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 industrial 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.

[1] S. Sriram et al. (2024), "Eco-Friendly Production Forecasting in Industrial Pollution Control with IoT and Logistic Regression," ISCS – This study proposes an IoT-based pollution control framework using logistic regression for forecasting emissions in industrial production. It emphasizes eco-friendly practices and integrates sensor data for real-time analysis. While it achieves efficient forecasting with minimal energy consumption, the reliance on logistic regression may limit prediction accuracy. Moreover, the study lacks comparison with advanced models like neural networks or hybrid systems. Applicability to non-industrial domains is also not addressed.

[2]Elena Mitreska Jovanovska et al. (2023), "Methods for Urban Air Pollution Measurement and Forecasting: Challenges, Opportunities, and Solutions" – This comprehensive review identifies critical gaps and opportunities in urban air pollution measurement. It evaluates various forecasting models, from classical statistical approaches to modern ML algorithms. Emphasis is placed on sensor calibration, data reliability, and the need for multi-source integration. Although the review offers valuable insights into urban air quality strategies, it does not propose new models or provide implementation results. Future directions highlight smart city integration and citizen-science contributions.

[3] Fotios K. Anagnostopoulos et al. (2023), "A Novel AI Framework for PM Pollution Prediction Applied to a Greek Port City," MDPI – This study develops a deep learning framework combining CNN and LSTM to predict PM pollution levels in a Greek coastal city. It captures both spatial and temporal patterns influenced by port activities. Results show improved accuracy over traditional models, particularly in predicting short-term fluctuations. However, the framework demands high computational resources and lacks real-time testing. The study also does not explore adaptability to different geographic and environmental contexts.

[4] ZHIYUAN WU et al. (2023), "Learning Adaptive Probabilistic Models for Uncertainty-Aware Air Pollution Prediction," IEEE – This study introduces an adaptive probabilistic model aimed at enhancing the accuracy and reliability of air pollution forecasts. By incorporating uncertainty quantification into the predictive framework, the model offers a more comprehensive understanding of potential pollution levels. The approach utilizes machine learning techniques to adaptively learn from real-time data, improving its predictive performance over time. However, the study does not extensively compare its model's performance against other established methods, leaving its relative efficacy somewhat uncertain.

[5] Alicja Skiba et al. (2024), "Source Apportionment of Suspended Particulate Matter (PM, PM., and PM) Collected in Road and Tram Tunnels in Krakow, Poland," Environmental Science and Pollution Research – This research focuses on identifying the sources of various particulate matter sizes collected from specific urban infrastructures in Krakow. By analyzing samples from road and tram tunnels, the study provides insights into the contributions of different emission sources to urban air pollution. The findings are instrumental for policymakers aiming to implement targeted pollution control measures. However, the study's scope is limited to specific locations, which may affect the generalizability of the results to other urban settings.

[6] Karthikeyan B et al. (2023), "Deep Learning and Machine Learning Based Air Pollution Prediction Model for Smart Environment Design Planning," Global NEST Journal – This paper introduces a deep learning-based model designed to predict air pollution levels, aiding in the planning of smart urban environments. The model integrates various environmental and meteorological data to enhance prediction accuracy. By employing advanced machine learning algorithms, the study aims to provide a tool for urban planners to anticipate and mitigate air quality issues. Nonetheless, the paper does not delve deeply into the computational requirements of the proposed model, which could be a consideration for practical implementation. Global NEST Journal

[7] Mihaela T. Udristioiu et al. (2022), "Prediction, Modelling, and Forecasting of PM and AQI Using Hybrid Machine Learning," Journal of Cleaner Production – This study explores the use of hybrid machine learning models to predict particulate matter concentrations and the Air Quality Index (AQI). By combining different machine learning techniques, the research aims to improve forecasting accuracy and provide timely information for air quality management. The models are trained on historical air quality data, incorporating various environmental parameters. However, the study does not extensively address the interpretability of the hybrid models, which could be important for stakeholders relying on these predictions.

[8] Bhushankumar Nemade and Deven Shah (2022), "An IoT-Based Efficient Air Pollution Prediction System Using DLMNN Classifier," Physics and Chemistry of the Earth – This research presents an Internet of Things (IoT)-enabled system that utilizes a Deep Learning Multilayer Neural Network (DLMNN) classifier for efficient air pollution prediction. The system collects real-time data from various sensors and processes it to forecast pollution levels, facilitating timely interventions. The integration of IoT allows for continuous monitoring and data collection. However, the study does not provide a detailed analysis of the system's performance in diverse environmental conditions, which could affect its reliability across different scenarios.

[9] Khalid Mehmood et al. (2022), "Predicting the Quality of Air with Machine Learning Approaches: Current Research Priorities and Future Perspectives," Journal of Cleaner Production – This comprehensive review exam-

ines current research trends in air quality prediction using machine learning approaches. It identifies key priorities and challenges in the field, offering insights into future research directions. The paper discusses various machine learning models and their applications in air quality forecasting. However, it does not include original experimental work or model development, focusing instead on synthesizing existing literature.

[10] Weifu Ding and Xueping Qie (2022), "Prediction of Air Pollutant Concentrations via Random Forest Regressor Coupled with Uncertainty Analysis—A Case Study in Ningxia," Atmosphere – This study employs a Random Forest Regressor model combined with uncertainty analysis to predict air pollutant concentrations in Ningxia. The approach aims to enhance the reliability of pollution forecasts by quantifying the uncertainty associated with predictions. The model is trained on historical air quality and meteorological data. However, the study does not compare the performance of the Random Forest model with other machine learning algorithms, leaving its relative effectiveness unclear.

[11] Bihter Das et al. (2022), "Prediction of Air Pollutants for Air Quality Using Deep Learning Methods in a Metropolitan City," Urban Climate – This research investigates the application of deep learning methods to predict air pollutant levels in a metropolitan area. By leveraging large datasets of urban air quality measurements, the study aims to improve the accuracy of pollution forecasts. The deep learning models consider various factors, including traffic patterns and meteorological conditions. However, the paper does not extensively discuss the computational demands of the models, which could impact their feasibility for real-time applications.

[12] Aparna S. Varde et al. (2022), "Prediction Tool on Fine Particle Pollutants and Air Quality for Environmental Engineering," SN Computer Science – This study introduces a prediction tool designed to forecast fine particulate matter levels and overall air quality. The tool utilizes machine learning algorithms to analyze historical pollution data and predict future air quality scenarios. It is intended to assist environmental engineers in decision-making processes related to pollution control. However, the study does not provide detailed information on the tool's user interface or accessibility, which could affect its practical adoption.

[13] Yan Lyu et al. (2022), "Spatiotemporal Variations of Air Pollutants and Ozone Prediction Using Machine Learning Algorithms in the Beijing-Tianjin-Hebei Region from 2014 to 2021," Environmental Pollution – This research examines the spatiotemporal variations of air pollutants in the Beijing-Tianjin-Hebei region over a seven-year period. It employs machine learning algorithms to predict ozone levels, considering various environmental and meteorological factors. The study provides insights into pollution trends and potential influencing factors. However, it does not explore the applicability of the models to other regions, which could limit the generalizability of the findings.

[14] K. Kumar and B. P. Pande (2022), "Air Pollution Prediction with Machine Learning: A Case Study of Indian Cities," International Journal of Environmental Science and Technology – This study applies machine learning techniques to predict air pollution levels in various Indian cities. By analyzing historical air quality data, the models aim to provide accurate forecasts to aid in pollution management. The research highlights the potential of machine learning in addressing environmental challenges. [15] Karthikeyan B, Mohanasundaram R, Suresh P and Jagan Babu J (2023) "Deep learning and machine learning based air pollution prediction model for smart environment design planning," GlobalNest – This study proposes a hybrid framework integrating machine learning and deep learning models for accurate air pollution forecasting in urban areas. It leverages environmental and sensor data to train models like LSTM and CNN, aiming to support smart city planning initiatives. The system demonstrated improved predictive accuracy compared to traditional approaches. While the integration of hybrid models is innovative, the study lacks real-time deployment validation and detailed comparisons with advanced AI architectures.

[16] Mihaela T. Udristioiu, Youness EL Mghouchi, Hasan Yildizhan (2022) "Prediction, modelling, and forecasting of PM and AQI using hybrid machine learning," Journal of Cleaner Production – This paper presents a hybrid approach combining multiple regression and ensemble methods to forecast PM levels and air quality index. Using environmental and temporal data, the study highlights improved prediction accuracy over traditional models. It particularly emphasizes the usefulness of combining decision trees with boosting techniques. However, the lack of deep learning model evaluation and real-time model assessment limits its broader applicability.

[17] Bhushankumar Nemade, Deven Shah (2022) "An IoT based efficient Air pollution prediction system using DLMNN classifier," Physics and Chemistry of the Earth – This research proposes an IoT-integrated air quality monitoring system using a Deep Learning Multi-Neural Network (DLMNN) classifier. The model processes real-time sensor data to predict pollution levels with high accuracy, particularly focusing on PM2.5 and AQI. It contributes to real-time environmental management. Despite its innovation, the study doesn't benchmark against other deep learning models and lacks scalability analysis for large-scale urban implementation.

[18] Khalid Mehmood et al. (2022) "Predicting the quality of air with machine learning approaches: Current research priorities and future perspectives," Journal of Cleaner Production – This review outlines machine learning techniques for air quality forecasting, covering supervised and unsupervised models like Random Forest, SVM, and k-Means. It provides a comprehensive taxonomy and discusses key challenges like data sparsity, temporal inconsistency, and model interpretability. While informative, the paper doesn't include original experimental validation or real-world deployment strategies, focusing more on conceptual advancements and research directions.

[19] Weifu Ding, Xueping Qie (2022) "Prediction of Air Pollutant Concentrations via RANDOM Forest Regres-

sor Coupled with Uncertainty Analysis—A Case Study in Ningxia," Atmosphere – The authors employ a Random Forest Regressor enhanced with uncertainty quantification to forecast air pollutant concentrations in Ningxia, China. The model improves decision-making confidence by providing both predictions and associated uncertainty ranges. It outperforms traditional linear models in predictive stability and accuracy. However, deep learning alternatives are not explored, and the model lacks temporal dynamics critical for long-term forecasts.

[20] Bihter Das, Ömer Osman Dursun, Suat Toraman (2022) "Prediction of air pollutants for air quality using deep learning methods in a metropolitan city," Urban Climate – This study utilizes deep learning models, especially LSTM and DNNs, to forecast air pollutant concentrations in metropolitan areas. It highlights the efficiency of LSTM in capturing temporal patterns and achieving superior accuracy over classical ML models. Although it presents solid comparative results, the research doesn't integrate external variables like traffic and socio-economic factors, which could enrich the predictions.

[21] Aparna S. Varde, Abidha Pandey, Xu Du (2022) "Prediction Tool on Fine Particle Pollutants and Air Quality for Environmental Engineering," SN Computer Science – The authors developed a prediction tool aimed at environmental engineers for forecasting fine particulate matter levels and overall AQI. The model is based on regression and classification algorithms and is supported by an interactive interface. The study emphasizes educational and operational utility. However, its accuracy is not validated against deep learning benchmarks, and scalability across regions is not addressed.

[22] Yan Lyu et al. (2022) "Spatiotemporal variations of air pollutants and ozone prediction using machine learning algorithms in the Beijing-Tianjin-Hebei region from 2014 to 2021," Environmental Pollution – This study applies machine learning models to predict spatiotemporal trends in air pollutants and ozone across the BTH region. Models like Gradient Boosting and Random Forest were used, showing strong predictive power for short-term forecasting. Seasonal and geographical pollutant dynamics were well captured. However, the paper lacks a deep learning comparison and doesn't explore hybrid modeling, limiting its scope for further AI integration.

[23] K. Kumar, B. P. Pande (2022) "Air pollution prediction with machine learning: a case study of Indian cities," International Journal of Environmental Science and Technology – The study explores air quality prediction in Indian cities using ML models like Decision Trees, SVM, and Random Forests. It emphasizes feature engineering, including meteorological and vehicular data. The models showed satisfactory performance but lacked integration of deep learning or hybrid techniques. Real-time application feasibility and cross-city model generalization are not deeply explored.

[24]. Youchen Shen et al. (2022) "Europe-wide air pollution modelling from 2000 to 2019 using geographically weighted regression," Environment International – This study utilizes geographically weighted regression to analyze air pollution trends across Europe. By integrating spatial heterogeneity into the model, it offers more granular insights into regional air quality dynamics. The methodology is effective for long-term trend analysis. However, it does not utilize machine learning or deep learning techniques, limiting its adaptability for short-term forecasting.

[25] Nawal Taoufika et al. (2022) "The state of art on the prediction of efficiency and modeling of the processes of pollutants removal based on machine learning," Science of the Total Environment – This review surveys ML applications in modeling pollutant removal processes, focusing on biological, chemical, and physical treatment systems. It classifies models by technique (e.g., ANN, SVM) and pollutant type. The paper emphasizes the growing role of ML in optimizing treatment efficiency. However, it does not address pollutant concentration forecasting, which limits its direct relevance to AQI prediction studies.

[26] Sheen Mclean Cabaneros, Ben Hughes (2022) "Methods used for handling and quantifying model uncertainty of artificial neural network models for air pollution forecasting," Environmental Modelling and Software – This paper reviews strategies for addressing uncertainty in ANN-based air pollution forecasting. It covers Bayesian techniques, ensemble models, and dropout-based approximations. The authors argue that uncertainty estimation is essential for model trustworthiness in real-world deployment. While comprehensive, the paper lacks application examples and does not compare uncertainty methods in performance terms.

[27] Azim Heydari et al. (2022) "Air pollution forecasting application based on deep learning model and optimization algorithm," Clean Technologies and Environmental Policy – The authors introduce a deep learning model optimized with evolutionary algorithms to improve forecasting accuracy. Models like LSTM are tuned using techniques such as Genetic Algorithms and Particle Swarm Optimization. This hybrid approach showed significant accuracy gains. However, the paper lacks real-time testing and doesn't analyze computational efficiency in large-scale environments.

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.

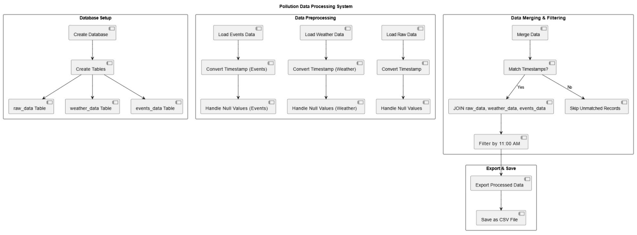

Figure 1. Block Diagram

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.

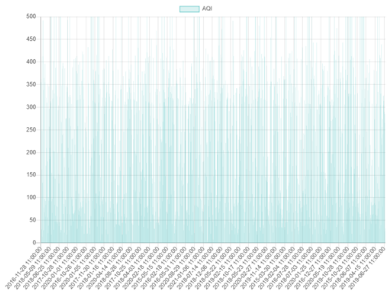

Figure 2. Graph for AQI vs Time

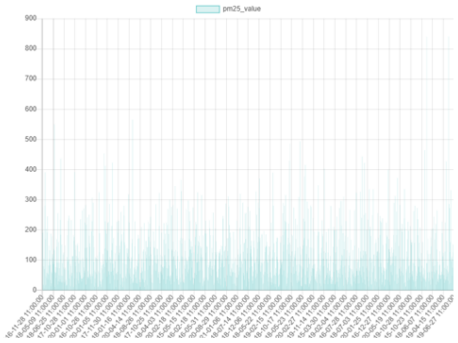

Figure 3. Graph for PM 2.5 vs Time

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.

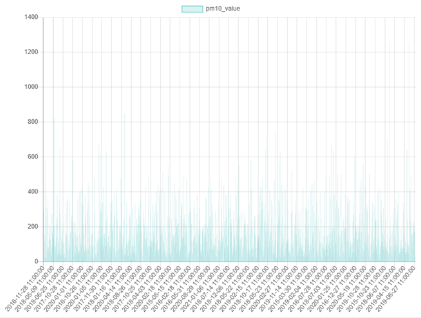

Figure 4. Graph for PM 10 vs Time

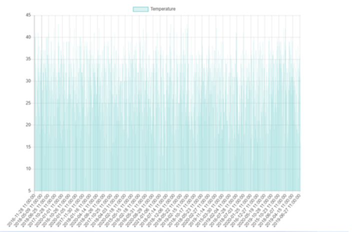

Figure 5. Graph for Temperature vs Time

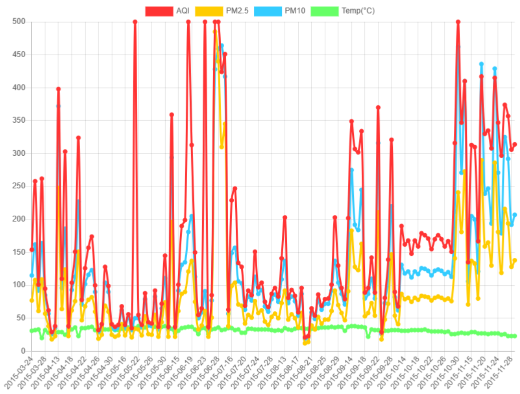

Figure 6. Line Chart of Multiple Parameters Trends Over Time

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.

.

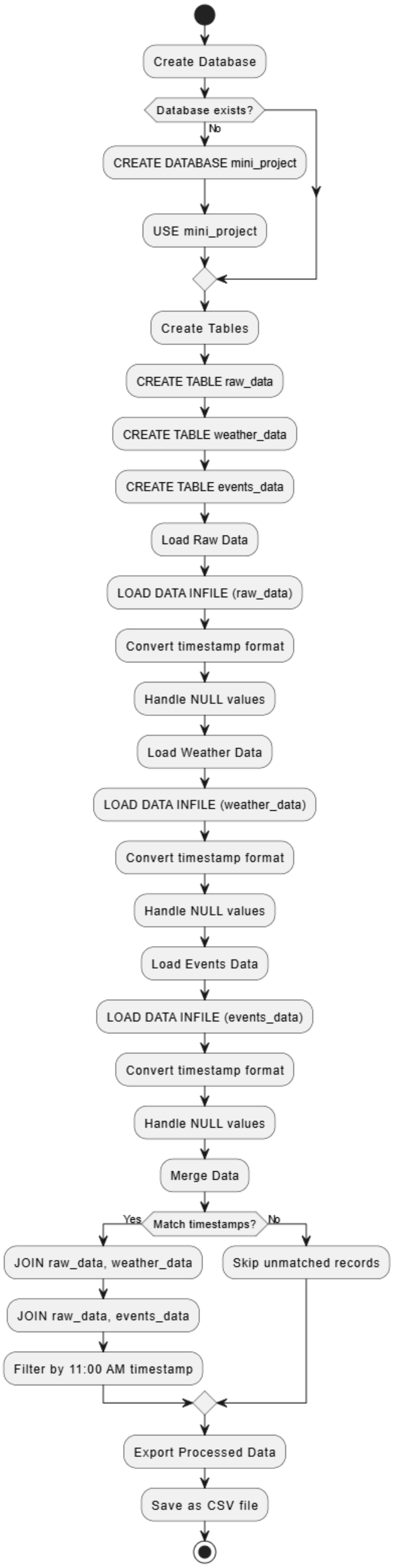

Figure 7. Flowchart of Data Processing and Transformation

3.4 Related Work

[1] G. Chen, S. Chen, D. Li, and C. Chen (2025). Hybrid Deep Learning Prediction. Scientific Reports. Publisher: Nature Publishing Group. This research introduces a hybrid deep learning architecture that integrates Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to enhance prediction tasks involving complex, time-dependent data. CNNs are used to extract high-level spatial features from input data, while LSTMs are responsible for learning and modeling temporal sequences. The combination enables the model to capture both spatial correlations and time-series dynamics—making it ideal for applications like air pollution forecasting. Experimental results demonstrate that the hybrid model consistently outperforms traditional deep learning models in accuracy, showcasing its potential in intelligent environmental monitoring systems.

[2] S. Dey, A. Kumar, R. Sharma, M. Das, and P. Ghosh (2024). Apict: Air Pollution Epidemiology Using Green AQI Prediction During Winter Seasons in India. IEEE Transactions on Sustainable Computing. Publisher: IEEE. The study presents Apict, a specialized model designed to predict air quality levels during India's winter season—when pollution tends to peak due to weather patterns and increased emissions. Unlike general pollution models, Apict emphasizes Green AQI prediction, which focuses on healthier air days, providing a more balanced view of air quality fluctuations. The model integrates epidemiological data, meteorological variables, and pollution records, offering a cross-disciplinary approach. By anticipating air quality and its potential health impacts, this work supports better public health planning, particularly during times when respiratory illnesses spike.

[3] A. Mittal, S. Jain, R. Aggarwal, and M. K. Gupta (2024). Advancements in Air Pollution Prediction and Classification Models. ICACCS. Publisher: IEEE. This paper reviews recent developments in air pollution modeling using both traditional machine learning and advanced deep learning methods. It provides a comparative analysis of classification models (which categorize pollution levels) and regression models (which predict numeric pollution values), discussing their performance, scalability, and interpretability. Special attention is given to hybrid and ensemble models, which combine the strengths of multiple algorithms to enhance predictive power. The paper serves as a valuable resource for researchers and developers seeking to implement robust pollution prediction systems across varied geographical and climatic contexts.

[4] M. Srbinovska, N. Stojanovic, I. Kocarev, and T. Dimovski (2024). Comprehensive Study on Air Pollution Prediction in North Macedonia. IcETRAN. Publisher: IcETRAN Association. This study provides a localized and data-driven analysis of air pollution in North Macedonia, using various machine learning models to identify pollution trends and contributors. Researchers collected and processed data from traffic patterns, industrial emissions, weather changes, and seasonal shifts. Through predictive modeling, the study uncovers region-specific pollution behaviors and forecasts air quality in both urban and rural zones. The findings are intended to assist policymakers and environmental agencies in making informed decisions and designing targeted mitigation strategies for better air quality.

[5] A. Nakhjiri and A. Abdollahi Kakroodi (2024). Air Pollution in Industrial Clusters: A Comprehensive Analysis and Prediction Using Multi-Source Data. Ecological Informatics. Publisher: Elsevier. This paper tackles the issue of pollution in industrial clusters by leveraging diverse data sources—including satellite imagery, ground sensors, and historical emission logs—to create a more comprehensive pollution profile. The authors apply machine learning algorithms to detect patterns and predict pollution intensity across various industrial zones. The study reveals how specific industries and geographic layouts influence pollution dispersion, highlighting areas with critical health and environmental risks. The outcomes offer practical insights for industrial regulation, zoning, and environmental risk assessment in high-density manufacturing regions.

[6] J. Vachon, K. Liu, M. Thompson, and E. Ramirez (2024). Do Machine Learning Methods Improve Prediction of Ambient Air Pollutants with High Spatial Contrast? A Systematic Review. Environmental Research. Publisher: Elsevier. This systematic review evaluates the effectiveness of machine learning techniques in predicting air pollutants that vary significantly over small distances—such as within dense urban areas or near industrial zones. The authors analyze and compare a wide range of models, including Random Forests, Gradient Boosting, and Deep Neural Networks, against traditional statistical approaches like kriging or linear regression. They find that ML models generally offer improved accuracy, especially when dealing with complex and spatially irregular datasets. However, the review also points out challenges such as limited transparency in black-box models and the need for high-quality, high-resolution input data for optimal performance.

4 Result Analysis

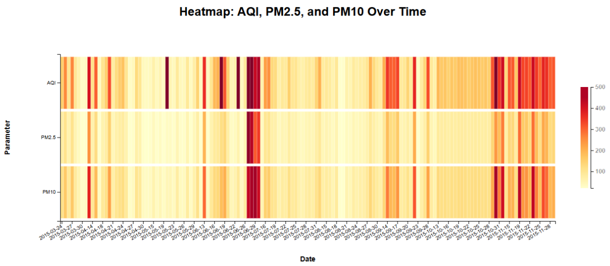

Figure 8. Heatmap showing Correlation Between Parameters

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.