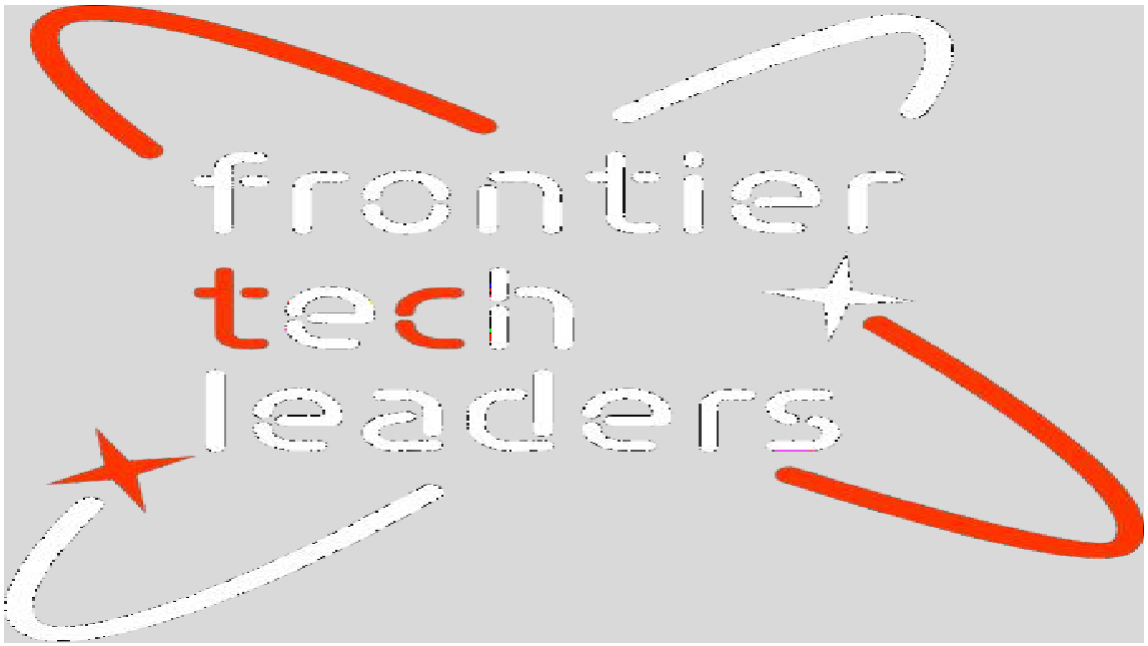

Project Title: Predicting Air Quality for Sustainable Urban Living

Prepared by: Group 4

1. Dawit Teklebrehan
2. Shemils Tilahun
3. Robel Ermiyas
4. Robel Roba
5. Tewodros Gebretsadik
6. Negassa Retta

April 2025

Table of Contents

Literature Review

1. Introduction

Urban air pollution has emerged as one of the leading global environmental concerns. Rapid urbanization and industrial activities have contributed to deteriorating air quality in cities worldwide. This research project aims to address the issue by developing predictive models that can forecast air quality using machine learning and diverse datasets. The review of existing literature is crucial to understand how previous studies have approached air pollution prediction and to identify opportunities where our project can contribute.

2. Organization

The literature is reviewed under the following thematic areas:
- Public health and environmental implications of air pollution
- Advances in machine learning for air quality prediction
- Integration of sensor networks and remote sensing for data collection

3. Summary and Synthesis

A) Health and Environmental Impacts:

Studies by the World Health Organization (2021) indicate that long-term exposure to fine particulate matter (PM2.5) and nitrogen dioxide (NO2) is directly linked to respiratory, cardiovascular, and neurological disorders. The Global Burden of Disease study also estimates that air pollution causes more than 7 million premature deaths annually. These statistics underscore the urgent need for predictive and preventive tools.

B) Machine Learning in Air Quality Prediction:

Zhang et al. (2018) developed an LSTM-based model to forecast PM2.5 levels in Beijing, showcasing its superior accuracy compared to traditional regression models. Similarly, Gupta et al. (2020) employed ensemble learning methods like Random Forests and Gradient Boosting, integrating meteorological features to enhance prediction reliability.

C) Sensor Networks and Remote Sensing:

Li et al. (2019) emphasized the synergy between low-cost sensor networks and remote sensing data such as Aerosol Optical Depth (AOD) from NASA MODIS to provide high spatial coverage. The fusion of ground and satellite data significantly improves modeling granularity.

4. Conclusion

The literature demonstrates a clear relationship between environmental data and predictive accuracy. While traditional statistical models offer baseline predictions, machine learning—especially time-series deep learning—enables more precise, dynamic forecasting. Our project extends the current knowledge by merging diverse datasets and technologies into an operational forecasting dashboard.

5. References

- World Health Organization. (2021). *Ambient air pollution: A global assessment of exposure and burden of disease*. World Health Organization.

- Zhang, Y., Wang, S., & Li, H. (2018). PM2.5 forecasting using LSTM neural networks. *Atmospheric Environment, 199*, 21–29. https://doi.org/10.1016/j.atmosenv.2018.01.012

- Gupta, R., Sharma, A., & Verma, R. (2020). Machine learning approaches to forecasting air pollution in urban cities. *Environmental Monitoring and Assessment, 192*(4), 205. https://doi.org/10.1007/s10661-020-8142-3

- Li, X., Zhou, Y., & Chen, M. (2019). Combining IoT sensor networks and remote sensing for air quality monitoring. *Sensors, 19*(18), 3874. https://doi.org/10.3390/s19183874

Data Research

1. Introduction

Accurate air quality prediction depends on the availability of comprehensive, high-resolution, and multi-source data. The goal of this section is to identify, describe, and analyze data sources that will be used to develop predictive models for urban air pollution. Data exploration is essential for understanding patterns, preparing input features, and ensuring the integrity of machine learning models.

2. Organization

This data research is organized into three main parts:

- Data source description

- Preprocessing and feature engineering

- Insight generation and correlation analysis

3. Data Description

We rely on a combination of publicly available datasets from reputable sources:

- OpenAQ: Offers hourly air pollution readings (PM2.5, PM10, CO, NO2, SO2, O3) across thousands of global cities.

- NASA MODIS: Provides Aerosol Optical Depth (AOD) measurements through satellite remote sensing.

- NOAA (National Oceanic and Atmospheric Administration): Meteorological data including temperature, humidity, wind speed, pressure, and rainfall.

These datasets are accessed in CSV and NetCDF formats and contain tens of thousands of data points across spatial and temporal dimensions.

4. Data Analysis and Insights

After initial preprocessing, we conducted exploratory data analysis to detect trends, gaps, and correlations:

- Missing data was addressed using time-based interpolation and forward/backward filling.

- Pollutants like PM2.5 showed seasonal spikes in colder months due to low atmospheric dispersion.

- Wind speed negatively correlates with pollution concentration, reinforcing its role in dispersion modeling.

- AOD values from satellite observations showed strong positive correlation ($r > 0.7$) with PM2.5, validating the integration of remote sensing.

We also created derived features such as AQI (Air Quality Index) and wind direction bins to improve model learning.

5. Conclusion

The combined dataset—spanning ground sensors and satellite inputs—provides a robust foundation for machine learning models. The insights drawn from temporal and spatial patterns will inform the architecture of our LSTM-based predictor. The preprocessing pipeline ensures model readiness and improved generalizability.

6. References

- OpenAQ. (n.d.). *OpenAQ platform*. https://openaq.org/

- NASA. (n.d.). *MODIS (Moderate Resolution Imaging Spectroradiometer) data*. https://modis.gsfc.nasa.gov/

- NOAA National Centers for Environmental Information. (n.d.). *Climate data online*. https://www.ncei.noaa.gov/

Technology Review

1. Introduction

The successful implementation of air quality prediction models relies on the appropriate use of data science technologies. This section explores the tools and frameworks used in our project and evaluates their applicability, advantages, and limitations. A sound technological foundation ensures our model is scalable, interpretable, and easily deployable for public use.

2. Technology Overview

Our project incorporates the following technologies:

- LSTM (Long Short-Term Memory): A type of recurrent neural network designed to handle sequential and time-series data.

- Random Forest: An ensemble machine learning algorithm used for classification, regression, and feature importance analysis.

- TensorFlow/Keras: A deep learning framework for building and training neural networks.

- Pandas and NumPy: Core Python libraries for data manipulation and numerical computations.

- Streamlit: A Python-based open-source app framework for interactive data visualization dashboards.

Each technology is selected based on its relevance to data processing, predictive modeling, or user interaction.

3. Relevance to Our Project

LSTM is central to our prediction engine as it captures temporal patterns in pollutant data. Random Forest is used to identify key features and as a baseline model to compare with deep learning approaches. TensorFlow/Keras provides the infrastructure for training deep models efficiently. Pandas and NumPy streamline preprocessing and transformations, while Streamlit converts model outputs into real-time visual insights for urban policymakers.

4. Comparison and Evaluation

Technology	Strengths	Limitations	Suitability
LSTM	Excellent for time-series, memory retention	Requires large datasets, risk of overfitting	High
Random Forest	Interpretable, handles missing data	Not ideal for temporal sequences	Medium
TensorFlow	Flexible, supports complex architectures	Steep learning curve	High
Streamlit	Easy UI creation, rapid deployment	Lacks multi-user capabilities	High
Pandas/NumPy	Efficient, industry-standard for data manipulation	Memory-intensive for large datasets	High

5. Use Cases and Examples

- LSTM has been used by environmental agencies in China and India for AQI forecasting.

- The U.S. Environmental Protection Agency (EPA) employs sensor-based dashboards to alert citizens in real time.

- Streamlit has been widely used by academic and municipal projects to visualize pollution and climate indicators.

6. Identify Gaps and Research Opportunities

While LSTM provides strong performance, its black-box nature poses challenges to interpretability. Tools like SHAP can enhance transparency. Streamlit is effective for MVPs but may need integration with more robust platforms like Dash or Flask for scalability. There's also an opportunity to explore real-time API data feeds and edge-device deployment in future iterations.

7. Conclusion

The technological stack used in this project ensures a strong balance between accuracy, usability, and scalability. LSTM models, when supported by tools like TensorFlow and Streamlit, provide an end-to-end framework for predicting and visualizing urban air quality in a practical, actionable manner.

8. References

- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
- Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
- TensorFlow. (n.d.). *TensorFlow documentation*. https://www.tensorflow.org/
- Streamlit. (n.d.). *Streamlit documentation*. https://docs.streamlit.io/
- McKinney, W. (2010). Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56). https://doi.org/10.25080/Majora-92bf1922-00a