UNIVERSITATEA NAȚIONALĂ DE ȘTIINȚĂ ȘI TEHNOLOGIE
POLITEHNICA BUCUREȘTI
FACULTATEA DE AUTOMATICĂ ȘI CALCULATOARE
DEPARTAMENTUL DE CALCULATOARE

PROIECT DE DIPLOMĂ

Predicția Calității Aerului

Alexandru Axenia

Coordonator științific:

Conf. Dr. Ing. Laura Ruse
Conf. Dr. Ing. Dan Tudose
Ing. Abhinuv Pitale

BUCUREȘTI
2025

NATIONAL UNIVERSITY OF SCIENCE AND TECHNOLOGY
POLITEHNICA BUCHAREST
FACULTY OF AUTOMATIC CONTROL AND COMPUTERS
COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

DIPLOMA PROJECT

Air Quality Prediction

Alexandru Axenia

Thesis advisor:

Conf. Dr. Ing. Laura Ruse
Conf. Dr. Ing. Dan Tudose
Ing. Abhinuv Pitale

BUCHAREST
2025

CONTENTS

SINOPSIS

Această lucrare propune și implementează un sistem de predicție a calității aerului utilizând tehnici moderne de învățare automată, aplicate pe date istorice colectate de la stații de monitorizare. Studiul se concentrează pe estimarea concentrațiilor zilnice pentru poluanții PM1, PM2.5 și PM10, folosind atât rețele neuronale Long Short-Term Memory (LSTM) specializate în prelucrarea seriilor de timp, cât și modele de regresie clasică și de tip ensemble. Pentru a asigura o evaluare robustă a performanțelor, a fost utilizată o strategie de testare cu fereastră glisantă (rolling window) pe mai multe subseturi temporale, iar acuratețea modelelor a fost măsurată prin Mean Absolute Error (MAE). Rezultatele au indicat că modelul LSTM a obținut cele mai bune performanțe în captarea tendințelor de evoluție a poluanților, în timp ce alte modele, deși competitive ca scor numeric, au prezentat limitări în surprinderea variațiilor rapide. În plus, a fost dezvoltat un modul interactiv care permite selecția stației de monitorizare și a modelului de prognoză, facilitând compararea directă a scenariilor de predicție. Lucrarea evidențiază importanța utilizării metodelor secvențiale în analiza datelor de mediu și propune direcții de îmbunătățire viitoare pentru mărirea acurateței de predicție.

ABSTRACT

This paper proposes and implements an air quality forecasting system using modern machine learning techniques applied to historical monitoring data. The study focuses on estimating daily concentrations of PM1, PM2.5, and PM10 pollutants by employing both Long Short-Term Memory (LSTM) neural networks, specialized for time series processing, and classical as well as ensemble regression models. To ensure a robust performance assessment, a rolling window testing strategy was applied across multiple temporal subsets, and model accuracy was evaluated using the Mean Absolute Error (MAE) metric. Results indicated that the LSTM model consistently achieved the best performance in capturing pollutant trends, while other models, although competitive in numerical scores, showed limitations in tracking rapid variations. Additionally, an interactive module was developed, allowing users to select the monitoring station and forecasting model, facilitating direct comparisons of prediction scenarios. This work highlights the importance of sequential methods in environmental data analysis and suggests potential future improvements for enhanced forecasting accuracy.

1 INTRODUCTION

1.1 Background and Motivation

Air pollution represents one of the most pressing environmental and public health challenges of the modern era. The increasing concentration of airborne particulate matter, particularly PM1, PM2.5, and PM10, has been directly linked to respiratory illnesses, cardiovascular diseases, and reduced life expectancy [1]. Reliable air quality monitoring and forecasting systems are essential tools for public authorities and communities, enabling timely preventive actions and informed decision-making [2].

In recent years, the availability of open source environmental datasets and advancements in machine learning have created new opportunities for developing predictive models capable of estimating future pollutant levels [3]. Unlike traditional statistical models, machine learning approaches - especially deep learning architectures like Long Short-Term Memory (LSTM) networks - can capture complex, non-linear patterns in time series data, improving the accuracy of air pollution forecasts [4]. This project was motivated by the need to investigate and compare various machine learning methods for air quality prediction, while also enhancing the accessibility of such models through interactive tools for practical use.

1.2 Problem Statement

Despite the critical importance of air pollution forecasting, many existing models suffer from limited accuracy, poor generalizability across different monitoring stations, and an inability to capture short-term fluctuations in pollutant concentrations. Traditional error metrics, such as Mean Absolute Error (MAE) or Mean Squared Error (MSE), often fail to reflect a model's effectiveness in tracking environmental trends [5]. Additionally, most implementations lack user-friendly tools for interactive scenario testing and model comparison.

This study addresses these challenges by designing, implementing, and evaluating multiple machine learning models - including LSTM, Linear Regression, Random Forest Regressor and Support Vector Regression - for multi-station, multi-pollutant air quality prediction. Furthermore, it introduces an interactive application module, allowing users to select both the station and the forecasting model, providing a practical tool for real-time analysis.

1.3 Objectives of the Project

The primary objectives of this thesis are:

- To develop and evaluate several machine learning models for predicting daily PM1, PM2.5, and PM10 concentrations based on historical data from multiple monitoring stations.
- To compare model performances using appropriate numerical metrics, such as MAE, and through graphical trend analysis.
- To implement a rolling window validation strategy to assess model stability and generalizability over time.
- To create an interactive application interface that allows users to select monitoring stations and prediction models, generating 30-day forecasts for comparative analysis.
- To identify the limitations of existing evaluation approaches and propose potential improvements for future research.

1.4 Structure of the Thesis

The thesis is structured as follows:

- **Chapter 2** reviews the current state of research regarding air quality forecasting methods and the role of machine learning in environmental data analysis.
- **Chapter 3** describes the datasets used, including data sources, pre-processing steps, and descriptive analysis.
- **Chapter 4** outlines the experimental methodology, including the performance metrics.
- **Chapter 5** presents the machine learning models implemented, detailing their architectures, configurations, and training procedures, including the rolling window validation strategy.
- **Chapter 6** discusses the results obtained for each model, with both quantitative error analysis and visual trend comparisons.
- **Chapter 7** analyzes the strengths and limitations of the project, reflecting on the findings and suggesting directions for future improvements.
- **Chapter 8** concludes the thesis by summarizing the main findings and presenting future work directions, including recommendations for model improvement, integration of external factors, development of trend-focused evaluation metrics and the exploration of advanced forecasting architectures.

2 STATE OF THE ART

2.1 Air Pollution and Its Impact

One of the greatest risks to human health and ecosystems is air pollution. The World Health Organization (WHO) reports that exposure to fine particles and dangerous gases causes an estimated 7 million premature deaths per year[1]. It also estimates that 99 percent of the people on the planet breathe air that contains more pollutants than WHO guidelines, including sulfur dioxide (SO2), nitrogen dioxide (NO2), particle matter (PM) and ozone (O3)[2]. Asthma, chronic obstructive pulmonary disease (COPD), lung cancer, cardiovascular diseases such as ischemic heart disease and stroke, as well as neurological and developmental problems are among the health effects of air pollution [6].

PM1, PM2.5 and PM10 pollutants - which are microscopic particles smaller in diameter than 1, 2.5 and 10 micrometers - are extremely dangerous because they can enter the lungs and even the bloodstream [7].

Air pollution not only directly affects human health, but also harms plants, reduces agricultural production, and deteriorates urban infrastructure. Urban regions are particularly vulnerable due to large concentrations of pollutants from transportation, heating systems, industrial activities, and human density [8].

Due to these complex effects, the ability to monitor and predict air pollution levels has become a critical component of public health management, policymaking, and urban planning. Accurate forecasting systems enable governments and communities to implement timely interventions, issue warnings, and plan mitigation strategies aimed at reducing population exposure and long-term health risks.

2.2 Existing Monitoring and Forecasting Solutions

Air quality is usually monitored by stations that collect air pollution concentrations. Platforms such as the Air Quality Open Data Platform (AQICN) provide air quality data collected from thousands of stations around the world[3]. In addition to monitoring data, various models have been created to predict pollution. Models like CALPUFF and CMAQ [9] simulate pollutant

[1] https://www.who.int/news/item/25-03-2014-7-million-premature-deaths-annually-linked-to-air-pollution

[2] https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

[3] https://aqicn.org

dispersion using emission and meteorological data. For predicting pollution in time, models like ARIMA and SARIMA have been widely used [10]. Today, approaches using machine learning have demonstrated better performance than classical models.

Numerous more methods have been created to measure and forecast air pollution in real time, in addition to platforms like AQICN and ground-based monitoring stations. Networks like the United States Environmental Protection Agency's AirNow system, which compiles air quality data from more than 4,000 sites across North America and offers public forecasts for important pollutants like PM2.5 and ozone, are frequently relied upon by national and regional governments[4].

On a bigger scale, remote sensing technologies are also crucial for air quality monitoring. Aerosol optical depth (AOD) and air pollutants concentrations such as nitrogen dioxide (NO2) and ozone (O3) can be measured using satellite-based instruments such as NASA's MODIS (Moderate Resolution Imaging Spectroradiometer) and ESA's Sentinel-5P [11]. In particular, in places with limited monitoring networks, these remote observations are being used more and more to supplement data from ground stations.

On the forecasting side, in addition to classical dispersion models like CALPUFF and CMAQ, novel machine learning and data-driven approaches have evolved. To generate daily predictions of air pollution at the global and regional levels, systems like the Copernicus Atmosphere Monitoring Service (CAMS) use chemical transport models, satellite data, and ground observations[5].

These integrated systems and methodologies (Table 1) offer comprehensive tools for managing air pollution and safeguarding public health, which is a major improvement over data from individual stations.

Table 1: Overview of data sources used for air quality monitoring

Data Source	Type	Spatial Coverage	Variables Monitored
AQICN	Ground stations	Global	PM1, PM2.5, PM10, NO2
AirNow	Ground stations	North America	PM2.5, O3
Copernicus CAMS	Ground + satellite	Global	PM2.5, NO2, O3, SO2
Sentinel-5P	Satellite	Global	NO2, SO2, CO, O3, AOD
MODIS	Satellite	Global	Aerosol Optical Depth

2.3 Deep Learning in Time Series Forecasting

Deep learning has emerged as a highly favored and efficient method for forecasting time series, especially in domains where data trends are intricate, nonlinear, and affected by various interacting elements. In contrast to conventional statistical models like ARIMA and SARIMA,

[4]https://www.airnow.gov

[5]https://atmosphere.copernicus.eu

deep learning models can autonomously discover complex temporal relationships and layered feature representations directly from unprocessed sequential data [12].

Recurrent Neural Networks (RNNs) are widely utilized within deep learning frameworks for modeling sequences that depend on time. Nevertheless, standard RNNs struggle with capturing long-range dependencies due to challenges like vanishing and exploding gradients. To resolve these issues, Long Short-Term Memory (LSTM) networks were developed. LSTM units (Figure 1) feature gating mechanisms (input, output, and forget gates) that manage the flow of information and sustain a long-term memory state, allowing the network to retain significant patterns across lengthy sequences [13].

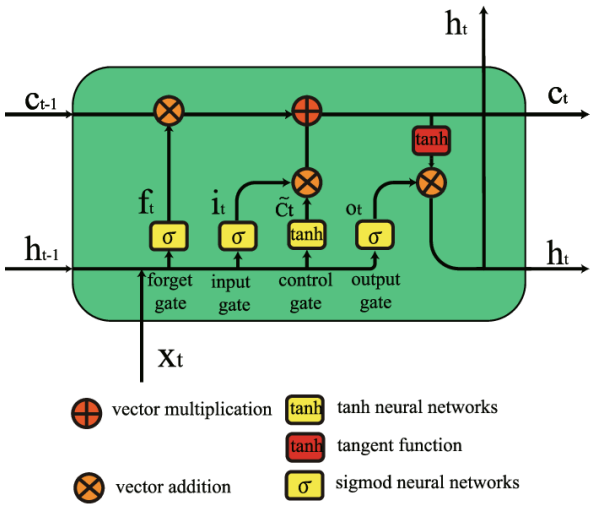

Figure 1: Structure of a typical LSTM cell[6]

In air pollution forecasting, LSTM networks have shown notable success in predicting concentrations of PM2.5, PM10, and NO2 based on historical pollution levels and external variables such as meteorological data. One of the advantages of LSTM is its flexibility to integrate heterogeneous data sources - including temperature, humidity, wind speed, and traffic density - which can significantly improve forecast accuracy.

More recent developments, such as Bidirectional LSTM (BiLSTM) and Attention-based LSTM (AttLSTM) models, have further enhanced prediction performance by enabling the network to access both past and future contexts within a sequence or by prioritizing important timesteps during training. These improvements have established deep learning as a state-of-the-art solution for sequential data analysis, particularly in environmental and public health domains.

[6] https://www.researchgate.net/figure/LSTM-cell-architecture_fig5_352658938

2.4 Comparative Analysis of Machine Learning Models for Pollution Prediction

In recent years, several studies have compared different machine learning models for the prediction of air pollution. Algorithms such as Linear Regression (LR), Support Vector Regression (SVR), Random Forest Regression (RFR), and LSTM have been tested on multiple datasets [14]. The results show that each model has its own advantages and limitations (Table 2). Linear models are easy to interpret but face difficulties in handling complex patterns, SVR handles small and medium datasets well, Random Forest is excellent for capturing interactions between features, and LSTM usually performs best in long-term forecasts, although it requires careful tuning and more resources. This project contributes to this growing research by comparing these models for daily PM1, PM2.5, and PM10 predictions, using real-world data collected from AQICN stations.

Table 2: Comparative analysis of machine learning models for air pollution prediction

Model	Advantages	Limitations
LR	Simple, fast to train, and highly interpretable. Suitable for problems with linear relationships.	Poor performance on complex or nonlinear data. Sensitive to outliers.
SVR	Good performance on small to medium datasets. Robust to outliers. Can model non-linear relationships using kernel functions.	Sensitive to the choice of kernel and parameters. Computationally expensive on large datasets.
RFR	Handles non-linearity and complex relationships well. Robust to noise and outliers. Reduces overfitting compared to single decision trees.	Less interpretable. Requires careful tuning of the number of trees and depth. Can be slow when using many trees.
LSTM	Capable of capturing long-term dependencies and temporal patterns in sequential data. Mitigates vanishing gradient issues in time series forecasting.	Requires large amounts of data. Computationally intensive. Difficult to tune and interpret.

2.5 Recent Trends in Air Pollution Forecasting

In recent years, air pollution forecasting has rapidly evolved through the integration of advanced machine learning, deep learning, and remote sensing technologies. Several trends have emerged that aim to improve the accuracy of prediction, spatial coverage, and model interpretability.

Multisource Data Fusion: One of the most significant developments involves combining heterogeneous data sources such as ground-based monitoring stations, satellite observations,

mobile sensors, and meteorological forecasts. Data fusion techniques allow models to capture both local pollutant variations and regional transport patterns [15]. For example, satellite-derived Aerosol Optical Depth (AOD) can complement PM2.5 measurements in areas with sparse station coverage, while traffic and industrial emission data enhance urban-scale predictions [16], [17].

Transfer Learning and Ensemble Methods: Transfer learning has started to gain traction in environmental modeling by enabling pre-trained models on large urban datasets to be fine-tuned for smaller or rural areas with limited historical data [18]. In parallel, ensemble learning methods - combining predictions from multiple algorithms - have proven effective in reducing individual model biases and improving generalization capabilities. Techniques such as stacking or boosting have been applied to air quality time series with encouraging results [19].

3 DATA COLLECTION AND PREPROCESSING

3.1 Data Acquisition

For this study, air pollution data was collected from AQICN[1]. This platform aggregates real-time and historical air quality data from thousands of monitoring stations around the world. To ensure the reliability and consistency of the forecasts, a careful selection of monitoring stations was made.

The selection criteria included the following.

- Availability of at least two consecutive years of daily data for each selected station.
- Minimal number of daily observations missing to reduce the amount of imputed data.
- Availability of particulate matter (PM) indicators: PM1, PM2.5 and PM10.

Each pollutant was stored in a separate CSV file for every monitoring station, containing daily median values and timestamps. This structure was chosen to maintain flexibility in the data processing pipeline and to enable isolated model evaluations per pollutant and per station.

3.2 Handling Missing Data and Interpolation Techniques

The obtained data had missing values due to equipment failures, transmission mistakes, or other operational problems, just like the majority of real-world environmental datasets. A missing data imputation approach was used since training sequential models like LSTM requires continuous time series.

By estimating intermediate values between known observations, the linear interpolation approach was used to fill in the missing values. The simplicity, low computational cost and applicability of this approach for datasets with small to moderate gaps in consecutive records led to its selection. Interpolation was applied only when the number of consecutive missing days did not exceed a predefined threshold (e.g. 14 days), to avoid introducing excessive bias.

In addition to the interpolation process, older data that lacked sufficient continuity with the recent observations were excluded. For example, for the station shown in Figure 2, only data starting from February 2023 were used.

[1] https://aqicn.org

[2] `https://aqicn.org/station/romania-bucharest-aleea-politehnicii/ro/#/z/15`

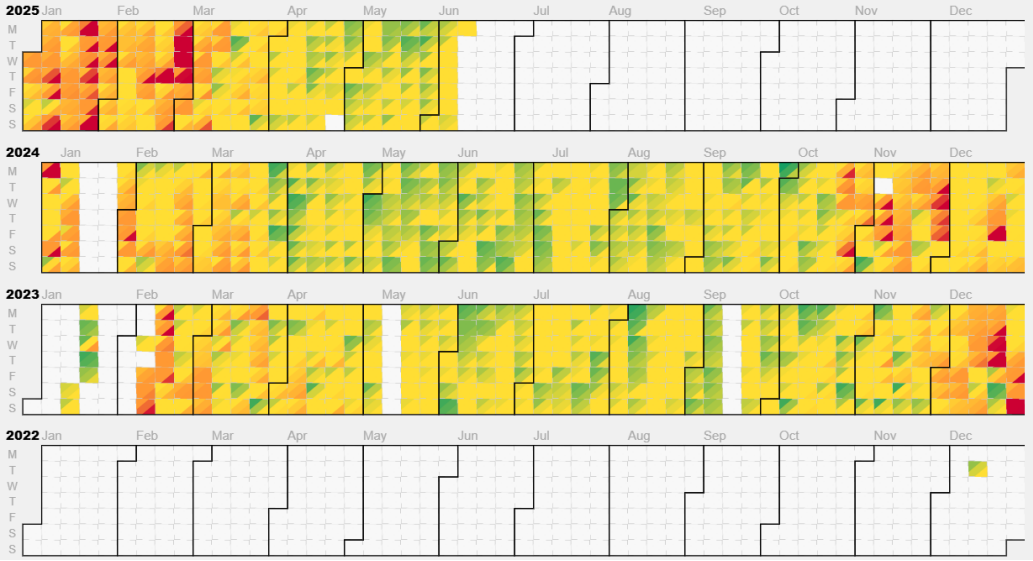

Figure 2: Visual representation of PM2.5 data from a station[2]

3.3 Data Preprocessing and Correlation Analysis

The dataset used in this study was collected from four air quality monitoring stations, each providing separate CSV files containing daily median values for PM1, PM2.5 and PM10 concentrations. The preprocessing workflow involved several steps designed to clean, align, and prepare the data for modeling.

- **Loading and Merging**: For each station, three CSV files corresponding to PM1, PM2.5, and PM10 were loaded and merged based on the common date column.
- **Handling Missing Data**: Since the monitoring period was not continuous for all pollutants and days, missing values were addressed by generating a complete date range between the earliest and latest records, then reindexing the dataframes and performing linear interpolation on missing values.
- **Correlation Analysis**: An exploratory correlation matrix was generated including PM1, PM2.5, PM10, atmospheric pressure, relative humidity, and temperature (Figure 3). The analysis revealed a very strong positive correlation between PM1, PM2.5, and PM10, while the other variables showed weak correlations with the particulate matter concentrations. Consequently, only the three PM variables were selected as input features for the forecasting models, leveraging their mutual correlation structure to potentially improve multivariate time series forecasting performance.

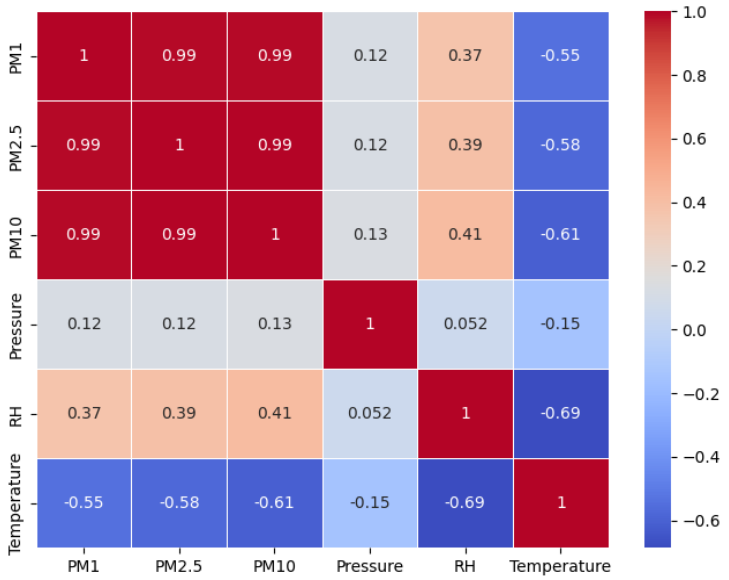

Figure 3: Heatmap of variable correlations

3.4 Data Splitting Strategy for Robust Testing

To ensure a robust and thorough evaluation of the models, the preprocessed dataset for each station was further divided into five progressively reduced subsets. This incremental reduction aimed to assess the model performance not only on the full dataset but also on shorter time series, simulating scenarios with limited historical data availability.

- **Subset Division**: The full dataset for each station was divided into five subsets, where each subsequent subset contained 30 fewer days than the previous one (Table 3). This produced sets of gradually decreasing size, with subset 1 containing the full available data, and subset 5 having the fewest data points.
- **Training and Testing Splits**: For each subset, the data was split into a training set (all but the last 30 days) and a testing set (the final 30 days). A sliding window approach was employed for time series sequence generation, using sequences of 10 consecutive days to predict the following day's values for PM1, PM2.5, and PM10.
- **Visualization of the Experimental Setup**: A schematic diagram illustrating the division of each station's data into the five subsets is presented in Figure 4. This visualization clarifies the incremental reduction strategy and the consistent separation of training and testing periods for each subset.
- **Reasoning for Subset Testing**: This experimental setup was chosen to assess the models' robustness under different data availability scenarios and to observe how the amount of historical data influences forecasting performance. It also enabled the evaluation of the temporal generalization capacity of the models on unseen data.

Table 3: Number of samples in each subset for different datasets

Dataset	Subset 1	Subset 2	Subset 3	Subset 4	Subset 5
Dataset 1	847	817	787	757	727
Dataset 2	798	768	738	708	678
Dataset 3	1200	1170	1140	1110	1080
Dataset 4	1898	1868	1838	1808	1778

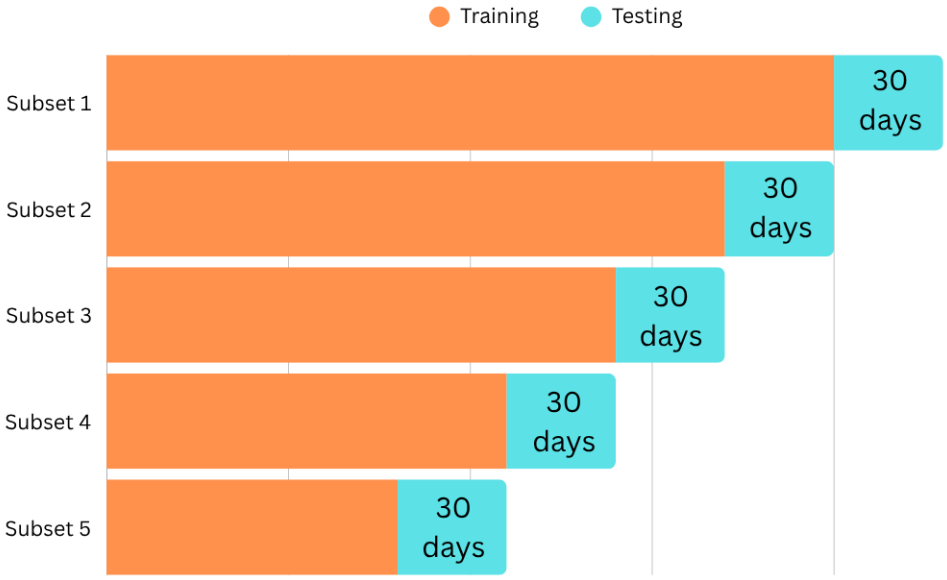

Figure 4: 5-Subset data split diagram

4 METHODOLOGY

This chapter presents the methodological framework adopted in this study to predict air pollution levels using a variety of machine learning models. The approach involves preparing time series data from multiple monitoring stations, training prediction models for different indicators of particulate matter (PM1, PM2.5, PM10), and evaluating their performance using appropriate metrics.

4.1 Long Short-Term Memory (LSTM) Network

LSTM networks are a special type of recurrent neural network (RNN) created to work with large sequences of data. In this study, LSTM models were chosen because they are very good at remembering patterns and connections that happen over longer periods of time. This makes them a great option for predicting air quality, where past pollution levels can influence future values.

LSTM networks were employed to predict air pollution levels for PM1, PM2.5 and PM10. The input to the LSTM consisted of sliding windows of 10 consecutive days of historical pollutant concentrations, used to predict the pollutant values for the next 30 days (Figure 5).

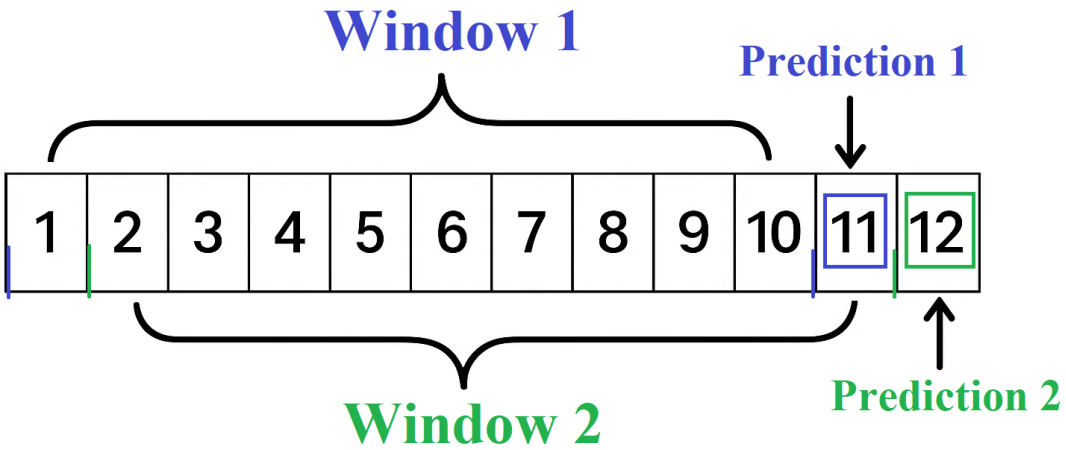

Figure 5: Sequential data windows for input

The decision to use a 10-day input window was based on an empirical analysis of the dataset, where it was observed that pollutant concentration patterns exhibited a relatively weekly trend. By capturing data over a 10-day period, slightly longer than one week, the model could better account for typical weekly fluctuations and possible anomalies, such as weekend traffic effects or specific weather patterns influencing pollutant dispersion.

For model evaluation, predictions were generated on the last 30 days of each data subset for every monitoring station and pollutant type. This testing strategy ensured that the model was consistently evaluated on recent, unseen data while preserving the chronological integrity of the time series. The implementation used a rolling window approach, incrementally moving the training window forward after each prediction to cover the entire test set.

The architecture of the LSTM model consisted of one LSTM layer followed by a Dense output layer with 30 neurons, corresponding to the 30-day prediction horizon.

4.2 Additional Prediction Models

To benchmark the performance of the LSTM network, three classical machine learning models were implemented and evaluated under identical conditions: Linear Regression, Random Forest Regressor, and Support Vector Regression. These models were selected for their distinct characteristics and frequent use in time series and environmental prediction problems.

4.2.1 Linear Regression

Linear Regression (Figure 6) is a basic yet widely used method that shows how one value depends on one or more other values by fitting a straight line through the data points. In this project, it was applied to the training data arranged in sequences, where the model used information from the past 10 days to predict the pollution level for the next day, following the same approach as the LSTM model.

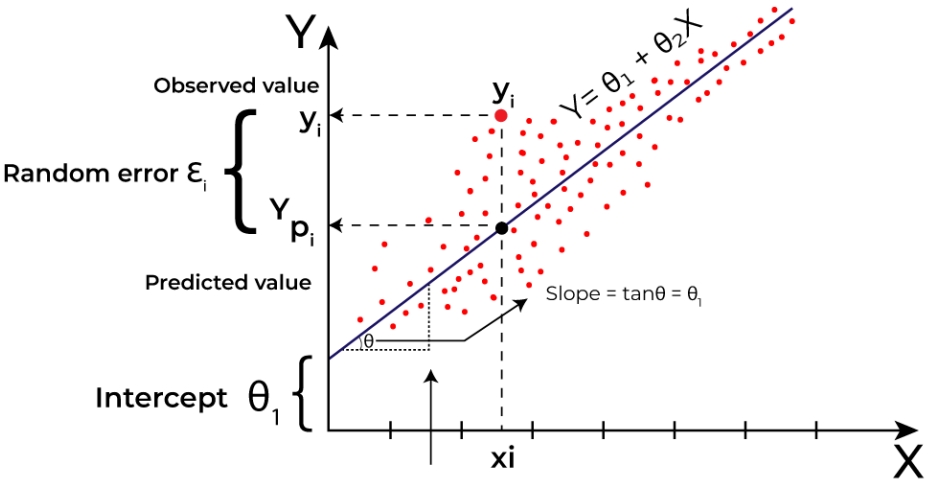

Figure 6: Linear regression model visualization[1]

[1] https://www.geeksforgeeks.org/machine-learning/ml-linear-regression

4.2.2 Random Forest Regressor

Random Forest (Figure 7) is a machine learning method that creates many decision trees and then combines their results to make a final prediction. By averaging the predictions from all the trees, it improves accuracy and reduces the risk of overfitting [20]. Because Random Forest can detect complex, non-linear connections between data points, it serves as a valuable model to compare against the LSTM when predicting air pollution levels [21].

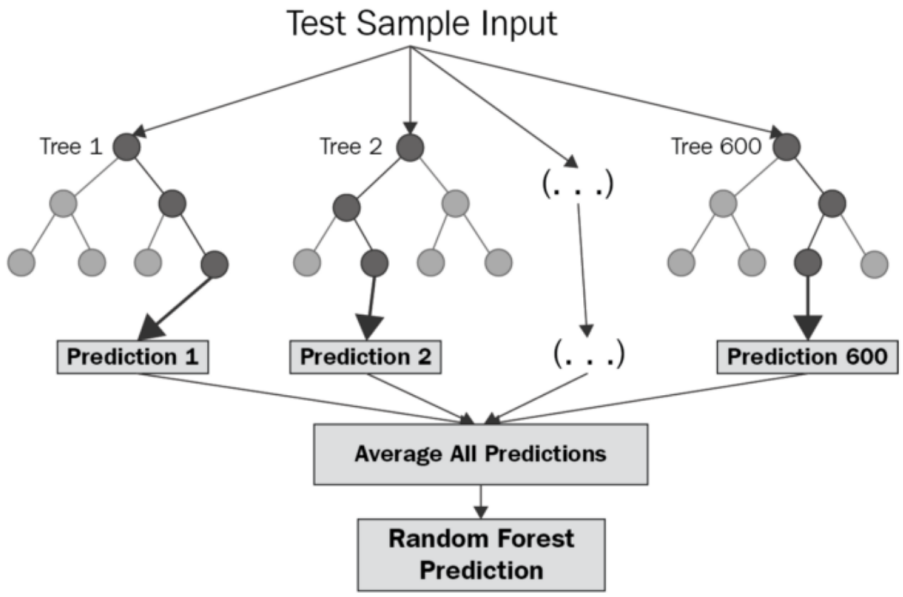

Figure 7: Random forest regressor model visualization[2]

4.2.3 Support Vector Regression

Support Vector Regression (Figure 8) is a technique based on Support Vector Machines (SVM), but instead of classifying data, it's used to predict continuous values. It works by finding a function that stays as close as possible to the real values, within a certain margin of error. SVR was chosen for this study because it's reliable when working with small or medium-sized datasets and can handle both simple and more complex relationships in data using special functions called kernels [22].

[2] https://www.keboola.com/blog/random-forest-regression

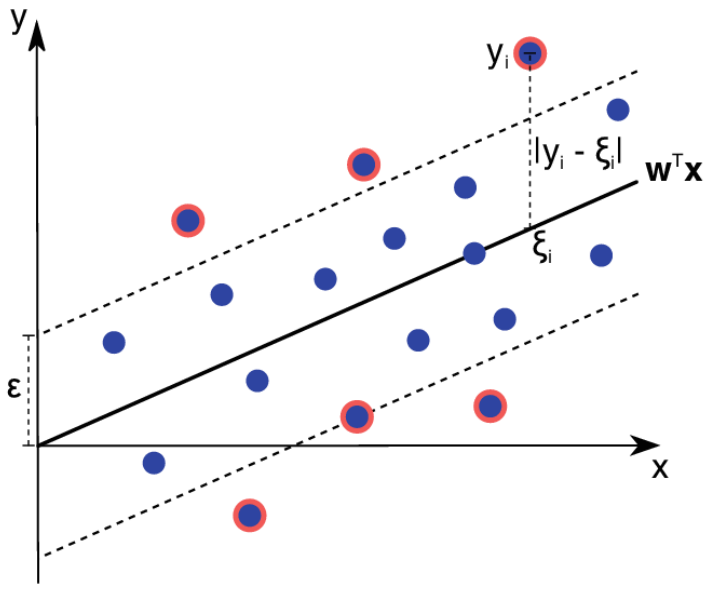

Figure 8: Support vector regression model visualization[3]

4.3 Performance Metrics and Evaluation Criteria

To evaluate the effectiveness of the prediction models, the Mean Absolute Error (MAE) was utilized as the main numerical assessment metric. MAE evaluates the average size of the absolute deviations between predicted and actual values, offering a clear error measurement in the same units as the target variable. It is a commonly used metric for regression tasks because of its straightforwardness and direct interpretability.

However, while MAE offers a useful quantitative assessment of prediction accuracy, it does not capture the temporal dynamics or trend alignment between predicted and actual pollutant concentrations. Therefore, visual inspection of the prediction plots was considered equally, if not more, important in this study.

The key objective of air pollution forecasting models is not solely to minimize numerical errors but to accurately capture trends, seasonal patterns, and fluctuations in pollutant levels over time. By carefully analyzing prediction plots for each station and pollutant type, it was possible to evaluate how well each model replicated the actual evolution of pollutant concentrations, particularly during peak events or sudden drops.

This combined evaluation strategy - integrating both numerical metrics like MAE and graphical trend analysis - ensured a more comprehensive and meaningful assessment of the models' real-world forecasting capabilities.

[3]https://www.researchgate.net/figure/Support-vector-regression-SVR-Illustration-of-an-SVR-regression-function-represented_fig12_248396465

4.4 Prediction Workflow for Multiple Stations and Models

To ensure consistency and reproducibility, a structured workflow was developed for handling data from multiple air quality monitoring stations. For each station and pollutant type (PM1, PM2.5, PM10), the following steps were performed:

- Data preprocessing, including missing value handling and Min-Max normalization.
- Creation of sequential data windows for input and corresponding future values.
- Model training and testing using a rolling window approach.
- Performance evaluation using MAE and result visualization.

This standardized methodology allowed for consistent model comparison and assessment of predictive performance across different locations and pollutant types.

5 IMPLEMENTATION

This section presents the technical implementation of the proposed forecasting workflow, detailing the used models, the sequential testing strategy adopted, and the methodology used for handling multi-station air quality datasets. All experiments were conducted using Python, with libraries such as TensorFlow, scikit-learn, pandas, and matplotlib.

5.1 LSTM Model Implementation

The core of the forecasting system was built using a Long Short-Term Memory (LSTM) neural network, specifically tailored for time series data. The implementation followed a structured sequence: data preprocessing, windowed sequence generation, model training, and prediction.

The input data consisted of air quality records, separately stored in CSV files for each pollutant (PM1, PM2.5, and PM10) and monitoring station. Each dataset was loaded, resampled to daily averages, and cleaned to remove missing values. MinMaxScaler from the scikit-learn library was applied to normalize the data between 0 and 1, a crucial step for ensuring stable convergence during neural network training.

A sliding window technique was used to create sequences of 10 consecutive days as input features, with the subsequent day's value as the prediction target. The datasets were split into training (80%) and validation (20%) subsets. The LSTM architecture consisted of the following layers:

1. **Input Layer**: Input(shape=(seq_length, 3))
   Accepts input sequences of shape (seq_length, 3), where "seq_length" is the number of past consecutive days used as input (default is 10), and 3 corresponds to the three pollutant variables: PM1, PM2.5 and PM10.
2. **First LSTM Layer**: LSTM(50, return_sequences=True)
   An LSTM layer with 50 hidden units (memory cells), "return_sequences=True" ensures that the output retains a sequence of hidden states for each time step, which is necessary since it will be passed to a second LSTM layer.
3. **First Dropout Layer**: Dropout(0.2)
   A dropout regularization layer with a dropout rate of 20%, applied to the outputs of the first LSTM layer to mitigate overfitting.
4. **Second LSTM Layer**: LSTM(50, return_sequences=False) Another LSTM layer with 50 hidden units, "return_sequences=False" means this layer only outputs the final hidden state after processing the entire sequence, which is then fed directly to the output layer.

5. **Second Dropout Layer**: Dropout(0.2)
   Another dropout layer with a 20% dropout rate, applied to the output of the second LSTM layer.
6. **Dense (fully connected) Output Layer**: Dense(3)
   A fully connected output layer with 3 neurons, corresponding to the predicted values for PM1, PM2.5 and PM10. No activation function is specified, so a linear activation is implicitly used, appropriate for a regression problem.

The model is compiled with:

- **Adam optimizer** with a learning rate of 0.001
- **Mean Squared Error (MSE)** as the loss function.
- **Mean Absolute Error (MAE)** as an additional evaluation metric.

The training process involved 50 epochs with a batch size of 16. The model's performance was assessed using the test set, and the predicted values were transformed back to their original scale by utilizing the fitted scaler.

Predictions for the following 30 days were produced through an iterative approach, where each newly predicted value was added to the input sequence, enabling the model to forecast one step ahead in a continuous manner. This approach mirrored real-world scenarios in which each day's prediction relies on the most recent data available.

5.2 Rolling Window Testing Approach

To improve the reliability of the forecast assessment, a rolling window testing strategy was implemented. This approach involved moving the training and testing windows through the dataset over time.

Specifically, each model was trained and tested on 5 subsets, which means that testing was done on the last 150 days (3.4). This method ensured the evaluation of the model's performance over multiple overlapping time intervals, providing a more comprehensive picture of predictive consistency.

Mean Absolute Error (MAE) was computed for each window, and average MAE values across all rolling iterations were reported for a robust comparison between models.

5.3 Implementation of Additional Regression Models

In addition to the LSTM model, several classical and ensemble machine learning models were implemented for comparative analysis. These included:

- **Linear Regression (LR)**: A baseline regression model using ordinary least squares to fit a linear relationship between the input sequences and the target value.
- **Random Forest Regressor (RFR)**: An ensemble learning method based on multiple decision trees, configured with 1000 estimators. The model was trained using the same input sequences as LSTM and predictions were generated through a similar rolling strategy.
- **Support Vector Regression (SVR)** : Implemented with an RBF kernel, testing various hyperparameters and finding the best ones (C=1, epsilon=0.01) to optimize the model's accuracy.

5.4 Multi-Station Data Handling

Given that the air quality data originated from multiple monitoring stations, a modular data handling system was designed to process each station's data independently while maintaining a consistent workflow.

Each monitoring station had separate CSV files for PM1, PM2.5, and PM10 measurements. The implementation included custom functions for batch loading, cleaning, and scaling the datasets. Preprocessing pipelines were applied station-wise, enabling independent sequence generation and scaling for each pollutant.

A dedicated visualization module was implemented to compare actual and predicted values for each station and pollutant over the test intervals. Graphs were generated using matplotlib and seaborn, presenting both the predicted values over the 30-day forecasting horizon and error trends across rolling windows.

In addition, to improve usability and flexibility, an interactive menu system was developed at the end of each model implementation. This system allows the user to select the desired monitoring station for which to generate 30 day predictions, displaying the corresponding forecast plots. Furthermore, a final interactive menu enables users not only to choose the target station, but also to select the specific predictive model (LSTM, Linear Regression, Random Forest Regressor or Support Vector Regression) (Figure 9) to be applied for the forecasting task (Figure 10). This feature facilitates direct, side-by-side model comparisons and scenario-based forecasting simulations for any of the available monitoring stations.

```
=== PREDICTION ===
Stations available: 1, 2, 3, 4
Models available:
 1 - LSTM
 2 - Linear Regression
 3 - Random Forest
 4 - SVR
Choose station (1-4): 1
Choose the prediction model (1-4): 3
```

Figure 9: Example of using the interactive menu

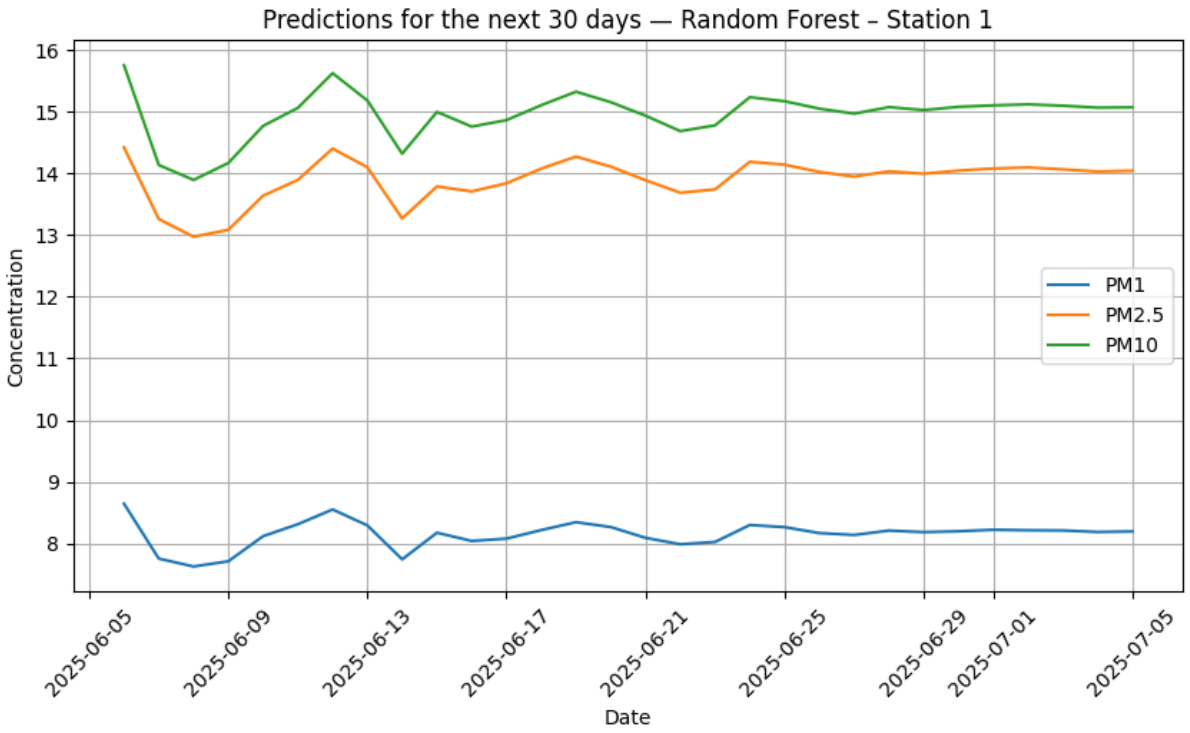

Figure 10: Prediction obtained by the Random Forest model for Station 1

This modular, multi-station architecture ensures scalability, allowing new stations and models to be integrated with minimal adjustments to the codebase while preserving consistency in preprocessing, model training, and evaluation workflows.

6 TESTING AND EVALUATION

6.1 LSTM Performance Results

The LSTM model was tested on each of the five data subsets corresponding to the four monitoring stations. For each subset, predictions for the last 30 days were generated and compared to the real values for PM1, PM2.5, and PM10. The results were visualized using time-series plots, where real values and predicted values were plotted for direct visual comparison. The LSTM model generally succeeded in capturing trends in the data, although, in some cases, slight delays in identifying trend changes were observed (Figure 11).

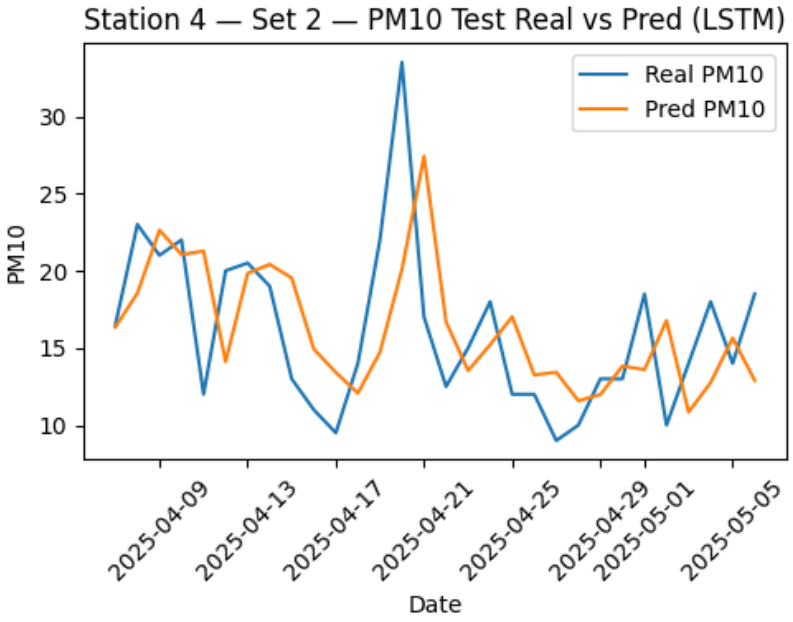

Figure 11: Slight delays in LSTM model's predictions

The Mean Absolute Error (MAE) was computed for each pollutant in every test subset. These values were recorded and later compared to the other models' results. Although MAE provides a useful numerical indicator of model accuracy, it was observed that it does not fully capture the model's ability to predict trend direction and magnitude shifts.

6.2 MAE Results Summary and Model Comparison

To systematically compare model performance, the MAE values obtained for each pollutant and subset were organized into four tables, one for each predictive model (LSTM, Random Forest, Support Vector Regression, and Linear Regression). Each table (Tables 4, 5, 6, 7) contains four columns (representing the stations) and three rows (representing the average values of the MAE on the 5 subsets).

Table 4: Mean Absolute Error (MAE) values for LSTM

Metric	Station 1	Station 2	Station 3	Station 4
PM1 MAE	3.43726	1.95124	3.99474	5.23082
PM2.5 MAE	5.49444	2.61686	4.8910	8.4951
PM10 MAE	7.55694	2.6277	5.69932	9.94908

Table 5: Mean Absolute Error (MAE) values for LR

Metric	Station 1	Station 2	Station 3	Station 4
PM1 MAE	3.6075	2.0864	4.3515	5.67234
PM2.5 MAE	5.1900	2.9759	5.5470	7.93036
PM10 MAE	7.4162	2.9980	5.8496	8.77152

Table 6: Mean Absolute Error (MAE) values for RFR

Metric	Station 1	Station 2	Station 3	Station 4
PM1 MAE	3.7010	2.33178	4.5805	5.81324
PM2.5 MAE	6.18624	2.9930	5.2340	9.73458
PM10 MAE	7.82016	2.8079	5.86934	10.23906

Table 7: Mean Absolute Error (MAE) values for SVR

Metric	Station 1	Station 2	Station 3	Station 4
PM1 MAE	4.8874	2.7272	4.9154	6.9802
PM2.5 MAE	7.4716	2.7978	5.9340	10.9854
PM10 MAE	9.5818	3.2766	6.9682	11.8712

The following is a visual representation of the MAEs for each monitoring station and the average MAE across all stations, for each pollutant and for each predictive model (Figures 12, 13, 14).

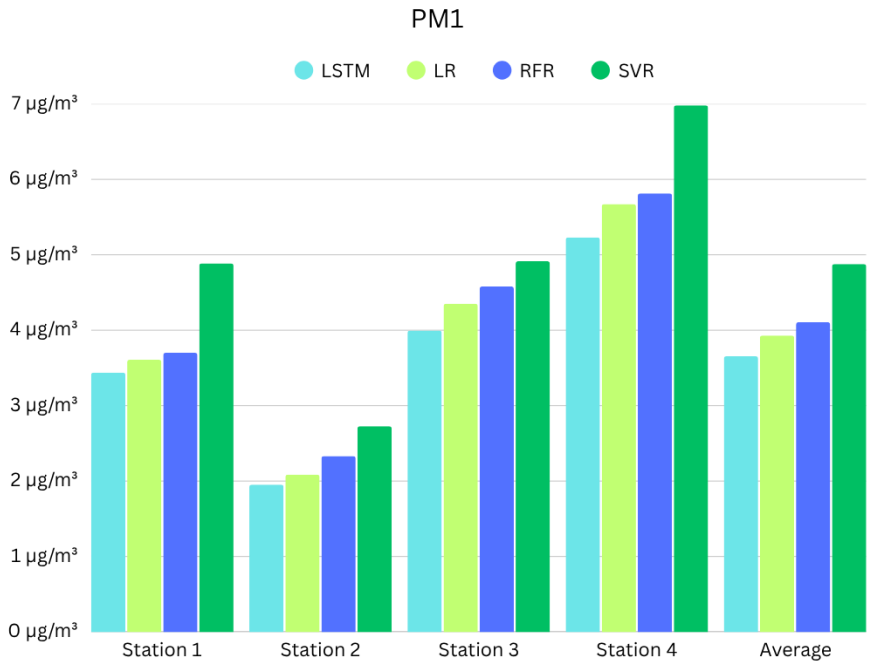

Figure 12: Performance of each prediction model for PM1

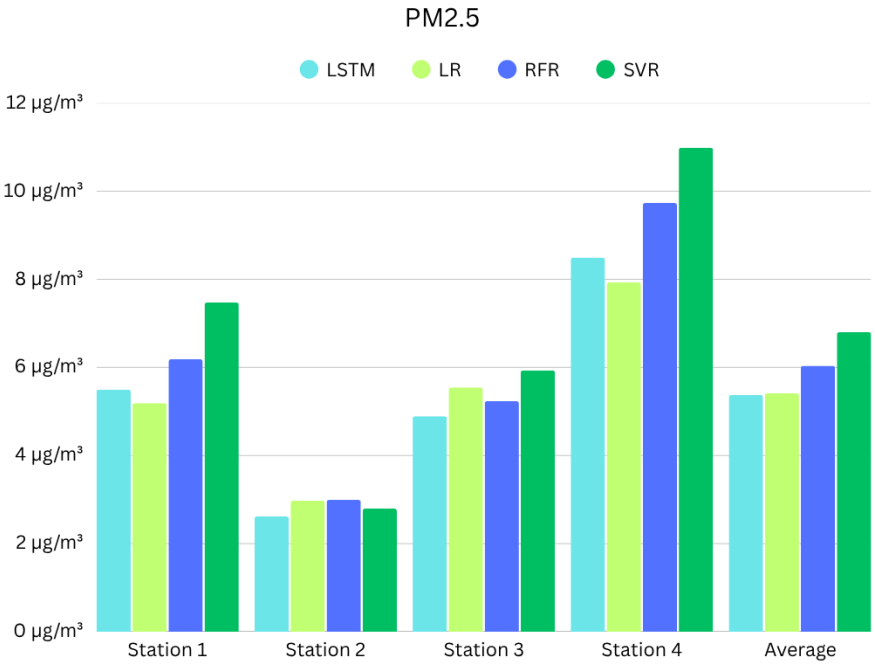

Figure 13: Performance of each prediction model for PM2.5

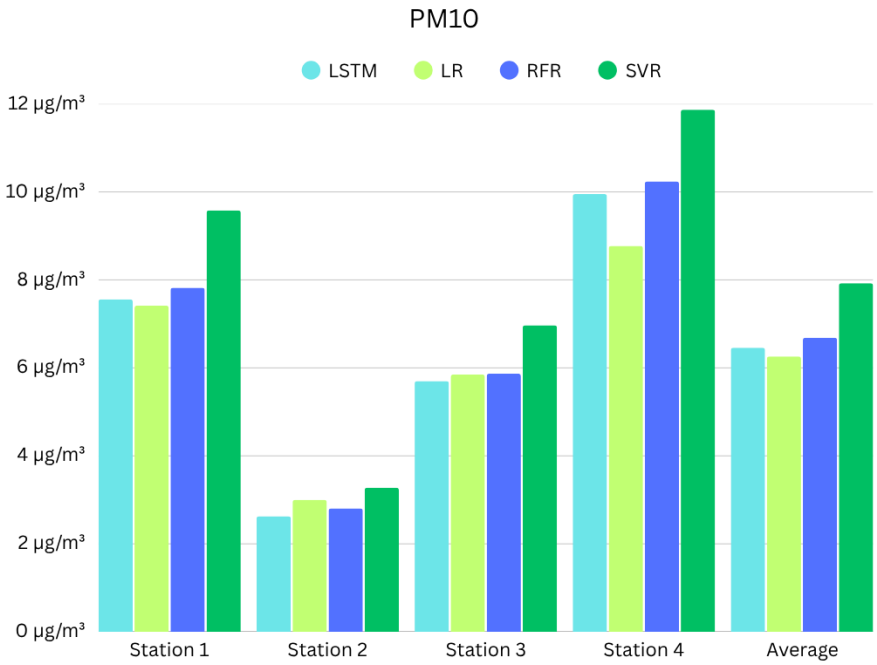

Figure 14: Performance of each prediction model for PM10

The LSTM model achieved the lowest average MAE values across the majority of pollutants and stations, demonstrating its superior capability to capture temporal dependencies and non-linear patterns inherent in air pollution time series data. Specifically, LSTM recorded an average MAE of 3.65351 for PM1, 5.37435 for PM2.5, and 6.45826 for PM10. Notably, LSTM outperformed all other models for PM1 and PM2.5 across all stations.

The Linear Regression model exhibited competitive performance, especially for PM10, where it achieved an average MAE of 6.25883, slightly surpassing LSTM's 6.45826. This outcome suggests that for certain pollutants with less pronounced non-linear temporal behavior, simpler linear models can remain surprisingly effective. For PM1 and PM2.5, however, LR trailed behind LSTM, with average MAEs of 3.92944 and 5.41082 respectively.

The Random Forest Regressor model ranked third overall, with average MAEs of 4.10663 for PM1, 6.03696 for PM2.5, and 6.68412 for PM10. Although RFR benefitted from its ensemble-based capacity to reduce variance and model complex interactions, it struggled to consistently outperform LSTM and LR, particularly for PM2.5 predictions.

Finally, the Support Vector Regression model consistently recorded the highest error values across all pollutants and stations, with averages of 4.87755 for PM1, 6.79720 for PM2.5, and 7.92445 for PM10. This outcome highlights the limitations of SVR in handling multi-station, multi-pollutant sequential data, likely due to its static kernel-based structure, which lacks intrinsic temporal memory components.

In conclusion, the LSTM model proved to be the most reliable and accurate overall, particularly excelling in PM1 and PM2.5 forecasting. The Linear Regression model, while simpler, demonstrated strong results for PM10, marginally outperforming LSTM for this pollutant.

These findings reaffirm the importance of selecting models tailored to both the nature of the data and the specific temporal characteristics of the pollutant.

It is also important to highlight the differences in prediction accuracy between the four monitoring stations. Station 2 consistently recorded the lowest MAE values across all models and pollutants. This outcome can be attributed to the fact that Station 2 is located in a less polluted area, where pollutant concentrations are generally lower and exhibit more stable, predictable patterns over time. As a result, all models - particularly LSTM - were able to capture these trends with greater accuracy.

Conversely, Station 4 registered the highest error values across all models and pollutants. This can be explained by its placement in a highly polluted urban area, characterized by frequent, abrupt fluctuations in air quality due to intense industrial activity, heavy traffic, and variable meteorological conditions. Such environments pose significant challenges for predictive models, as rapid, irregular changes in pollutant levels are inherently more difficult to forecast, especially using time series models sensitive to trend continuity.

These findings emphasize the importance of considering local environmental conditions and pollution profiles when deploying air quality forecasting systems, as model performance can vary substantially depending on the characteristics of the monitored area.

In addition, to provide a qualitative comparison, a sample prediction graph was included for each model, allowing a visual assessment of how accurately each model captured both the pollutant level trends and values.

Figures 15, 16, 17, 18 display sample predictions for each model.

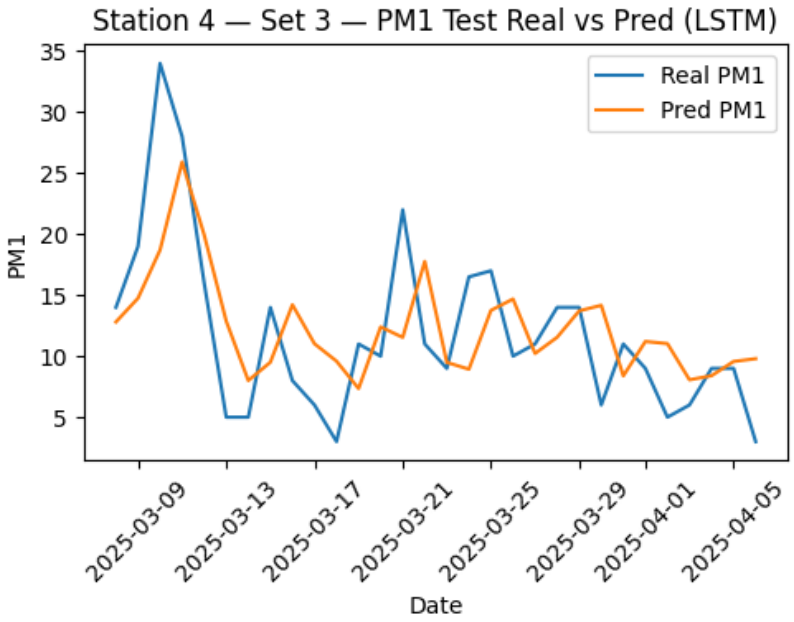

Figure 15: Example of LSTM prediction

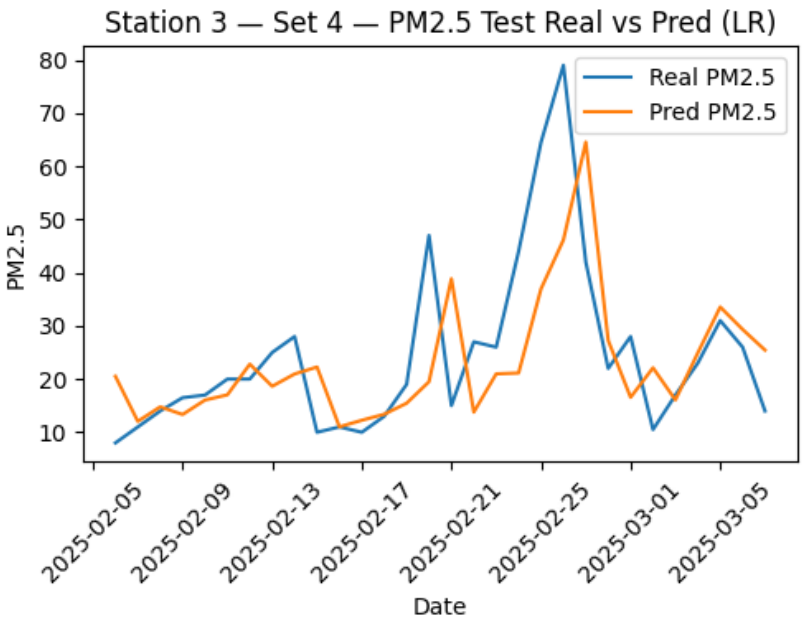

Figure 16: Example of LR prediction

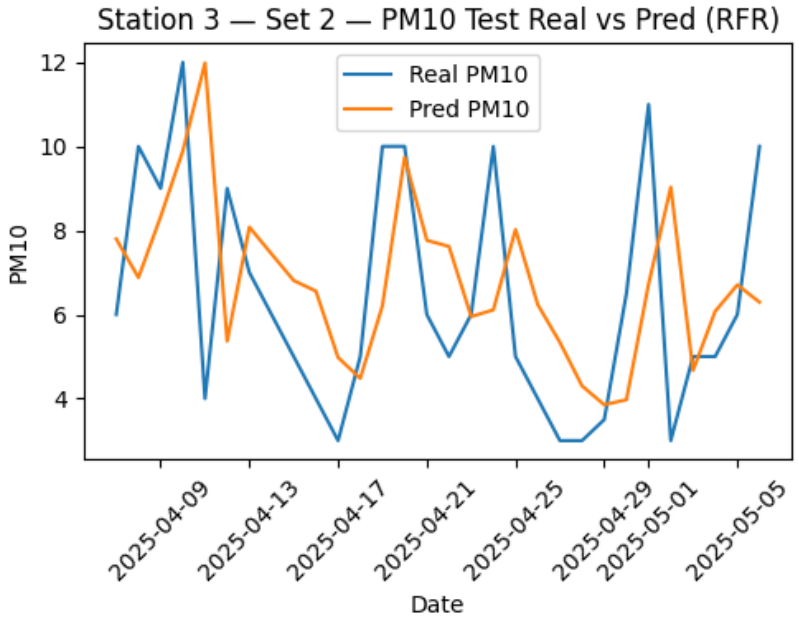

Figure 17: Example of RFR prediction

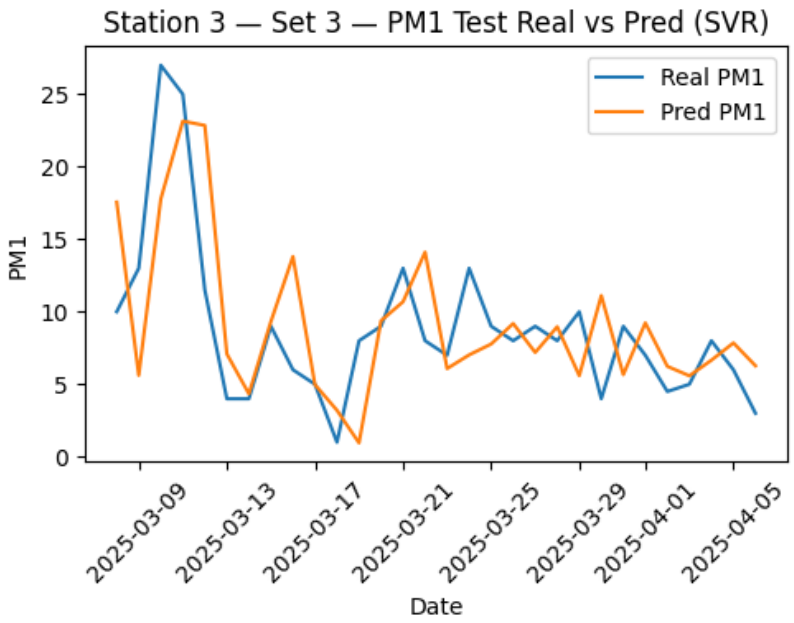

Figure 18: Example of SVR prediction

In addition, the visual inspection of the prediction graphs aligns closely with the numerical performance results obtained through the MAE metric. The models that achieved lower MAE values, particularly the LSTM model, produced forecast curves that closely followed the actual pollutant concentration trends, with smoother transitions and smaller deviations from the real data. In contrast, models with higher MAE values, such as RFR and SVR, exhibited visibly larger prediction errors and struggled to capture rapid fluctuations, especially in highly variable pollutant profiles like PM10.

This visual consistency between the plotted forecasts and the quantitative evaluation metrics reinforces the validity of the obtained results. It also highlights the value of combining both numerical metrics and graphical analyses in evaluating time series forecasting models, as certain subtle behaviors are more easily observed through visualization than through aggregate error values alone.

6.3 Multi-Station Prediction Results

The model performances were also analyzed across the four monitoring stations. While some variability in performance was observed between stations, the overall ranking of models by MAE remained consistent, with LSTM performing best and SVR the weakest.

An important feature of the implementation was the addition of an interactive menu, allowing the user to select a station and model to generate 30 day predictions. This feature facilitated comparative analysis and model benchmarking directly from the application, streamlining both experimentation and operational use.

7 DISCUSSION

7.1 Analysis of Results

Although numerical metrics such as MAE, MSE and RMSE are widely used for regression model evaluation, this study highlights their limitations when applied to time-series forecasting for environmental data. In several instances, the LSTM model accurately identified pollutant level trends but with slight delays. This led to higher MAE values despite visually accurate trend predictions.

In particular, models that predict average or smoothed values may appear better by error metrics but neglect critical fluctuations in pollutant levels [23]. Trend-sensitive metrics like Dynamic Time Warping (DTW) distance have been proposed to complement traditional evaluation [24].

In contrast, some models produced lower MAE values by predicting values close to the mean, without effectively capturing fluctuations and trend shifts. This observation suggests that, while numerical error metrics offer a useful benchmark, they do not fully reflect a model's ability to detect and follow environmental trends, which are critical in air pollution monitoring and decision-making.

The most reliable evaluation in this context was the graphical analysis of the predicted versus actual pollutant levels over time. Visualizing trends allowed for the identification of model behavior.

7.2 Strengths and Limitations

The multi-model, multi-model, and multi-subset testing structure offered comprehensive insights into model performance and generalizability. The interactive menu for station and model selection significantly improved usability, providing a practical tool for rapid scenario testing.

Limitations include the exclusive reliance on MAE as a quantitative metric and the absence of a dedicated trend-following score. The models also did not incorporate external factors, such as meteorological conditions, which are known to affect air pollutant concentrations.

7.3 Potential Improvements

Future enhancements could include implementing custom trend detection metrics, such as Dynamic Time Warping (DTW) distance or trend correlation measures. Incorporating exogenous variables like temperature, humidity, and wind speed could improve prediction accuracy and trend alignment.

Another potential improvement involves using ensemble models combining LSTM with decision tree-based models, or experimenting with sequence-to-sequence architectures and attention-based models such as Transformer-based regressors.

8 CONCLUSIONS AND FURTHER WORK

8.1 Final Conclusions

This thesis investigated the performance of multiple machine learning models for air pollution forecasting, focusing on daily predictions of PM1, PM2.5, and PM10 concentrations from four urban monitoring stations. The study demonstrated that deep learning architectures, particularly Long Short-Term Memory (LSTM) networks, consistently outperformed classical regression models and methods such as Random Forest Regressor and Support Vector Regression.

The project revealed several key insights:

- Using only numerical metrics, like Mean Absolute Error (MAE), limits our understanding of a model's effectiveness in tracking variations and trends in pollutants.
- Visual trend assessments emerged as a useful supplementary evaluation method, showcasing situations where LSTM models effectively captured the trend's direction and shape, despite higher numerical errors.
- The multi-model, multi-station experimental design offered a thorough evaluation of the models' generalizability and robustness.
- Integrating an interactive prediction menu significantly improved the project's practical functionality, allowing for quick model comparisons and real-time scenario analysis.

In summary, the thesis adds to the expanding research on data-driven forecasting of air quality and suggests a practical approach for both researchers and decision makers.

8.2 Future Work and Recommendations

Although the results obtained in this study are promising, several areas for future improvement have been identified:

- **Incorporation of Exogenous Variables**: Future models should include external factors such as temperature, humidity, wind speed, and atmospheric pressure, which are known to influence air pollution levels.
- **Trend-Focused Evaluation Metrics**: Developing and integrating custom metrics like Dynamic Time Warping (DTW) distance or trend correlation coefficients would offer a more realistic assessment of a model's trend-following capabilities.

- **Ensemble and Hybrid Models**: Combining the strengths of LSTM with decision tree-based models or exploring Transformer-based sequence models could enhance both accuracy and stability [25].
- **Extended Prediction Horizons**: Future work could evaluate the models' performance for longer-term forecasts (60 or 90 days) and assess prediction reliability over extended periods.
- **Geospatial Interpolation and Spatial Modeling**: Integrating data from additional monitoring stations or applying spatial interpolation techniques could improve regional predictions and fill data gaps in areas without direct measurements.

By addressing these directions, future research can further enhance the precision, applicability, and interpretability of air pollution forecasting models, contributing to smarter urban environmental management systems.

BIBLIOGRAPHY

[1] J. Lelieveld, J. S. Evans, M. Fnais, D. Giannadaki, and A. Pozzer. The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature*, 525(7569):367–371, 2015.

[2] Yang Zhang, Marc Bocquet, Vivien Mallet, Christian Seigneur, and Alexander Baklanov. Real-time air quality forecasting, part i: History, techniques, and current status. *Atmospheric Environment*, 60:632–655, 2012.

[3] Yu Zheng, Furui Liu, and Hsun-Ping Hsieh. U-air: when urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, page 1436–1444, New York, NY, USA, 2013. Association for Computing Machinery.

[4] Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. A comparison of arima and lstm in forecasting time series. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1394–1401, 2018.

[5] Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006.

[6] Dean E. Schraufnagel and John R. Balmes et al. Health benefits of air pollution reduction. *Annals of the American Thoracic Society*, 16(12):1478–1487, 2019. PMID: 31774324.

[7] Ioannis Manisalidis, Elisavet Stavropoulou, Agathangelos Stavropoulos, and Eugenia Bezirtzoglou. Environmental and health impacts of air pollution: A review. *Frontiers in Public Health*, Volume 8 - 2020, 2020.

[8] Sunil Gulia, S.M. Shiva Nagendra, Mukesh Khare, and Isha Khanna. Urban air quality management-a review. *Atmospheric Pollution Research*, 6(2):286–304, 2015.

[9] Daewon Byun and Kenneth L. Schere. Review of the governing equations, computational algorithms, and other components of the models-3 community multiscale air quality (cmaq) modeling system. *Applied Mechanics Reviews*, 59(2):51–77, 03 2006.

[10] George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. *Time Series Analysis: Forecasting and Control, 5th Edition*. Wiley, Hoboken, New Jersey, 2015.

[11] P Veefkind and I. Aben et al. Tropomi on the esa sentinel-5 precursor: A gmes mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications. *Remote Sensing of Environment*, 120:70–83, 05 2012.

[12] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 11 1997.

[14] Muhammad Haseeb, Zainab Tahir, Syed Amer Mahmood, Hania Arif, Khalid F. Almutairi, Walid Soufan, and Aqil Tariq. Comparative analysis of machine learning models for predicting pm2.5 concentrations using meteorological and chemical indicators. *Journal of Atmospheric and Solar-Terrestrial Physics*, 263:106338, 2024.

[15] Tongwen Li, Huanfeng Shen, Qiangqiang Yuan, Xuechen Zhang, and Liangpei Zhang. Estimating ground-level pm2.5 by fusing satellite and station observations: A geo-intelligent deep learning approach. *Geophysical Research Letters*, 44(23):11,985–11,993, 2017.

[16] Aaron van Donkelaar, Randall V. Martin, Michael Brauer, and Brian L. Boys. Use of satellite observations for long-term exposure assessment of global concentrations of fine particulate matter. *Environmental Health Perspectives*, 123(2):135–143, 2015.

[17] Pawan Gupta and Sundar A. Christopher. Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: 2. a neural network approach. *Journal of Geophysical Research: Atmospheres*, 114(D20), 2009.

[18] Jianjun Ni, Yan Chen, Yu Gu, Xiaolong Fang, and Pengfei Shi. An improved hybrid transfer learning-based deep learning model for pm2.5 concentration prediction. *Applied Sciences*, 12(7), 2022.

[19] Gokulan Ravindiran, K. Karthick, Sivarethinamohan Rajamanickam, Deepshikha Datta, Bimal Das, G. Shyamala, Gasim Hayder, and Azees Maria. Ensemble stacking of machine learning models for air quality prediction for hyderabad city in india. *iScience*, 28(2):111894, 2025.

[20] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[21] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *Forest*, 23, 11 2001.

[22] Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, August 2004.

[23] Fotios Petropoulos and Daniele Apiletti et al. Forecasting: theory and practice. *International Journal of Forecasting*, 38(3):705–871, 2022.

[24] Toni Giorgino. Computing and visualizing dynamic time warping alignments in r: The dtw package. *Journal of Statistical Software*, 31(7):1–24, 2009.

[25] Magesh T, Supriya S, Yuvaprakash A, Vishvapriya R T, Nisha C, and Govindaraajan P. Hybrid weather forecasting: Integrating lstm neural networks and random forest models for enhanced accuracy. In *2024 International Conference on Recent Advances in Electrical, Electronics, Ubiquitous Communication, and Computational Intelligence (RAEEUCCI)*, pages 1–5, 2024.