM, enabling adaptive model fusion tailored to varying environmental and operational conditions. Unlike static weighting or simple averaging, dynamic weighting adapts to data-specific characteristics, enhancing overall prediction performance.

5.2.5. Ensemble Learning Model Evaluation

After selecting the base learners, this study developed an ensemble learning model based on fuzzy clustering and dynamic weighting strategies. The ensemble model integrates the strengths of General Regression Neural Net-

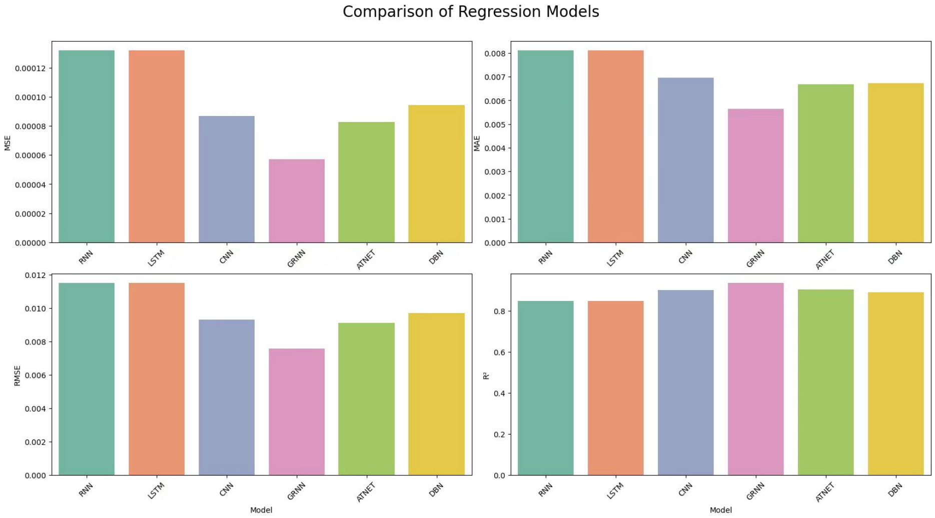


Figure 11: Comparison of individual models.

work (GRNN), Convolutional Neural Network (CNN), and Attention Regressor to enhance overall prediction accuracy and robustness.

The experimental results demonstrate that the ensemble model significantly outperformed individual models across all evaluation metrics. Specifically, the ensemble model achieved a Mean Squared Error (MSE) of 0.0002, Mean Absolute Error (MAE) of 0.0052, Root Mean Squared Error (RMSE) of 0.0130, and an R^2 score of 0.9622. These results highlight the ensemble model's superior accuracy and reliability in predicting in-vehicle $PM_{2.5}$ concentrations.

Compared to individual models, the ensemble model exhibited substantially lower error metrics and a markedly improved R^2 score, indicating stronger explanatory power for data variability. This improvement is primarily attributed to the effective combination of fuzzy clustering and dynamic weighting strategies. Fuzzy clustering partitions input samples into multiple subspaces, ensuring greater similarity in environmental conditions or operational states within each subspace. Dynamic weighting adjusts the contributions of base learners based on their prediction errors in each subspace, ensuring optimal performance under varying conditions.

GRNN: GRNN excels at capturing local nonlinear relationships, accurately reflecting subtle fluctuations in in-vehicle $PM_{2.5}$ concentrations.

CNN: CNN effectively extracts spatial features from multi-sensor data,

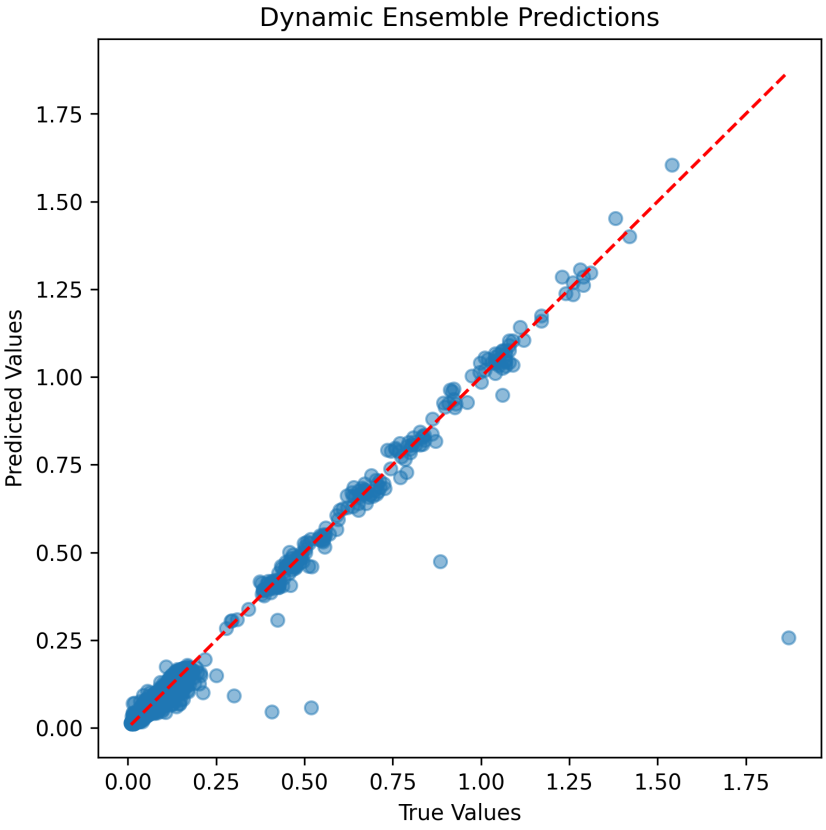


Figure 12: Dynamic Ensemble Learning Predictions.

identifying synchronized patterns such as the interplay between temperature and $PM_{2.5}$ concentrations.

Attention Regressor: By incorporating attention mechanisms, the Attention Regressor dynamically focuses on critical features and key moments, enhancing its adaptability to complex environmental variations.

The ensemble model successfully combines the advantages of these three base learners, significantly improving overall predictive performance. The high R^2 value of 0.9622 indicates an excellent fit to the actual variations in $PM_{2.5}$ concentrations, effectively capturing the main trends and fluctuations in the data. Additionally, the low MSE and MAE further demonstrate the ensemble model's outstanding ability to minimize prediction errors, validating its reliability and practicality in real-world applications.

6. Conclusion

This study presents a novel framework for predicting in-vehicle $PM_{2.5}$ concentrations, addressing the complexities and importance of ensuring air quality within vehicles. By integrating fuzzy clustering with dynamically weighted neural networks, the proposed approach achieves significant improvements in prediction accuracy and robustness.

Leveraging a comprehensive dataset that encompassed temporal, geographical, meteorological, and traffic-related variables, as well as atmospheric pollutants, the research highlights key factors influencing in-vehicle $PM_{2.5}$ levels. Statistical analyses identified external $PM_{2.5}$ concentrations (r = 0.65), Air Quality Index (AQI, r = 0.60), relative humidity (r = -0.30), average speed (r = -0.40), and stops (r = 0.35) as significant determinants. These insights informed the selection of features and optimization of predictive models, addressing the nonlinear and multidimensional nature of the task.

The predictive framework incorporates three base learners—General Regression Neural Networks (GRNN), Convolutional Neural Networks (CNN), and Attention Regressor—each selected for its unique strengths. GRNN captured fine-grained nonlinear relationships, CNN effectively extracted spatial features, and the Attention Regressor dynamically focused on key temporal and feature-based variations. These models were integrated using a dynamic weighting mechanism informed by fuzzy clustering, enabling adaptive model performance across various subspaces of input data.

Experimental results demonstrated the effectiveness of the proposed ensemble framework, achieving a Mean Squared Error (MSE) of 0.0002, Mean Absolute Error (MAE) of 0.0052, Root Mean Squared Error (RMSE) of 0.0130, and a coefficient of determination (R^2) of 0.9622. These metrics represent a significant improvement over individual models, highlighting the model's capability to capture both global trends and local variations in in-vehicle $PM_{2.5}$ concentrations.

Furthermore, the results reveal critical findings regarding in-vehicle air quality. External $PM_{2.5}$ was identified as the dominant factor influencing in-vehicle concentrations, with high exposure levels observed during peak traffic hours and under open-window driving conditions. The framework's superior predictive performance underscores its practical applicability for urban air quality monitoring, real-time risk assessment, and health protection strategies.

References

[1] Adams, H.S., Nieuwenhuijsen, M.J., Colvile, R.N., McMullen, M.A.S., Khandelwal, P., 2001. Fine particle (pm2.5) personal exposure levels in transport microenvironments, london, uk. Science of the Total Environment 279, 29–44.

[2] An, X., Cao, Y., Wang, Q., Fu, J., Wang, C., Jing, K., Liu, B., 2022. Analysis of pollution characteristics and sources of pm2.5 components in urban areas of beijing. Environmental Science 43, 2251–2261.

[3] Bezdek, J.C., Ehrlich, R., Full, W., 1984. Fcm: The fuzzy c-means clustering algorithm. Computers & geosciences 10, 191–203.

[4] Bu, X., Xie, Z., Liu, J., Wei, L., Wang, X., Chen, M., Ren, H., 2021. Global pm2.5-attributable health burden from 1990 to 2017: Estimates from the global burden of disease study 2017. Environmental Research 197, 111123.

[5] Dietterich, T.G., 2000. Ensemble methods in machine learning , 1–15.

[6] Kumar, P., Patton, A.P., Durant, J.L., Frey, H.C., 2018. A review of factors impacting exposure to pm2. 5, ultrafine particles and black carbon in asian transport microenvironments. Atmospheric Environment 187, 301–316.

[7] Liu, K., Lin, X., Xu, J., Ma, F., Yang, W., Cao, R., Hu, X., Wei, Y., Jiang, L., Wang, Z., 2024. Investigating the influence of platform design on the distribution of traffic particulate matter at the bus stop. Building and Environment 255, 111395.

[8] Liu, W., Chen, R., 2022. A hybrid cnn-gru model for spatiotemporal prediction of pm2.5. Environmental Pollution 302, 119368.

[9] Luengo-Oroz, J., Reis, S., 2019. Assessment of cyclists' exposure to ultrafine particles along alternative commuting routes in edinburgh. Atmospheric Pollution Research 10, 1148–1158.

[10] Maji, K.J., Namdeo, A., Hoban, D., Bell, M., Goodman, P., Nagendra, S.S., Barnes, J., De Vito, L., Hayes, E., Longhurst, J., et al., 2021. Analysis of various transport modes to evaluate personal exposure to pm2. 5 pollution in delhi. Atmospheric Pollution Research 12, 417–431.

[11] Meng, Z., Li, Y., Chen, X., 2021. A hybrid ensemble model for predicting atmospheric pollutants. Environmental Modelling & Software 139, 104998.

[12] Mu, G., Wang, B., Cheng, M., He, X., Liu, W., Zhang, X., 2022. Long-term personal pm2.5 exposure and lung function alternation: A longitudinal study in wuhan urban adults. Science of The Total Environment 845, 157327.

[13] Ni, Y., Wu, Z., Xu, Y., 2020. Deep ensemble learning for spatiotemporal air pollution prediction. IEEE Access 8, 123456–123468.

[14] O'Shea, K., 2015. An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458 .

[15] Specht, D.F., et al., 1991. A general regression neural network. IEEE transactions on neural networks 2, 568–576.

[16] Webb, G.I., Zheng, Z., 2004. Multistrategy ensemble learning: Reducing error by combining ensemble learning techniques. IEEE Transactions on Knowledge and Data Engineering 16, 980–991.

[17] Wu, Z., Huang, N.E., 2009. Ensemble empirical mode decomposition: a noise-assisted data analysis method. Advances in adaptive data analysis 1, 1–41.

[18] Yang, F., Zhao, M., 2023. Ensemble empirical mode decomposition and grnn-based hybrid model for pm2.5 prediction. Environmental Modelling & Software 159, 105453.

[19] Yoo, D., Park, S., Lee, J.Y., Paek, A.S., So Kweon, I., 2015. Attentionnet: Aggregating weak directions for accurate object detection, in: Proceedings of the IEEE international conference on computer vision, pp. 2659–2667.

[20] Załuska, M., Gładyszewska-Fiedoruk, K., 2020. Regression model of pm2. 5 concentration in a single-family house. Sustainability 12, 5952.

[21] Zhang, H., 2020. Characteristics of air pollution and analysis of particulate matter sources in xi'an. Xi'an University of Architecture and Technology DOI: 10.27393/d.cnki.gxazu.2020.001128.

[22] Zhang, L., Huang, W., 2021. Urban pm2.5 concentration prediction via attention-based cnn-lstm. Applied Sciences 11, 3567.

[23] Zhou, Z.H., 2012. Ensemble methods: Foundations and algorithms. Chapman and Hall/CRC.