Pollution Predictor

ANKIT SARKAR[1], SHEETAL SINHA[2], RUTVIK AVINASH BARBHAI[3], and TUSAR KANTI MISHRA[4]

[1]**MANIPAL INSTITUTE OF TECHNOLOGY**

1 Abstract

Air pollution poses a severe threat to public health, ecosystems, and climate stability, driven by rapid industrialization, urbanization, and vehicular emissions. Accurate forecasting of air quality is essential for proactive mitigation strategies, public health awareness, and regulatory compliance. This study presents Pollution Predictor, a machine learning-based system that leverages historical pollution data to predict future air quality levels. The system analyzes key air quality parameters, including the Air Quality Index (AQI), PM2.5, and PM10 concentrations, using TensorFlow.js for real-time model inference. The model integrates advanced feature engineering techniques and deep learning algorithms to capture seasonal and temporal variations in pollution trends. The proposed approach is designed to provide an accessible, web-based interactive platform that enables users to visualize air quality predictions through heatmaps, enhancing interpretability and decision-making.

Recent research in air pollution prediction has explored hybrid deep learning techniques, regression-based models, and remote sensing data integration to improve forecasting accuracy. Studies have demonstrated the effectiveness of models such as CNN-LSTM hybrids, support vector regression (SVR), and ensemble learning in capturing spatial-temporal pollution patterns. However, challenges such as computational overhead, real-time deployment, and model generalization persist. The Pollution Predictor addresses these limitations by balancing predictive accuracy with computational efficiency, ensuring scalability across different geographic regions. Model performance is evaluated using key metrics such as Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) to ensure robustness and reliability. This research highlights the significance of predictive analytics in environmental monitoring and underscores the role of AI-driven solutions in sustainable urban development. The system serves as a foundation for future advancements in pollution forecasting, contributing to smart city initiatives and improved air quality management strategies.

2 Introduction

Air pollution has become a critical environmental issue, significantly affecting human health, ecosystems, and climate stability. The rapid industrialization, urbanization, and vehicular emissions have led to deteriorating air quality worldwide, necessitating effective monitoring and predictive systems to mitigate its impact. This research focuses on developing the Pollution Predictor, a machine learning-driven system that leverages historical air quality data to forecast future pollution levels. By analyzing key air quality parameters such as the Air Quality Index (AQI), PM2.5, and PM10 concentrations, the system aims to provide an early warning mechanism to reduce exposure to hazardous pollutants.

The proposed system utilizes TensorFlow.js for real-time model inference and offers a web-based interactive platform for users to access predictions seamlessly. The approach involves preprocessing large-scale environmental datasets, feature engineering, and the development of predictive models that capture seasonal and temporal variations in pollution levels. Visualization through heatmaps enhances interpretability and allows for comparative analysis over time. The system is designed to support policymakers, researchers, and the general public in mitigating pollution-related risks by providing accurate, data-driven insights.

Recent advancements in air pollution prediction have explored a variety of machine learning (ML) and deep learning (DL) techniques. For instance, hybrid models combining convolutional networks with recurrent layers have proven effective in capturing spatial-temporal correlations in air quality data [1]. Regression-based models, particularly LSTM, have demonstrated high accuracy, particularly in forecasting pollution during specific seasons, such as winter in India [2]. While significant progress has been made, challenges such as computational overhead, model generalization, and real-time deployment remain. The current research seeks to address these challenges by offering a practical and scalable solution for pollution forecasting.

Machine learning models, including support vector regression (SVR), random forests (RF), and deep learning models like CNN, LSTM, and RNN, have been applied to air quality prediction, yielding promising results. However, integrating real-time monitoring and enhancing the energy efficiency of these models remains an ongoing challenge [3][4]. Moreover, hybrid approaches and the incorporation of remote sensing technologies, such as satellite imagery, offer opportunities for improving prediction accuracy and emission tracking, especially in indus-

trial zones [5]. Despite these advancements, the real-time deployment of these models continues to be a barrier to widespread implementation [6].

This study emphasizes the potential of data science in solving pressing environmental concerns and highlights the importance of predictive analytics in public health management. By combining machine learning techniques with real-world data, the Pollution Predictor project demonstrates a step forward in improving air quality forecasting and supporting sustainable urban development.

Gang Chen, Shen Chen, Dong Li Cai Chen (2025), "Hybrid Deep Learning Prediction," Scientific Report – This study introduces a convolution-sequence hybrid model integrating KNN, STA-ConvNet, and ConvLSTM for air pollution forecasting. Using air quality data from Beijing (2020-2021), the model efficiently captures spatial-temporal correlations to enhance prediction accuracy. The study applies preprocessing techniques like linear interpolation and Min-Max normalization, with data split into training, validation, and testing sets. Evaluation using RMSE and accuracy metrics shows improvements in multi-step and trend predictions. While the model effectively reduces data redundancy and prevents gradient vanishing, it requires high computational power and tuning for different regions. The study highlights the importance of deep learning in air quality modeling but lacks real-time deployment discussion.

Sweta Dey et al. (2024), "Apict: Air Pollution Epidemiology Using Green AQI Prediction During Winter Seasons in India," IEEE Transactions on Sustainable Computing – This paper focuses on air quality forecasting in Indian cities using a regression-based hybrid model. Leveraging CPCB data from 2015-2020, it builds a customized dataset of over 5 billion samples to train machine learning (RF, KNN, SVR) and deep learning (CNN, MLP, LSTM, RNN) models. The study emphasizes energy-aware features, model pruning, and real-time energy monitoring for improved efficiency. Results indicate high accuracy (98.53 percent) with LSTM, striking a balance between performance and energy consumption. However, its winter-focused dataset limits generalizability to other seasons. While the paper advances sustainable air quality forecasting, computational overhead remains a challenge.

[1] Amit Mittal et al. (2024), "Advancements in Air Pollution Prediction and Classification Models: Exploring Emerging Frontiers," ICACCS – This research provides a systematic literature review of AI-based urban air quality prediction techniques. It compares machine learning and deep learning models using RMSE, MAE, PLCC, and R^2 evaluation metrics across various case studies. The study highlights the effectiveness of hybrid models combining CNN, LSTM, and Graph Neural Networks (GNNs) for accurate pollution classification and prediction. Additionally, it offers a taxonomy of AI-based models to support future research. However, the paper lacks original experimental validation, relying primarily on past studies. It also discusses the limitations of real-time implementation and data challenges in large-scale deployment.

[2] Mare Srbinovska et al. (2024), "Comprehensive Study on Air Pollution Prediction in North Macedonia: Insights from LASSO Modeling," IcETRAN – This study analyzes air quality trends in North Macedonia using historical data and meteorological parameters. It applies LASSO regression to handle high-dimensional data, identifying key predictors of air pollution levels. The experimental design integrates tree-based and linear regression models to compare pre- and post-COVID-19 pollution trends. The study provides robust insights into pollution sources and their correlations. However, it lacks integration with real-time monitoring and is geographically limited. Additionally, it does not compare LASSO's performance with deep learning models, which could offer enhanced accuracy.

[3] Julien Vachon et al. (2024), "Do Machine Learning Methods Improve Prediction of Ambient Air Pollutants with High Spatial Contrast? A Systematic Review," Environmental Research – This systematic review examines 38 studies comparing ML-based and statistical models for air pollution prediction. It evaluates models based on R^2 and RMSE metrics, focusing on pollutants such as NO2, BC, and UFP across various locations. The study finds ML models outperform traditional statistical methods in spatial-temporal air pollution modeling. However, it notes limited integration of deep learning techniques like LSTM. Additionally, while the review provides valuable comparative insights, it does not contribute original model development. The study also lacks discussion on the real-world deployment of ML models in pollution monitoring.

[4] Armin Nakhjiri Ata Abdollahi Kakroodi (2024), "Air Pollution in Industrial Clusters: A Comprehensive Analysis and Prediction Using Multi-Source Data," Ecological Informatics – This research focuses on industrial pollution forecasting using Sentinel-5P satellite imagery and auxiliary datasets. It employs the Exponential Smoothing Model (ESM) for short-term predictions while categorizing industrial zones based on emission types. The study highlights the benefits of integrating remote sensing and predictive modeling for improved emission tracking. Seasonal forecasting further enhances pollution control strategies. However, the model lacks a comparison with deep learning techniques that could improve accuracy. Moreover, the study is region-specific (Tehran province), limiting its broader applicability.

These studies collectively contribute to the advancement of air pollution prediction, highlighting the role of hybrid deep learning models, regression techniques, systematic reviews, and remote sensing in environmental monitoring. Despite significant progress, challenges such as computational overhead, model generalization, and real-time implementation remain areas for future research and innovation.

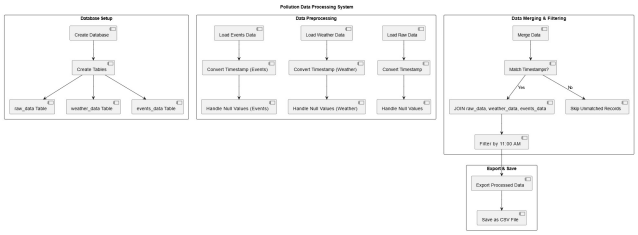

Figure 1. Block Diagram

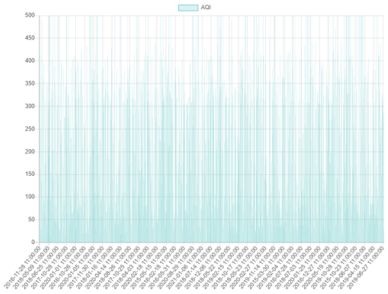

Figure 2. Graph for AQI vs Time

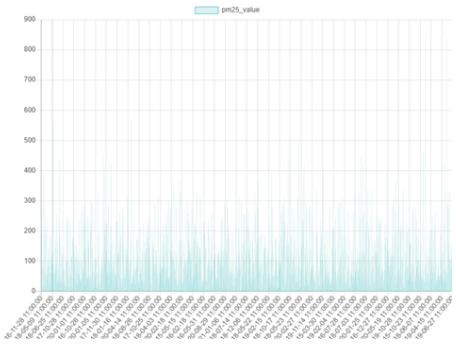

Figure 3. Graph for PM 2.5 vs Time

3 Proposed Method

3.1 Data Pre-Processing

There are multiple datasets:

- One dataset contains values of weather parameters (temperature, humidity, etc.) for one location over a period of 10 years.
- Another dataset contains values of pollution parameters (AQI, PM2.5, and PM10) over the last 10 years for the same location.
- The third dataset contains values of congestion index and special events occurring in the same location over a period of 10 years.

Different sources were used to compile the dataset:

- The pollution dataset includes recorded air quality parameters from government and independent monitoring stations.
- The weather dataset is compiled from meteorological observations and historical weather databases.
- The traffic dataset incorporates congestion indices and records of special events from transportation departments and event organizers.

By integrating these sources, we ensure a comprehensive dataset that effectively captures the interactions between environmental conditions, pollution levels, and human activities.

3.2 Data Transformation

To ensure data consistency and improve visualization:

- Data was sorted based on parameter values.
- Repeating values for different days were normalized.
- A new parameter, "special events," was added to analyze its effect on pollution levels.

For better data visualization:

- Graphs were created with:
  - **X-axis:** Date
  - **Y-axis:** The measured parameter (AQI, PM2.5, PM10, Temperature)
- The **Chart.js** library was used to generate the graphs.

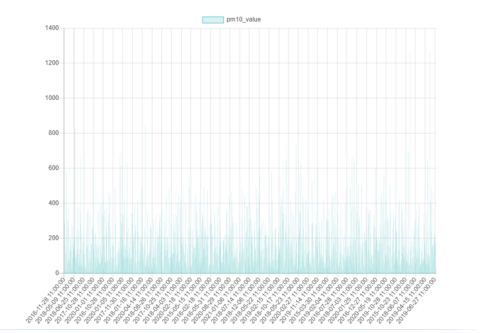

Figure 4. Graph for PM 10 vs Time

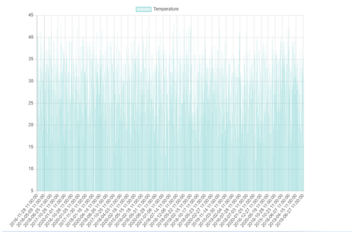

Figure 5. Graph for Temperature vs Time

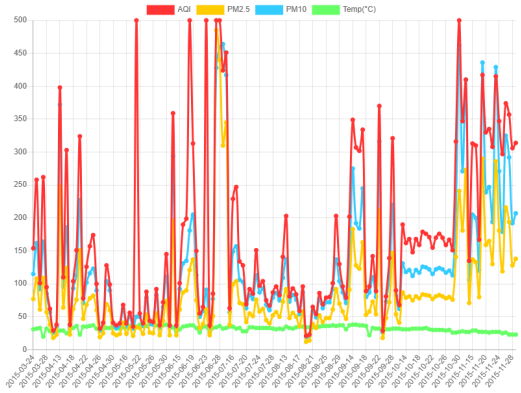

Figure 6. Line Chart of Multiple Parameters Trends Over Time

3.3 Flowchart

The flowchart presented illustrates a structured approach for managing and processing datasets in a database system for a mini-project. The workflow begins with verifying the existence of the database, followed by the creation of the mini project database if it does not exist. Subsequently, three tables—raw data, weather data, and events data—are initialized to store relevant information. Data ingestion follows, where raw data is imported into the raw data table, ensuring that timestamps are standardized and missing values are appropriately handled. The same preprocessing steps are applied to the weather data and events data tables to maintain data consistency. These preprocessing steps ensure data integrity, enabling accurate analysis in subsequent stages.

Once data preparation is complete, the merging process is executed based on timestamp alignment. Records with unmatched timestamps are excluded to maintain data reliability. The raw data table is joined with the weather data and events data tables, ensuring a comprehensive dataset for further analysis. Additionally, a filtering mechanism is applied to extract data recorded at 11:00 AM, optimizing the dataset for specific analytical objectives. Finally, the processed data is exported and stored as a CSV file, facilitating ease of access and further computational processing. This systematic approach enhances data accuracy, integrity, and usability, making it well-suited for analytical and research applications.

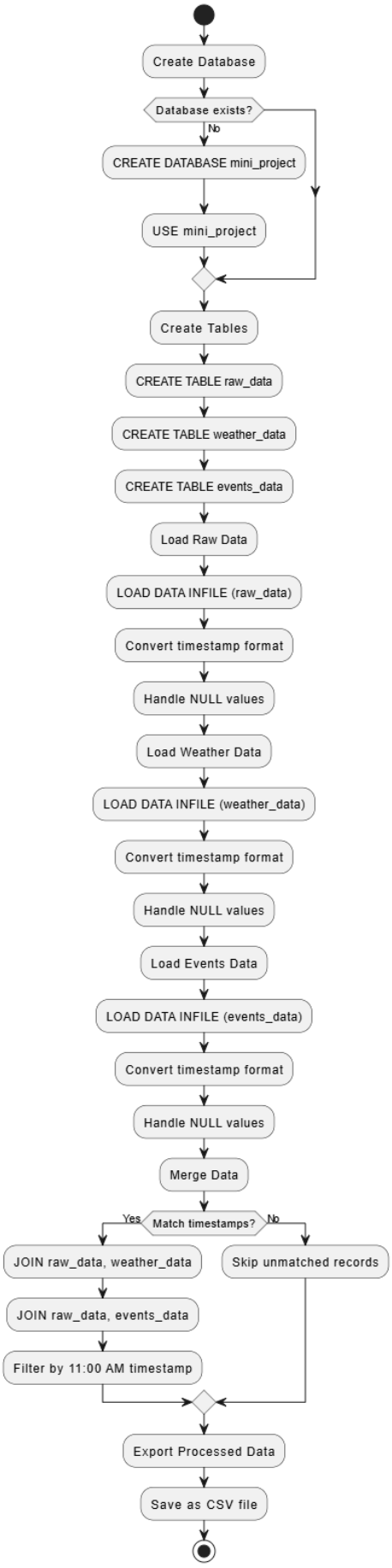

Figure 7. Flowchart of Data Processing and Transformation

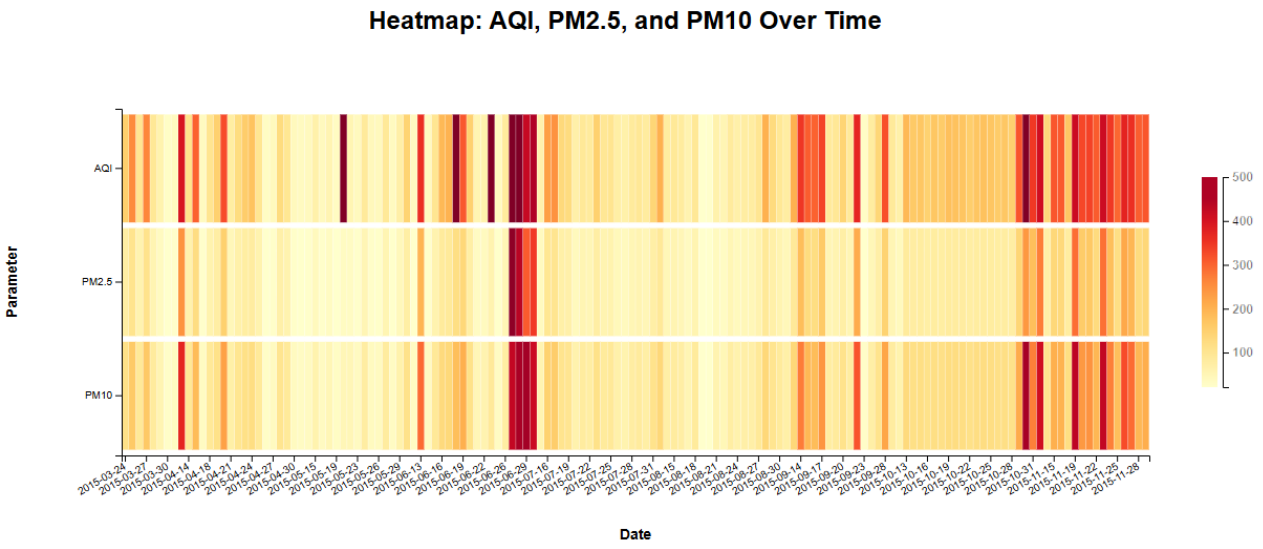

Figure 8. Heatmap showing Correlation Between Parameters

4 Result Analysis

4.1 Heatmap

To identify correlations among different parameters, a heatmap was generated as shown in figure 8. The heatmap visualizes the variations in Air Quality Index (AQI), PM2.5, and PM10 levels over time, providing a comprehensive representation of pollution trends. The x-axis represents the date, while the y-axis categorizes the parameters, namely AQI, PM2.5, and PM10. The color intensity corresponds to the concentration levels, with lighter shades indicating lower values and darker shades representing higher pollutant concentrations. Notable spikes in AQI, PM2.5, and PM10 can be observed at various intervals, suggesting periods of high pollution. The structured visualization facilitates the identification of temporal patterns and anomalies, which can be critical for environmental analysis and policy making.

4.2 Conclusion

In this project, we focused on preprocessing the dataset to ensure data quality, consistency, and usability for visualization and analysis. The key preprocessing steps included:

- Handling missing values and standardizing data formats.
- Removing outliers and optimizing data volume.
- Resolving encoding issues to ensure a clean and structured dataset.

By refining the dataset, we improved visualization clarity, identified trends, and reduced errors. Proper preprocessing also made the data suitable for further statistical analysis and machine learning applications.

5 References

[1]G. Chen, S. Chen, D. Li, and C. Chen, "Hybrid Deep Learning Prediction," Scientific Report, 2025.

[2]S. Dey et al., "Apict: Air Pollution Epidemiology Using Green AQI Prediction During Winter Seasons in India," IEEE Transactions on Sustainable Computing, 2024.

[3] A. Mittal et al., "Advancements in Air Pollution Prediction and Classification Models," ICACCS, 2024.

[4] M. Srbinovska et al., "Comprehensive Study on Air Pollution Prediction in North Macedonia," IcETRAN, 2024.

[5]A. Nakhjiri and A. Abdollahi Kakroodi, "Air Pollution in Industrial Clusters: A Comprehensive Analysis and Prediction Using Multi-Source Data," Ecological Informatics, 2024.

[6]J. Vachon et al., "Do Machine Learning Methods Improve Prediction of Ambient Air Pollutants with High Spatial Contrast? A Systematic Review," Environmental Research, 2024.

References