Analyzing Air Quality: Insights into the Impact of Weather and Traffic in Big Cities

Atreyo Das
das.at@northeastern.edu

Omer Seyfeddin Koc
koc.o@northeastern.edu

Shruti Suhas Kute
kute.s@northeastern.edu

Northeastern University, Boston, MA

Abstract

Urban air pollution is a critical public health concern, particularly in densely populated cities where traffic congestion and varying weather conditions significantly impact air quality. This study investigates the relationship between air pollution, traffic density, and meteorological conditions in three major U.S. cities: New York, Chicago, and Los Angeles. We employ advanced machine learning techniques, including Random Forest, XGBoost, ARIMA, and LSTM models, to identify key predictors of pollutant levels (PM2.5, PM10, and NO_2) and forecast future air quality trends. XGBoost emerged as the top-performing model for regression-based predictions, achieving high accuracy in predicting pollutant levels, while LSTM excelled in capturing long-term dependencies for time-series forecasting. The insights gained from this research offer actionable recommendations for policymakers to design targeted interventions, mitigate pollution, and improve public health in urban areas.

1 Introduction

Air pollution remains a critical public health and environmental concern, particularly in densely populated urban areas where vehicular emissions and fluctuating weather conditions contribute to high levels of pollutants such as PM2.5, PM10, and NO_2.Understanding the complex interactions between these factors is essential for developing accurate predictive models that can provide actionable insights to mitigate pollution and improve public health outcomes. This project focuses on analyzing the relationship between air quality, weather conditions, and traffic patterns in three major U.S. cities—New York, Chicago, and Los Angeles—by employing advanced machine learning techniques.

The primary objectives of the experimentation phase were twofold. The first task focuses on identifying predictors of pollutant levels (PM2.5, PM10, and NO_2) using regression models to evaluate the predictive power of weather parameters (e.g. temperature, precipitation, wind speed, humidity) and traffic data. The next stage of our project involves forecasting future pollutant levels by studying past trends using a time series model. To achieve these tasks, we used regression models like Random Forest Regressor and XGBoost models for first objective and, ARIMA and LSTM models were used for our latter objective. Comparing these models allow us to properly understand the nature of our time series data and help us understand which models are better suited to provide accurate results. XGBoost emerged as the most accurate model for regression tasks, consistently outperforming other models with higher R^2 values and lower error rates, while LSTM demonstrated superior performance in capturing long-term dependencies and sequential patterns in pollutant data. More detailed exploratory data analysis and model code can be found here[1].

2 Related Work

Air quality prediction has been extensively studied using a variety of machine learning and statistical approaches, with a focus on understanding the influence of meteorological and traffic-related factors on pollutant levels. Prior research has explored diverse methodologies ranging from traditional regression models to advanced deep learning techniques, providing a foundation for the experimentation conducted in this project.

In the paper (Castelli et al., 2020), the authors used historical air quality data from the U.S. Environmental Protection Agency (EPA), along with meteorological and environmental variables, to predict air quality in California using Support Vector Regression (SVR) with an RBF kernel. Although SVR demonstrated effectiveness in capturing nonlinear relationships, it exhibited limitations such as computational inefficiency for large datasets, sensitivity to hyperparameter tuning, and reduced interpretability. Another study (Liu et al., 2019) compared the performance of SVR and Random Forest Regression (RFR) using the Beijing Air Quality Dataset (2013–2018) and an Italian dataset (De Vito et al., 2009) containing multi-sensor air quality recordings. While SVR was effective in predicting the Air Quality Index (AQI) for Beijing, RFR performed better for NOX prediction in the Italian dataset. Despite its robustness, RFR's susceptibility to overfitting and the misrepresentation of feature importance for small datasets highlighted the need for careful model tuning, which we addressed by using hyperparameter optimization and cross-validation techniques in our experimentation.

Time-series models have proven effective in forecasting pollutant levels in several different studies. There are studies

[1] https://github.com/Shrutikute09/Capstone_2025

(Abhilash et al., 2018) that analyzed air pollution data from Bengaluru, India, using ARIMA models to predict SO_2, NO_2, and RSPM concentrations. Although ARIMA performed well in short-term forecasting when the data was stationary, its accuracy diminished for non-stationary data, limiting its reliability in capturing complex temporal patterns. Recognizing these limitations, we incorporated LSTM models, which are well-suited for handling non-linear and sequential dependencies in air quality data (Siami-Namini et al., 2019), to complement ARIMA's capabilities. This greater performance of LSTM can be observed when researchers used hourly air quality data from Kuwait, including ozone (O3) concentrations and meteorological variables, to build an LSTM-based recurrent neural network (Freeman et al., 2018). LSTM demonstrated superior performance by capturing long-term temporal dependencies, but it was prone to overfitting due to its complexity.

A comparative evaluation of different machine learning techniques was also conducted, where the authors analyzed air pollution data from five major Chinese cities (Guangzhou, Chengdu, Beijing, Shanghai, and Shenyang) between 2010 and 2015 (Ameer et al., 2019). Models such as Decision Tree, Random Forest, Gradient Boosting, and Multi-Layer Perceptron were compared, with Random Forest emerging as the most accurate model. However, its slightly higher processing time highlighted efficiency as a minor limitation. Our decision to include Random Forest and XGBoost in the experimentation phase was influenced by these findings, as both models are known for their strong predictive performance and robustness in handling structured environmental data.

While prior studies have demonstrated the effectiveness of individual models for air quality prediction, few have explored the integration of traffic and weather data to enhance model performance. Our project differentiates itself by incorporating traffic data, specifically for New York City, alongside meteorological parameters to improve the predictive accuracy of pollutant levels. Additionally, by combining ensemble models like Random Forest and XGBoost with time-series models such as ARIMA and LSTM, our approach provides a comprehensive evaluation of both regression-based and time-series forecasting techniques. By adopting this approach, our work not only contributes novel insights to air quality prediction but also enhances model accuracy and generalizability across diverse datasets and regions.

3 Objectives and Hypotheses

The primary objective of this project is to analyze the relationship between air quality, weather conditions, and traffic patterns in major U.S. cities and develop accurate predictive models for pollutant levels. By employing machine learning and time-series techniques, the project aims to identify key factors influencing air quality and build robust models capable of predicting pollutant concentrations (PM2.5, PM10, and NO_2) in urban environments. Specifically, the experimentation phase is designed to evaluate the effectiveness of different models in capturing these relationships and providing accurate forecasts of pollutant levels across multiple time periods and locations. The experimentation phase is guided by a set of well-defined hypotheses that align with the project's objectives. The first hypothesis posits that meteorological variables such as temperature, humidity, precipitation, and wind speed significantly influence pollutant concentrations, and these relationships can be effectively captured using machine learning regression models. The second hypothesis suggests that traffic density, particularly in densely populated areas like New York City, contributes significantly to increased levels of pollutants, making it an important feature in air quality prediction models. The third hypothesis proposes that time-series models such as LSTM and ARIMA will provide more accurate predictions for sequential air quality data compared to regression models by capturing long-term temporal dependencies.

By testing these hypotheses, the project aims to establish a comprehensive understanding of the factors influencing air quality in urban environments and identify optimal modeling approaches for predicting pollutant levels. The insights gained from this phase will not only validate the effectiveness of different machine learning and time-series models but also provide a foundation for improving air quality forecasting and informing policy interventions aimed at reducing pollution in metropolitan areas.

4 Experimental Setup

4.1 Hardware and Software

The experiments were conducted on a **MacBook Pro 14-inch (November 2024 model)** with the following specifications:

- **Chip:** Apple M4 Pro
- **Memory:** 24 GB RAM
- **Operating System:** macOS 15.3.2
- **Startup Disk:** Macintosh HD

Software tools and libraries used include:

- **Programming Language:** Python (Jupyter Notebook environment)
- **Machine Learning Libraries:** scikit-learn, XGBoost, TensorFlow/Keras (for LSTM), statsmodels (for ARIMA)
- **Data Processing:** Pandas, NumPy
- **Visualization Tools:** Matplotlib, Seaborn

4.2 Model Selection

Based on prior evaluations and performance comparisons, we selected two primary models for different tasks:

1. **XGBoost:** Chosen for regression-based predictions due to its high accuracy, ability to handle structured data, and computational efficiency. It consistently outperformed Random Forest in predictive performance.

2. **LSTM (Long Short-Term Memory):** Selected for time-series forecasting as it captured temporal dependencies better than traditional statistical models like ARIMA.

Other models considered but not finalized:

- **Random Forest:** Performed reasonably well but struggled with capturing time-series dependencies.
- **ARIMA:** Effective for short-term forecasting but lacked adaptability for dynamic urban air quality trends.

4.3 Justification for Model Selection

4.3.1 XGBoost - Best Regression Model

- **Superior Predictive Performance:** XGBoost consistently produced higher R^2 values and lower error rates, making it the most accurate choice for structured air pollution data.
- **Robust Feature Handling:** It effectively manages missing values and outliers, which is crucial for large-scale and diverse environmental datasets.
- **Computational Efficiency:** XGBoost is optimized for speed and scalability, allowing efficient model training even with large datasets.

4.3.2 LSTM - Best Time-Series Model

- **Effective Temporal Modeling:** LSTM excelled in capturing long-term dependencies and fluctuations in air pollution levels.
- **Deep Learning Flexibility:** It adapts well to complex sequential data, making it superior for time-series forecasting compared to ARIMA and traditional machine learning models.
- **Robust Sequential Predictions:** Unlike regression models, LSTM can predict future pollutant levels based on past trends, making it more suited for dynamic air quality forecasting.

This setup ensured efficient model training and evaluation while maintaining computational feasibility on the available hardware.

5 Dataset and Preprocessing

5.1 Data Collection

We collected information from several trustworthy sources in order to examine the connection between traffic, weather, and air quality in large cities. Every dataset offers insightful information on various facets of the urban environment. Due to the project's diverse objectives, the information gathered for it came from three main sources.

5.1.1 Air Quality

The United States EPA (US Environmental Protection Agency, 2025) monitors pollutants nationwide and offers comprehensive pre-generated datasets on its official website, detailing various pollutants dating back to 1980. These datasets are categorized into Average Yearly Summaries, Daily Levels, and Hourly Levels, further separated by pollutant type. For this project, we utilized fifteen different datasets focused on three pollutants: PM2.5 (non-FRM/FEM Mass), PM10, and NO2, collected on an hourly basis from 2020 to 2024.

The datasets were downloaded in compact CSV format for easier access. Each dataset contained approximately 1-4 million rows and around 24 columns. The columns included essential information such as the county name and number, the latitude and longitude of monitoring sites, the parameter collected along with its unit, and the time of data collection. This dataset provided access to pollutant levels at any hour for most counties in the United States.

5.1.2 Meteorological Information

Weather data was collected from Open-Meteo, an open-source weather API that offers free access for non-commercial use (Zippenfenig, 2023). The API compiles historical weather data by collecting observations from weather stations, aircraft, buoys, radars, and satellites, resulting in a comprehensive record of past weather conditions. The platform features a user-friendly interface that allows users to select a location by name or coordinates, specify a desired time range, and choose from a variety of hourly weather variables such as temperature, air pressure, humidity, and wind speed.

Using this website, we generated three weather datasets for Chicago, New York, and Los Angeles, covering the years 2020 to 2024. We selected four key parameters for analysis: temperature, relative humidity, precipitation levels (both rain and ice), and wind speed (measured 100 meters above ground). The resulting dataset contained over 43,000 rows and five columns, covering these parameters along with the corresponding timestamps.

5.1.3 Traffic Count

Although traffic datasets for U.S. cities are widely available, finding datasets with uniformly collected hourly traffic metrics across all the cities studied proved challenging. Consequently, we narrowed our focus to New York City. Traffic data was obtained from New York State's Official Website. The dataset, MTA Bridges and Tunnels Hourly Crossings (Metropolitan Transportation Authority, 2025), provides hourly data on bridge and tunnel crossings categorized by facility, direction, vehicle class, and payment method (such as EZ-Pass or Tolls). The data was available in multiple formats, including CSV, RDF, XML, and JSON.

The dataset is regularly updated and, at the time of collection, contained approximately 11 million rows. The eleven columns in the dataset provided detailed information such as the date, time, facility name, vehicle types, traffic counts, and other relevant attributes.

5.2 Data Processing

To ensure the data was suitable for analysis, multiple processing steps were undertaken. This involved combining data from different sources, imputing missing values, and transforming features to better capture trends and patterns.

5.2.1 Data Accumulation

A total of nineteen datasets were gathered for this project. To streamline analysis, essential information was extracted from these datasets and consolidated into a single unified dataset. The pollutant datasets were organized in an orderly format where each row represented the pollutant levels for a specific hour in the selected regions. To narrow the scope of our analysis, we focused on three specific counties: New York (primarily encompassing the Manhattan borough of New York City), Cook (part of the Chicago Metropolitan Area), and Los Angeles (representing the Greater Los Angeles Area). Additionally, the analysis was restricted to data collected during the month of January for each year under consideration. Upon reviewing the collected data, we identified that the EPA dataset lacked PM10 and NO_2 concentration levels for New York County, limiting our analysis in that region to PM2.5 levels only.

The weather data obtained from Open-Meteo was merged by matching records based on the corresponding date, hour, and county name. For New York's traffic data, we examined the facilities where the data was collected and filtered out entries that were not located within or directly connected to the county. We then aggregated traffic counts across the relevant facilities for the corresponding time period to compute the Average Traffic Count, which was subsequently merged into the dataset for New York entries.

As a result, the final aggregated dataset comprised 11,160 rows and 14 columns, encompassing all relevant air pollutant, weather, and traffic data required for the analysis.

5.2.2 Data Transformation

Several preprocessing techniques were applied to refine the dataset and ensure its suitability for analysis:

- **Linear Interpolation**: Minor gaps in the datasets were filled using linear interpolation to maintain continuity in the data without introducing artificial spikes.

- **Scaling**: The data was scaled using Standard Scaling to ensure features followed a consistent distribution, improving model performance.

- **Feature Engineering**: Additional features were generated to capture relevant patterns such as rolling average of the pollutant levels to highlight short-term pollutant trends while smoothing out random fluctuations and cyclic encoding of the "Day" variable into sine and cosine components to account for cyclical trends in weekly patterns.

6 Training and Validation Process

6.1 Training Setup

The training process was designed to ensure model robustness and accuracy. The key configurations included:

- **Batch Sizes:**
  - LSTM: 32
  - XGBoost and Random Forest: Not applicable (tree-based models do not use batch training)
  - ARIMA: Sequential forecasting (batch processing not required)
- **Number of Epochs:**
  - LSTM: 20 epochs (with early stopping based on validation loss)
  - XGBoost: Trained iteratively with optimized stopping criteria
  - Random Forest: Trained using an ensemble of 100 decision trees
  - ARIMA: Optimized using grid search for (p, d, q) parameters
- **Optimization Algorithms:**
  - LSTM: Adam optimizer (learning rate = 0.001)
  - XGBoost: Gradient Boosting with a learning rate of 0.05
  - Random Forest: Bootstrap aggregation
- **Regularization Techniques:**
  - XGBoost: L1 (Lasso) and L2 (Ridge) regularization applied to prevent overfitting.
  - LSTM: Dropout layers with a dropout rate of 0.2 to mitigate overfitting.
  - Random Forest: Limited tree depth to prevent excessive model complexity.

6.2 Validation Strategies

To ensure the reliability of our models, rigorous validation techniques were implemented:

- **Hyperparameter Tuning:**
  - XGBoost: Optimized using **Grid Search CV** to tune the number of estimators, learning rate, and tree depth.
  - LSTM: Manually tuned batch size, neuron count, and dropout rate.
  - Random Forest: Hyperparameter tuning performed using **Random Search CV**.
  - ARIMA: (p, d, q) parameters were optimized using **grid search** to minimize AIC and BIC scores.
- **Cross-Validation:**
  - **K-Fold Cross-Validation** (5-fold) was applied to XGBoost and Random Forest.

- **Time-Series Validation** (rolling window) was applied to ARIMA and LSTM to maintain temporal consistency.

- **Overfitting Prevention:**
  - **Early stopping** was used for LSTM to halt training when validation loss plateaued.
  - **Dropout regularization** (rate: 0.2) was applied to LSTM layers.
  - **Feature pruning** was applied in Random Forest and XGBoost to enhance model generalizability.

6.3 Final Testing

Once optimal hyperparameters were identified, the final model evaluation was performed on unseen test data:

- **Test Dataset Split:**
  - 80% Training, 20% Testing.
- **Performance Metrics:**
  - Mean Absolute Error (MAE)
  - Root Mean Squared Error (RMSE)
  - R^2 (coefficient of determination)
- **Model Performance:**
  - XGBoost had the highest R^2 score (**0.9019** for NO_2 prediction in Chicago).
  - LSTM performed best in sequential forecasting with an R^2 of **0.75** for PM2.5 in Chicago.
- **Reproducibility Measures:**
  - All models were trained using fixed random seeds to ensure consistency.
  - Code and datasets are available at https://github.com/Shrutikute09/Capstone_2025 for replication.

The results confirm that **XGBoost** was the most effective regression model, while **LSTM** was optimal for time-series forecasting. The integration of cross-validation, hyperparameter tuning, and robust testing strategies ensured the generalizability and reproducibility of our results.

7 Results

7.1 Quantitative Results

To evaluate model performance, we compared different machine learning models using key metrics such as **Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 (coefficient of determination)**. Below is a summary of our findings:

- **XGBoost** outperformed other models in predicting pollutant levels, achieving the highest R^2 scores and the lowest error rates.

Model	Performance Metrics
Random Forest	PM2.5 (R^2): 0.78 PM10 (R^2): 0.75 NO_2 (R^2): 0.80 MAE (PM2.5): 3.21 RMSE (PM2.5): 5.67
XGBoost	PM2.5 (R^2): 0.85 PM10 (R^2): 0.82 NO_2 (R^2): 0.88 MAE (PM2.5): 2.97 RMSE (PM2.5): 4.89
LSTM	PM2.5 (R^2): 0.76 PM10 (R^2): 0.73 NO_2 (R^2): 0.78 MAE (PM2.5): 3.42 RMSE (PM2.5): 5.92
ARIMA	PM2.5 (R^2): 0.65 PM10 (R^2): 0.61 NO_2 (R^2): 0.67 MAE (PM2.5): 4.12 RMSE (PM2.5): 6.34

Table 1: Comparison of Model Performance Metrics

- **Random Forest** performed well but had slightly higher errors compared to XGBoost.
- **LSTM** was effective for time-series forecasting, but its performance varied across different pollutants.
- **ARIMA** had the lowest accuracy as it struggled to capture complex relationships between air quality, traffic, and weather conditions.

7.2 Qualitative Analysis

In addition to numerical evaluation, we analyzed sample outputs to understand how well the models captured pollution trends.

- **Time-Series Forecasting:** LSTM and ARIMA were tested for future air quality predictions. While LSTM captured trends better than ARIMA, it sometimes struggled with sudden pollution spikes.
- **Feature Importance:** XGBoost and Random Forest provided insights into which features had the most impact on predictions. Traffic volume and temperature were the most significant predictors, confirming our hypothesis.
- **Data Challenges:** Some missing values required interpolation, and certain pollution readings showed unexpected fluctuations due to data collection inconsistencies.

7.3 Interpretation

The experimentation phase provided valuable insights into air pollution trends and predictive modeling. One of the key

successes was the successful development and evaluation of multiple machine learning models for air quality prediction. Among these, **XGBoost emerged as the best-performing model**, demonstrating the highest accuracy in predicting pollutant levels. Furthermore, our findings confirmed that **traffic and weather conditions significantly impact pollution levels**, aligning with our initial hypotheses.

Despite these successes, there were certain limitations in the study. One notable constraint was the **geographical bias** in the dataset, as data was collected from only three major cities—New York, Chicago, and Los Angeles—limiting its generalizability to other regions. Additionally, the availability of **traffic data was restricted to New York**, making it challenging to analyze its impact comprehensively across all cities. Another challenge was related to **time-series forecasting**, where the LSTM model required extensive tuning and struggled with short-term fluctuations in pollution trends.

In terms of alignment with our objectives and hypotheses, our results strongly support the idea that **weather conditions significantly influence air pollution**. We observed a clear correlation between pollutant levels and meteorological factors such as temperature and wind speed. Similarly, the hypothesis that **traffic congestion contributes to increased pollution** was validated, with traffic volume emerging as a key predictor. The application of **machine learning models for air quality forecasting proved to be effective**, with XGBoost delivering the most reliable predictions.

Overall, these findings reinforce the importance of integrating multiple environmental factors in pollution modeling and highlight opportunities for future research to further improve prediction accuracy and expand applicability across different urban environments.

8 Evaluation Metrics and Validation

8.1 Performance Metrics

To assess the effectiveness of our machine learning models for air quality prediction, we used the following evaluation metrics:

- **R^2 (Coefficient of Determination):** Measures how well the model explains variance in pollutant levels. A higher R^2 value (closer to 1) indicates better predictive accuracy. Since air quality prediction involves continuous numerical values (PM2.5, PM10, NO_2), R^2 helps determine how well the model fits the data.

- **Mean Absolute Error (MAE):** Calculates the average absolute difference between actual and predicted pollution values. Lower values indicate better model performance. MAE provides an easy-to-interpret error measurement, making it useful for understanding real-world prediction deviations.

- **Root Mean Squared Error (RMSE):** Similar to MAE but penalizes larger errors more heavily. Lower RMSE values indicate better predictive performance. RMSE is sensitive to large deviations, making it useful for identifying models that fail in extreme pollution scenarios.

- **Feature Importance Analysis (for Tree-Based Models):** Used in Random Forest and XGBoost to identify key factors influencing pollution levels. Helps interpret the most significant predictors, such as traffic volume, temperature, and wind speed. This analysis helps validate whether the model is using meaningful environmental factors for prediction.

8.2 Validation Techniques

To ensure the robustness of our models, we implemented the following validation techniques:

- **Train-Test Split (80-20 Split):** The dataset was divided into **80% training data** and **20% testing data** to evaluate generalization. This ensured that models were tested on unseen data to avoid overfitting.

- **K-Fold Cross-Validation (for XGBoost and Random Forest): 5-fold cross-validation** was applied to improve the reliability of results. Each model was trained and tested on different subsets, reducing the risk of overfitting.

- **Time-Series Validation (for LSTM and ARIMA):** Instead of random shuffling, we used a **rolling window validation** method. This ensured that models learned sequential patterns from past pollution data and helped validate how well the models could predict future air quality trends.

- **Hyperparameter Tuning (Grid Search & Random Search):**
  - **XGBoost & Random Forest:** Used **Grid Search CV** to optimize tree depth, learning rate, and number of estimators.
  - **LSTM:** Tuned batch size, number of neurons, and activation functions through manual experimentation.
  - **ARIMA:** Used **grid search** to select optimal (p, d, q) parameters.

These validation techniques ensured that our models provided **accurate and generalizable air quality predictions** across different environmental conditions.

9 Discussion and Iterative Improvements

9.1 Model Adjustments

Throughout the experimentation process, several adjustments were made to the model architecture, feature selection, and hyperparameters to improve performance. One significant refinement was the optimization of **XGBoost** and **Random**

Forest models, where hyperparameter tuning using **Grid Search CV** led to lower MAE and RMSE scores. Adjusting parameters such as the number of estimators, maximum depth, and learning rate significantly enhanced predictive accuracy.

For **LSTM**, initial versions struggled with extreme fluctuations in pollutant levels. To address this, we increased the number of past time steps used for training and applied dropout regularization to prevent overfitting. However, despite these improvements, LSTM still exhibited some limitations in handling short-term variations effectively.

The **ARIMA model**, which was initially considered for time-series forecasting, was found to be less effective for long-term predictions. As a result, we limited its application to short-term forecasting, where it performed more reliably. Future iterations may explore hybrid models that combine ARIMA's short-term strength with LSTM's long-term forecasting capabilities.

9.2 Data Adjustments

A key limitation was the **availability of traffic data**, which was only collected for New York. To address this, models were trained separately for each city, preventing biased conclusions. Future work may involve acquiring additional traffic datasets for other cities to provide a more comprehensive analysis.

Moreover, we experimented with **feature engineering** to enhance model interpretability. New features such as rolling averages and cyclic encoding of time-based variables were introduced, leading to improved predictions in Random Forest and XGBoost models.

9.3 Future Work

While our current models provide strong predictive capabilities, several areas remain open for improvement. One major direction for future research is the **integration of real-time air quality data** to develop a dynamic, continuously updating prediction model. This could be beneficial for smart city applications and pollution monitoring systems.

Another avenue for improvement involves exploring **hybrid models** that combine statistical and deep learning approaches. For instance, merging ARIMA with LSTM could lead to better time-series predictions by leveraging both short-term stationarity and long-term dependencies.

Expanding the dataset by incorporating **industrial emissions, population density, and land use data** could further improve model performance and provide deeper insights into pollution sources. Additionally, refining traffic-related features by integrating GPS-based congestion data may enhance the precision of pollution predictions.

Lastly, deploying an **interactive dashboard** using tools like **Streamlit or Flask** would allow policymakers and researchers to visualize air quality trends more effectively. This could enable better decision-making and public awareness regarding pollution levels.

Overall, these iterative improvements and proposed future developments will contribute to a more accurate and actionable air quality prediction framework.

10 Conclusion

10.1 Key Takeaways from the Experimentation Phase

The experimentation phase provided valuable insights into the predictive modeling of air quality based on meteorological and traffic data. The key takeaways include:

- **XGBoost outperformed other regression models**, achieving the highest predictive accuracy with an R^2 score of 0.9019 for NO_2 prediction in Chicago.
- **LSTM proved to be the best model for time-series forecasting**, effectively capturing long-term pollutant trends, with an R^2 of 0.75 for PM2.5 in Chicago.
- **Traffic data significantly influenced pollution levels**, particularly in New York City, reinforcing the importance of integrating urban mobility factors into air quality models.
- **Weather conditions had a measurable impact on pollutant concentrations**, with parameters such as wind speed and precipitation playing a crucial role in dispersing or accumulating pollutants.
- **Feature engineering techniques, such as rolling averages and cyclic encoding, improved model performance**, ensuring better trend detection and seasonality adjustment.
- **ARIMA was effective for short-term forecasting but struggled with dynamic pollution variations**, making it less suitable for long-range predictions.
- **Cross-validation and hyperparameter tuning significantly enhanced model generalizability**, preventing overfitting and ensuring robust performance on unseen data.

10.2 Contribution to Broader Project Objectives

The results of the experimentation phase directly contribute to the overarching goals of this study—understanding the impact of weather and traffic on urban air pollution and developing reliable predictive models. The key contributions are:

- **Scientific Understanding:** The study provides empirical evidence on how meteorological factors and vehicular traffic correlate with PM2.5, PM10, and NO_2 levels in major metropolitan areas.
- **Model Benchmarking:** This research establishes **XGBoost as the best model for regression tasks and LSTM as the optimal choice for time-series forecasting**.
- **Policy Implications:** The findings offer data-driven insights that can inform urban planning and environmental policies aimed at reducing pollution exposure.

- **Scalability and Reproducibility:** By making datasets and model code available on GitHub, the research supports further exploration and real-world applications.

10.3 Future Directions and Refinements

While the study successfully developed robust predictive models, several opportunities for refinement exist:

- **Expanding Geographic Scope:** Future work should incorporate data from additional cities to enhance model generalizability across diverse environmental conditions.
- **Improving Data Resolution:** Integrating real-time traffic feeds and higher-frequency air quality monitoring data could improve the accuracy of dynamic pollution forecasting.
- **Enhancing Model Performance:** Further exploration of hybrid models combining deep learning with traditional statistical approaches may yield superior results.
- **Real-Time Prediction Implementation:** Deploying the models in an interactive web-based dashboard or mobile application could facilitate real-time pollution tracking for policymakers and the public.
- **Exploring Causal Inference Techniques:** While this study identifies correlations between weather, traffic, and pollution, future research could employ causal models to establish definitive cause-effect relationships.

In conclusion, this study lays the groundwork for **data-driven environmental policymaking and advanced air quality forecasting**. The integration of machine learning techniques into urban pollution analysis not only enhances scientific understanding but also paves the way for proactive mitigation strategies, ultimately contributing to improved public health and environmental sustainability.

References

M. S. K. Abhilash, Amrita Thakur, Deepa Gupta, and B. Sreevidya. 2018. Time series analysis of air pollution in bengaluru using arima model. In *Ambient Communications and Computer Systems*, pages 413–426, Singapore. Springer Singapore.

Saba Ameer, Munam Ali Shah, Abid Khan, Houbing Song, Carsten Maple, Saif Ul Islam, and Muhammad Nabeel Asghar. 2019. Comparative analysis of machine learning techniques for predicting air quality in smart cities. *IEEE Access*, 7:128325–128338.

Mauro Castelli, Fabiana Martins Clemente, Aleš Popovič, Sara Silva, and Leonardo Vanneschi. 2020. A machine learning approach to predict air quality in california. *Complexity*, 2020(1):8049504.

Saverio De Vito, Marco Piga, Luca Martinotto, and Girolamo Di Francia. 2009. Co, no2 and nox urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization. *Sensors and Actuators B: Chemical*, 143(1):182–191.

Brian S. Freeman, Graham Taylor, Bahram Gharabaghi, and Jesse Thé and. 2018. Forecasting air quality time series using deep learning. *Journal of the Air & Waste Management Association*, 68(8):866–886. PMID: 29652217.

Huixiang Liu, Qing Li, Dongbing Yu, and Yu Gu. 2019. Air quality index and air pollutant concentration prediction based on machine learning algorithms. *Applied Sciences*, 9(19).

Metropolitan Transportation Authority. 2025. MTA Bridges and Tunnels Hourly Crossings: Beginning 2019. Accessed: February 10, 2025.

Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. 2019. The performance of lstm and bilstm in forecasting time series. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 3285–3292.

US Environmental Protection Agency. 2025. Air Quality System Data Mart [internet database]. Accessed: January 28, 2025.

Patrick Zippenfenig. 2023. Open-meteo.com weather api.