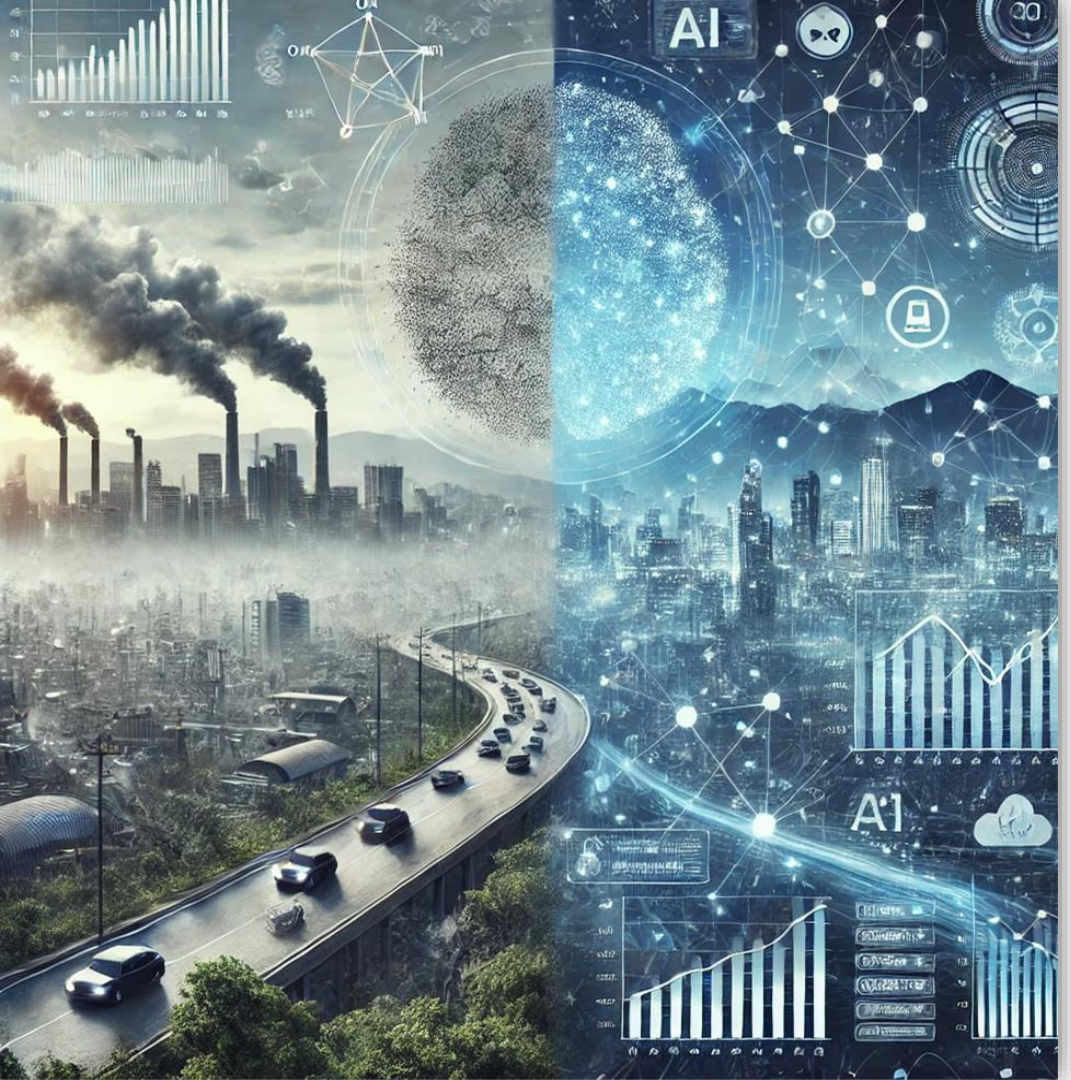


Predicting Air Quality:
An Applied AI Approach to Beijing's Pollution Data

Applied ML Final Project

Initial Feasibility Report
ENG M 680

Group members

Shuvo Biswas (1877743)
Nasim Zaman Piyas (1888376)
Maryum Ali (1888378)
Izuchukwu Ogbuigbo (1872279)
Manjot Singh (1885545)

Stakeholder Analysis

External

Internal

Problem Statement

What is Air Pollution?

•Air pollution involves harmful substances in the air, including particulate matter (PM2.5, PM10), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and ground-level ozone (O_3).

•Major sources of air pollutants include vehicle emissions, industrial activities, and natural events like wildfires.

•These pollutants pose serious health risks, contributing to respiratory and cardiovascular diseases in exposed populations. It also harms the environment by lowering air quality and negatively affecting ecosystems.

•Urban and industrial regions are particularly vulnerable, with significant public health risks and economic costs associated with poor air quality.

World Health Organization. Ambient (outdoor) air pollution. 2018, www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health.

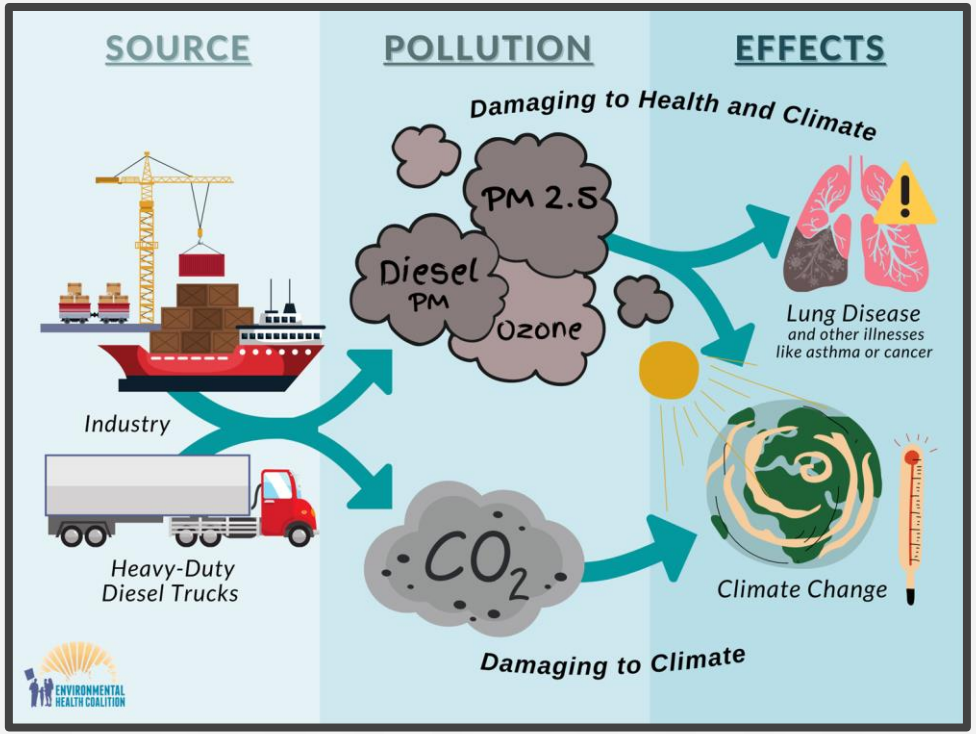


Figure 1. Various ways of Source, Pollution of Effects of Air Pollution

Problem Statement

Air pollution in Beijing

Main Sources:

- ❖ Industrial activities
- ❖ Vehicle emissions
- ❖ Coal burning

Primary Pollutants:

- ➢ PM2.5 (fine particulate matter)
- ➢ SO_2 (sulfur dioxide)
- ➢ NO_2 (nitrogen dioxide)

- **Seasonal Impact**: Winter weather patterns trap pollutants, worsening air quality
- **Health Impacts**: Increased respiratory and cardiovascular risks, especially for children and the elderly
- **Environmental & Economic Consequences**: Reduced visibility and acid rain formation. Moreover, Higher healthcare costs and decreased productivity due to health effects

Source: (Zhang et al. 706).

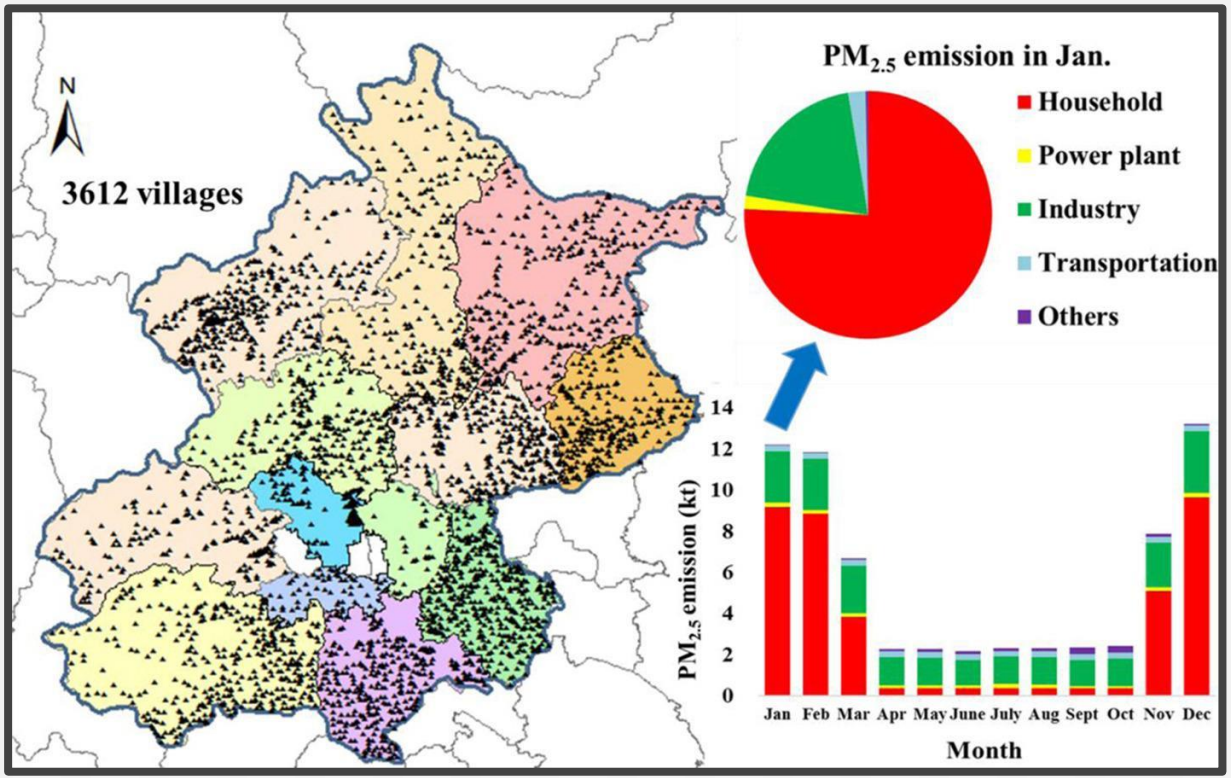


Figure 2. Emissions from household heating contributed to ~70% of $PM_{2.5}$ and ~60% of SO_2 emissions in winter

Problem Statement

Air pollution in Beijing

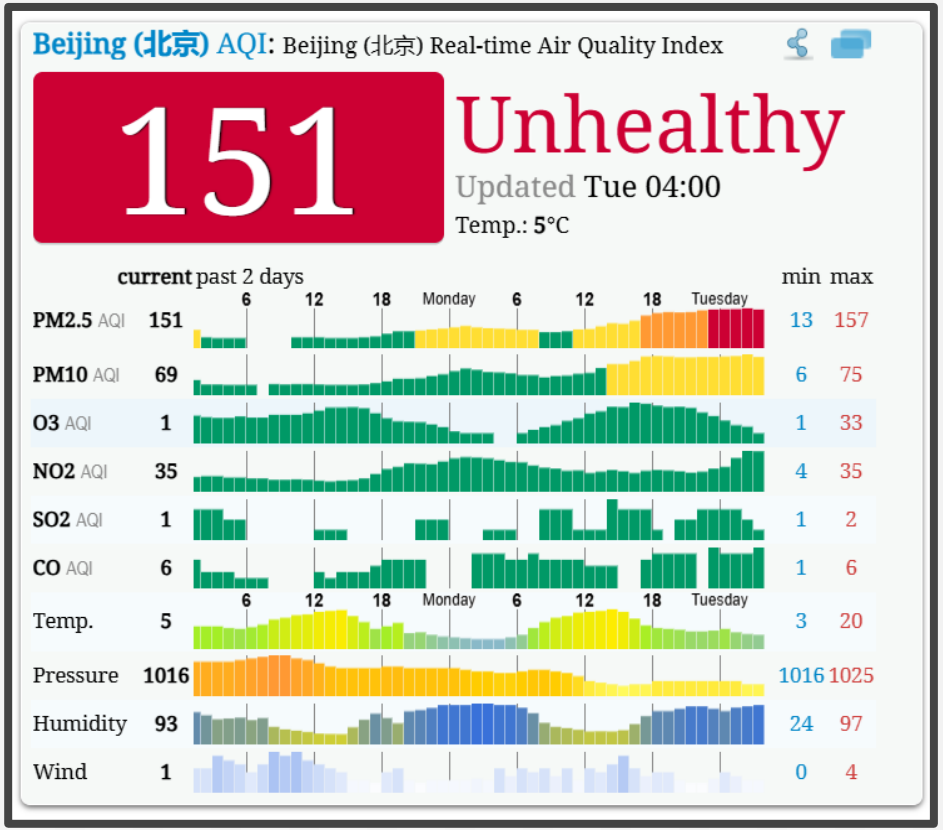


Figure 3. Real –Time Air Quality Index of Beijing

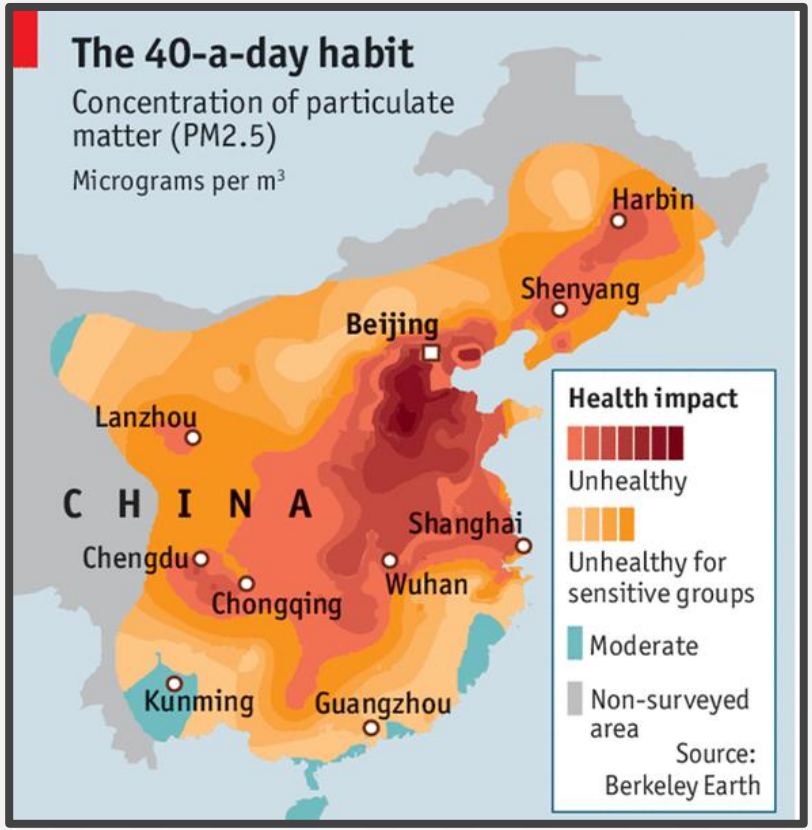


Figure 4. Concentration of Partculate Matter (PM2.5)

Problem Statement

Objectives and potential ROI

This project addresses the negative effects of air pollution on public health and quality of life in Beijing. By using AI to predict pollution levels, it offers valuable benefits for public health and economic efficiency. Government agencies, healthcare providers, and residents can use these forecasts to anticipate and reduce health risks.

Figure 5. Key stakeholders working to improve air quality and health outcomes

- **Public Health and Economic Impact**: Predictive air quality models enable timely health advisories, potentially reducing respiratory hospital visits by 10-15% and lowering pollution-related healthcare costs by 5%, offering substantial savings for Beijing's healthcare sector (Guan et al. 1392; Zhang et al. 707).

- **Environmental and Social Significance**: Improved air quality enhances life quality, supports higher productivity, and aligns with Beijing's urban sustainability goals, reducing pollution exposure risks and creating healthier public spaces (Wang et al. 854).

- **Business Relevance**: This project aids urban planning and public health management, attracts clean technology investments, and boosts Beijing's appeal as a livable city, which may support economic development and tourism (Maji et al. 225).

Responsible AI

Ethical Concerns and Biases

In our project, when exploring correlations between weather patterns and pollution levels in Beijing, several ethical considerations need to be addressed to align with Responsible AI principles. Here's a breakdown of potential ethical concerns:

Geographical Bias	Uneven data collection across monitoring sites may under represent some areas. This could lead to biased findings if pollution levels, and weather patterns are not consistently covered across all regions.
Temporal Bias	Data only spans March 2013 to Feb 2017, which may not capture recent trends in air quality or variation post 2017.
Data Collection Bias	If some monitoring stations have inconsistent data quality due to technical issues (e.g., equipment malfunctions or missing values), this could skew results.
Privacy	Although the data itself is non-personal, transparency is needed if findings influence public perception or policy.
Data Security	Proper storage and handling are essential, especially with any sensitive supplementary data.

Responsible AI

Mitigation Strategies

Bias Mitigation:

- **Normalize and Test**: Ensure consistent data coverage across locations and time; use statistical tests to confirm adequate representation.

Privacy and Security:

- **Transparent Data Use**: Document data sources and explain non-personal data usage.
- **Secure Data Storage**: Store data securely with restricted access.

Responsible Analysis:

- **Fairness Adjustments**: Apply resampling and normalization for balanced results.
- **Highlight Limitations**: Communicate any data limitations clearly in findings.

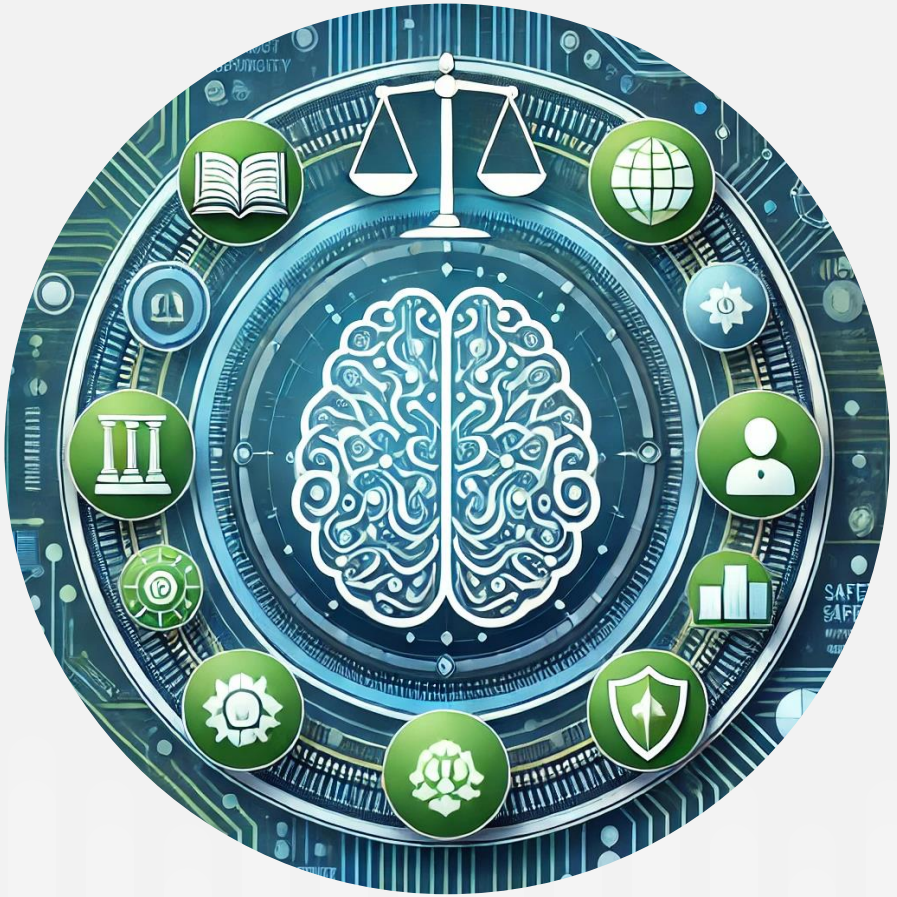
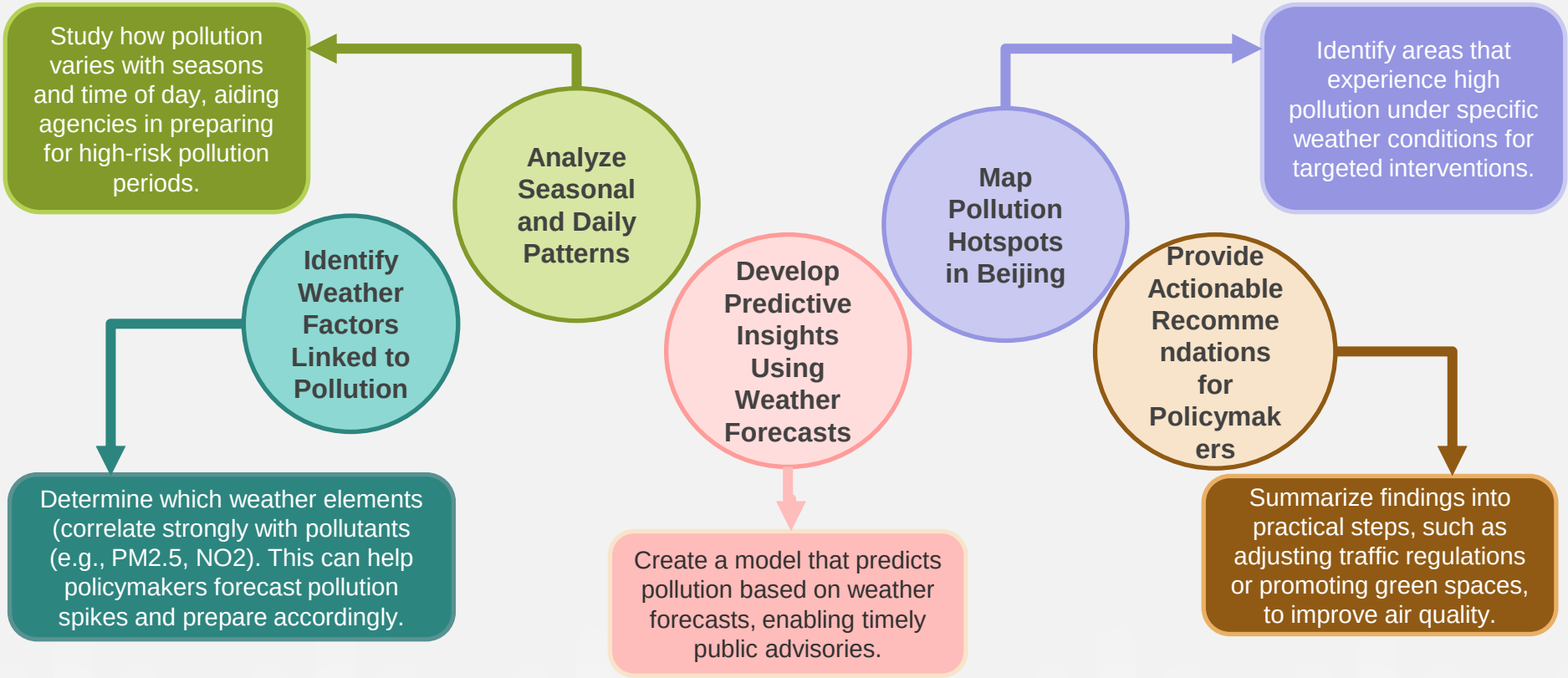


Figure 6. AI generated picture to visualize Mitigation Strategies

Project Goals



Alignment with Business Problem: These objectives offer insights for predicting pollution, guiding public health actions, and informing policies for better air quality management and urban health outcomes.

Data Context

Overview of the Dataset

- The air-quality data is sourced from the Beijing Municipal Environmental Monitoring Center, with weather information paired from the nearest meteorological station by the China Meteorological Administration.
- Data is collected from 12 different monitoring It contains includes pollutant concentrations (PM2.5, PM10, SO2, NO2, CO, O3) and meteorological factors (temperature, pressure, wind direction, humidity, etc.).
- Covers hourly data from March 2013 to February 2017
- Each station has the same number of samples and features which are 35064 and 18, respectively.

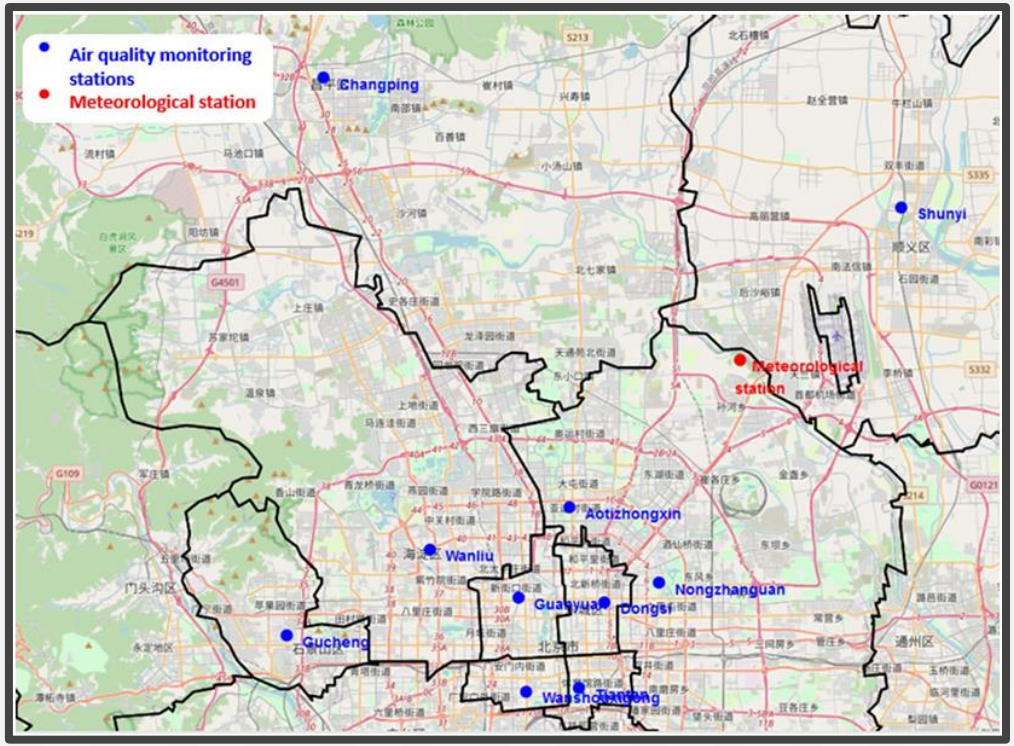


Figure 7. The distribution of monitoring stations in Beijing (Bekkar, A.,et al, 2021)

Data Context

Dataset Features

Sl No	2nd -5th	6th -11th	12th -18th
Category	Temporal Features	Air Pollutants	Weather Variables
Description	Aid in identifying seasonal patterns	Essential indicators of air quality and health risk	Influence on pollutant dispersion and concentration

Additional Datasets Integration

Purpose:

Helps create a more accurate and holistic air quality prediction model

Water-Soluble particles

Airflow Dynamics

Rain patterns

- Particles contribute to pollution levels that impact health in humid conditions.
- Wind speed, direction, and pressure changes affect pollutant dispersion.
- Rain reduces pollutant levels by washing particles out of the air

Feature	Description
No	Record index for each data entry
year	Year of the measurement
month	Month of the measurement
day	Day of the measurement
hour	Hour of the measurement
PM2.5	Concentration of PM2.5 particles (μg/m^3)
PM10	Concentration of PM10 particles (μg/m^3)
SO2	Concentration of sulfur dioxide (SO_2, μg/m^3)
NO2	Concentration of nitrogen dioxide (NO_2, μg/m^3)
CO	Concentration of carbon monoxide (CO, mg/m^3)
O3	Concentration of ozone (O_3, μg/m^3)
TEMP	Temperature in degrees Celsius (°C)
PRES	Atmospheric pressure in hectopascals (hPa)
DEWP	Dew point temperature in degrees Celsius (°C)
RAIN	Precipitation in millimeters (mm)
wd	Wind direction (categorical)
WSPM	Wind speed in meters per second (m/s)
station	Name of the monitoring station

*Table 1. D*ataset features and their descriptions

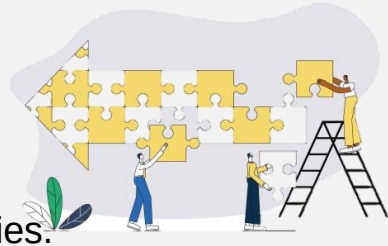
Data Context

➢ Completed Preprocessing

Missing Values:

- Replaced using monthly averages specific to each station.
- Used other stations' monthly averages when data was missing for an entire month.
- Applied overall feature averages if no station data was available.

➢ Future Preprocessing Steps



- **Outlier Handling**: Remove or adjust data inconsistencies to improve model reliability.
- **Feature Integration**: Add supplementary features for better model accuracy.
- **Time-Series Components**: Integrate time-series analysis to enhance predictive capabilities.

➢ Dataset Challenges

- **Temporal Variation**: Seasonal patterns add complexity to modeling.
- **Data Quality**: High volume of missing values requires careful handling.

Initial EDA

Handling the missing values

Handled missing values by replacing them with monthly averages specific to each station to minimize bias and ensure data consistency. (Example: Tiantan)

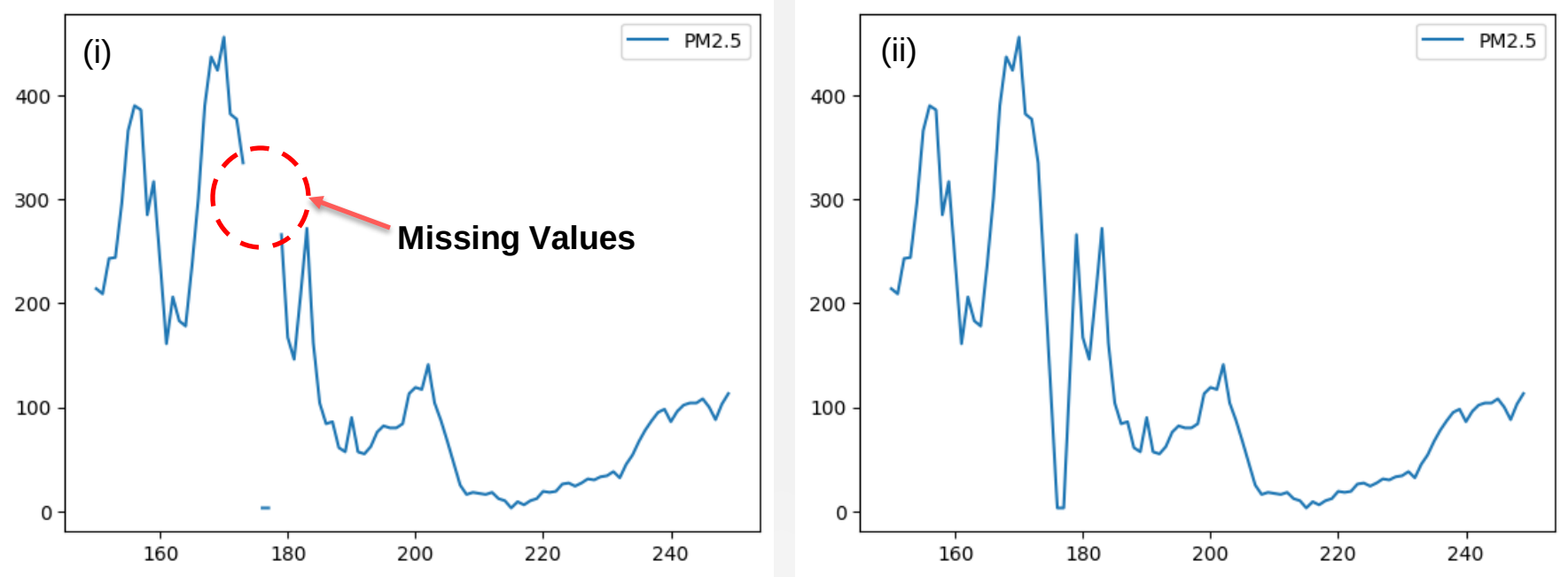


Figure 9: Values for the PM2.5 column plotted (i) before and (ii) after the replacement with monthly averages

Initial EDA

Encoding Categorical Variables: Wind Factor

Importance of Wind Factor in Analysis

- **Atmospheric Influence:**

Wind plays a key role in pollutant distribution and concentration.

- **Pollutant Dispersion**:

Wind speed affects how pollutants spread, while helps determine areas with higher PM2.5 levels.

0°	– north wind	(N)
22.5°	– north-northeast wind	(NNE)
45°	– northeast wind	(NE)
67.5°	– east-northeast wind	(ENE)
90°	– east wind	(E)
112.5°	– east-southeast wind	(ESE)
135°	– southeast wind	(SE)
157.5°	– south-southeast wind	(SSE)
180°	– south wind	(S)
202.5°	– south-southwest wind	(SSW)
225°	– southwest wind	(SW)
247.5°	– west-southwest wind	(WSW)
270°	– west wind	(W)
292.5°	– west-northwest wind	(WNW)
315°	– northwest wind	(NW)
337.5°	– north-northwest wind	(NNW)
360°	– north wind	(N)

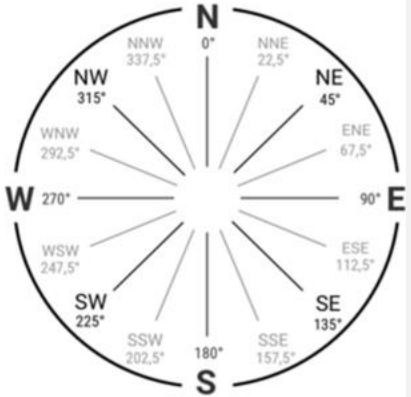


Figure 10: Wind Direction and Degree Values

Encoding Approach

- **Wind Direction as a Categorical Variable**:

Encode wind direction as a categorical feature to capture its influence on pollutant distribution.

- **Impact on PM2.5 Concentration**:

Proper encoding helps model the relationship between wind direction and pollution levels.

Initial EDA

Correlation Matrix: Key Insights

<u>High Correlation</u>:

PM10 and PM2.5: Strong correlation of 0.89, indicating these pollutants often increase together.

<u>Weather Factors</u>:

Temperature, Pressure, Dew Point: Show significant correlations with pollutant levels, suggesting weather conditions influence air quality.

<u>PM2.5 and CO</u>:

Correlation of 0.80: Indicates that CO levels significantly impact PM2.5 concentration, emphasizing CO's role in air pollution.

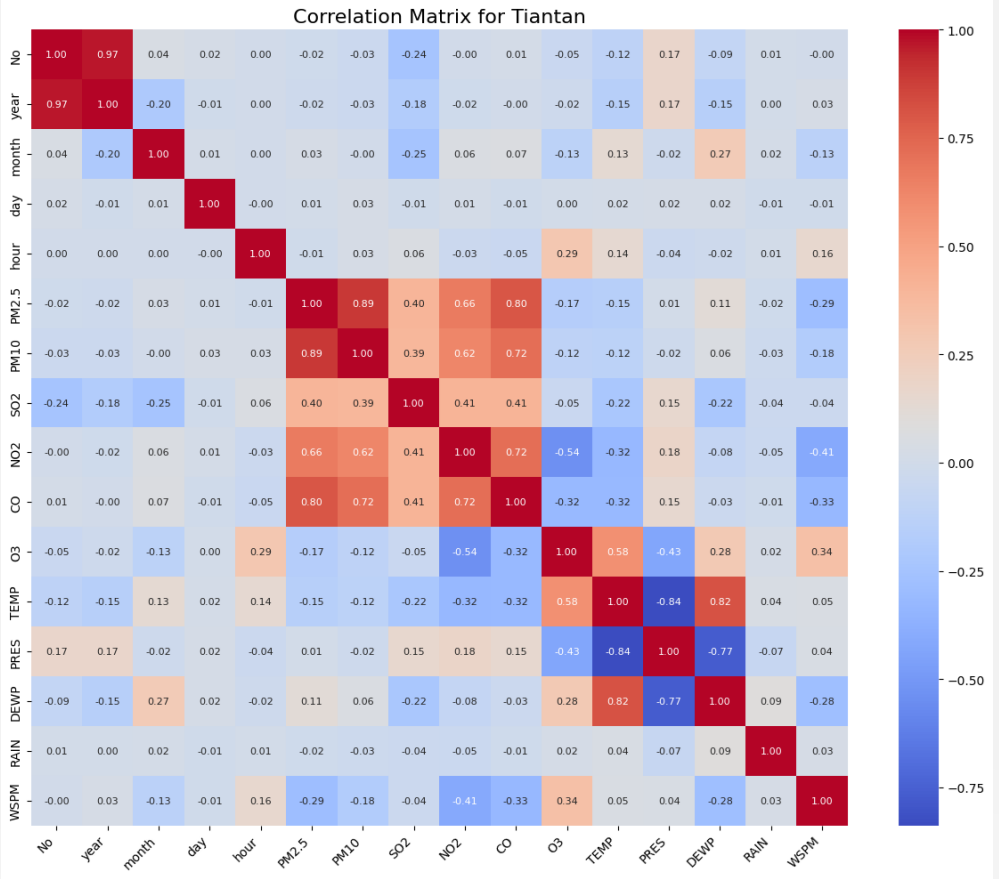


Figure 11: Correlation Matrix for Tiantan

Technical Description of Proposed Approach

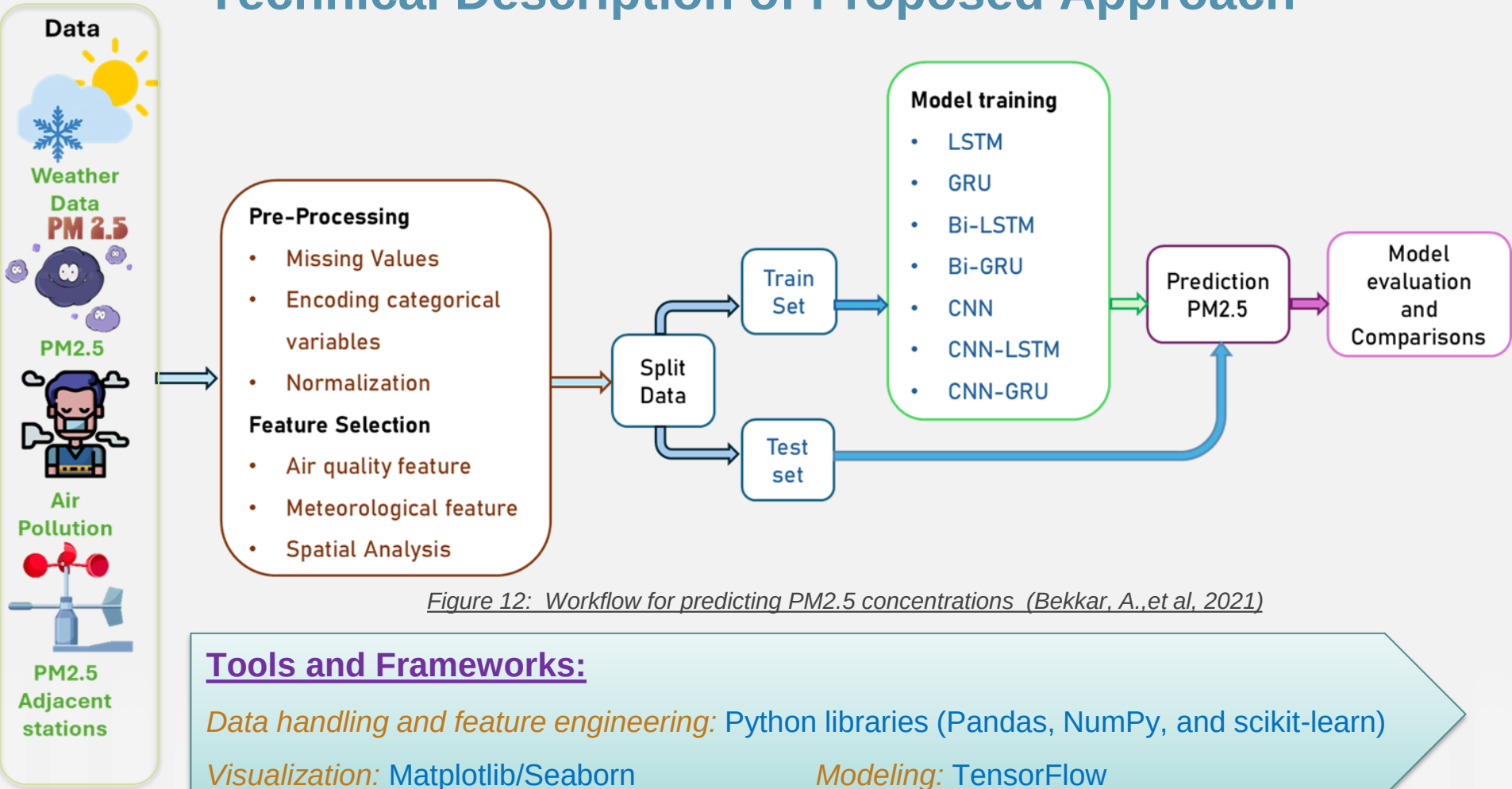


Figure 12: Workflow for predicting PM2.5 concentrations (Bekkar, A.,et al, 2021)

Tools and Frameworks:

Data handling and feature engineering: Python libraries (Pandas, NumPy, and scikit-learn)

Visualization: Matplotlib/Seaborn *Modeling:* TensorFlow


Thank you!