Proposal: A multi-modal approach for prediction of Air Quality Index using Machine Learning

Sachin Malego
st125171@ait.asia
Asian Institute of Technology
Thailand

Bidhan Bajracharya
st125287@ait.asia
Asian Institute of Technology
Thailand

Ishita Pradhan
st125024@ait.asia
Asian Institute of Technology
Thailand

Aman Oberoi
st125490@ait.asia
Asian Institute of Technology
Thailan

1. Problem Statement

Air pollution's impact on human health has become an increasingly urgent issue in recent years. Despite the availability of the Air Quality Index (AQI) to monitor pollution, there is limited understanding of the individual impacts of specific AQI components on overall air quality, as well as the causes of abrupt increases in these pollutants. A more granular understanding of these factors is essential for developing targeted interventions to mitigate pollution.

Key determinants of air quality in any given area include particulate matter, gaseous pollutants, and meteorological variables. Along with these various other factors, including automotive emissions, industrial activities, and meteorological changes, influence air pollution in distinct ways [1]. While automotive emissions, industrial activities, and weather patterns contribute to air pollution dynamics, the complex interactions among these sources and pollutants remain challenging to analyze. The primary contributors to the AQI—Particulate Matter 2.5 (PM2.5), PM10, Carbon Dioxide (CO2), Carbon Monoxide (CO), Sulfur Oxides (SOx), Nitrogen Oxides (NOx), Ozone (O3), and Ammonia (NH3)—demand a more granular understanding to develop effective solutions.

Figure 1: Primary contributors to the AQI

Therefore, the objective of this project is to gain a comprehensive understanding of how individual AQI components influence overall air quality and to uncover the underlying factors behind significant increases in specific pollutants. By providing insights into the causes of AQI spikes at particular times, this project aims to support proactive and informed approaches to air quality management. Specifically, the project aims to answer questions such as: How do meteorological conditions correlate with variations in PM2.5 levels? What sources contribute most significantly to spikes in individual AQI concentrations? By providing insights into the causes of AQI spikes at

particular times, this project aims to support proactive and informed approaches to air quality management, ultimately contributing to improved public health and environmental policies.

2. Market Analysis

The Air Quality Index (AQI) is a crucial measure that communicates air quality levels to the public. It aggregates data on key pollutants. As air pollution becomes a significant global concern, the AQI serves as an essential tool for informing public health policies, environmental regulations, and community awareness initiatives. Regulatory bodies, such as the U.S. Environmental Protection Agency (EPA) and the World Health Organization (WHO), set standards that define safe levels for these pollutants, influencing the demand for monitoring solutions.

The global air quality monitoring market was valued at approximately $4.9 billion in 2023 and is projected to grow to around $6.9 billion by 2028, with a compound annual growth rate (CAGR) of about 7% during this period. This growth is attributed to increasing government regulations, rising awareness of the health impacts of air pollution, and public-private funding initiatives aimed at enhancing air quality monitoring systems. In another estimate, the air quality monitoring market was valued at around $4 billion in 2023, with projections to reach approximately $7 billion by 2030, showing a CAGR of 8%. This upward trend is primarily driven by factors such as urbanization, stringent regulatory frameworks, and an increasing focus on public health concerning air pollution [2, 3, 4].

3. Business Model

Rising interest in environmental issues can create demand for content and platforms dedicated to educating people about air quality and its health effects. The proposed business model for the Air Quality Index (AQI) can be of providing monitoring services, data and analytics to address the growing concerns over air pollution, and provision of support to government regulations bodies in support to policy formulation. This model can include a subscription-based platform offering tiered plans for consumers, businesses, and government agencies, providing access to detailed AQI data, historical trends, and predictive analytics.

Additionally, revenue can be generated through data analytics services, which sell aggregated insights to organizations needing environmental assessments and compliance support. Partnerships with IoT device manufacturers can enhance monitoring capabilities, while educational programs can raise community awareness about air quality issues. Target markets include health-conscious individuals, corporations in regulated industries, and governmental bodies focused on public health initiatives.

The value proposition centers on providing actionable insights that empower users to make informed decisions to protect their health and enhance environmental sustainability. By combining technology, data analytics, and community engagement, this business model can foster a proactive approach to air quality management, ultimately improving public health outcomes and environmental conditions.

4. Financial Projections

For the proposed AQI monitoring service, this projection outlines anticipated financial performance over the first three quarters, emphasizing partnerships with Industries, UN (United Nations), I/NGOs, and Government agencies. By engaging these key stakeholders, the service aims to secure a diverse revenue stream through industry partnerships, UN and I/NGO collaborations, and government

contracts. These projections are based on market trends and demand for air quality solutions, as growing public concern over environmental issues drives investments in monitoring technologies. This approach not only ensures financial sustainability but also aligns with broader public health and environmental objectives. Below is the detailed financial projection reflecting these strategies.

- Industry Partnerships (CSR – Community Social Responsibility): For businesses, being able to monitor and report AQI data around industrial sites areas is essential to meeting regulatory standards, and it supports transparency initiatives that inform the public. Collaboration with 5 companies in Q1, generating $10,000 each for data services, increasing by 5% each quarter.
- UN and I/NGO Engagement: Partner with 10 NGOs in Q1, contributing $15,000 per organization for workshops and awareness campaigns, with a 10% increase in participation and funding each quarter.
- Government Contracts: Initial contracts worth $30,000 in Q1 for monitoring and reporting services, increasing by 20% each quarter.

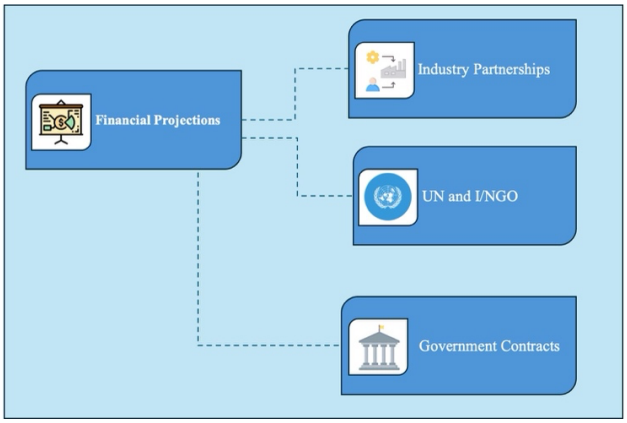

Figure 2: Stakeholders mapping for partnerships

Quarterly Breakdown:

Stakeholders	Q1	Q2	Q3
Industry Partnerships (CSR)	$50,000	$52,500	$55,125
UN and I/NGO Engagement	$150,000	$165,000	$181,500
Government Contracts	$30,000	$36,000	$43,200
Total Revenue	**$230,000**	**$253,500**	**$279,825**
Operating Costs	$90,000	$108,000	$130,000
Net Profit	**$140,000**	**$145,500**	**$149,825**

By the end of Q3, the total projected revenue reaches approximately $279,825, with a net profit of around $149,825. These financial projections show a steady growth trajectory, driven by diversified partnerships across Industry, UN and I/NGO, and Government sectors, positioning the AQI monitoring service for sustained success and community impact.

5. Risk Assessment and Mitigation

Predicting AQI is vital for safeguarding public health, enabling timely interventions, and helping policymakers make informed decisions about environmental regulations. With increasing urbanization and industrial activities, pollution levels have risen globally, necessitating robust, predictive models that can accurately forecast air quality levels. A multi-modal approach to AQI prediction leverages diverse data sources offering a comprehensive perspective on factors influencing air quality. ML (Machine Learning) models applied to such multi-modal data can detect patterns and forecast AQI with precision and provide early warning and help mitigate air pollution's adverse effects.

Despite the promise of such systems, the inherent complexity of using multiple data sources and ML techniques introduces several risks. These include data quality issues, model interpretability challenges, and the potential

for overfitting due to noise within certain data types. Furthermore, discrepancies between data sources can amplify prediction errors, and the reliance on sensitive input data, like traffic and industrial emission data, may raise concerns about data availability and privacy. An additional concern in this can be the mitigation for huge load of missing values for certain key factors contributing to AQI. Moreover, one stop machine learning model solution may not be efficient for such kind of dynamic factors as it varies depending on several important features as per the geographical, historical backgrounds and external sources. Another risk can be of class imbalance in certain features leading to biased prediction.

In order to mitigate this a through data collection, cleaning and pre-processing needs to be performed. If lots of missing values are prevalent in the chosen dataset then consider alternative datasets from reliable, geographically consistent sources. Techniques like data imputation, interpolation, or even real-time data sources may also help manage gaps effectively. Model must be trained according to the geographic area and system must be able to identify the area before performing prediction calculation. Implementing regularization techniques and cross-validation to control overfitting. Applying methods such as SMOTE (Synthetic Minority Over-sampling Technique) to balance classes and ensure a fair representation across diverse AQI factors can help mitigate biasness.

6. Solution

6.1 Overview

In recent years, there has been a growing focus on developing robust models for analyzing and predicting air pollutant concentrations. A multi-modal approach leveraging machine learning can enhance Air Quality Index (AQI) forecasting capabilities, providing invaluable tools for environmental management and public health initiatives. Prolonged exposure to elevated levels of air pollution can lead to severe health issues. Moreover, various pollutants contribute to significant environmental challenges such as the greenhouse effect, ozone depletion, haze, and acid rain. Thus, accurate forecasting of air quality is essential for effective pollution management and timely early warning systems.

Implementing a multi-modal machine learning framework for AQI prediction can integrate diverse data sources, such as meteorological data, pollutant concentration data, traffic, and industry emission data and others. By combining these diverse modalities, machine learning algorithms can be employed to develop a comprehensive AQI prediction model. This multi-faceted approach aims to produce accurate forecasts of air quality, enabling proactive management strategies and public health advisories.

A multi-modal machine learning approach for predicting the Air Quality Index represents a promising solution for addressing the critical challenges posed by air pollution. By harnessing the power of diverse data sources and advanced machine learning analytics, we can foster a healthier environment and safeguard public health.

6.2 Value Proposition

The multi-modal machine learning model for predicting the Air Quality Index (AQI) significantly enhances predictive accuracy by integrating diverse data sources, including meteorological conditions, pollutant concentrations, and

other factors. This precision allows for timely public health alerts, enabling authorities to implement proactive measures during high pollution events.

Our solution empowers industries, health authorities and policymakers to make informed, data-driven decisions regarding air quality management. By accurately forecasting pollution events, they can prioritize regulatory actions and allocate resources efficiently, promoting sustainable environmental practices that contribute to long-term ecological health.

In addition to supporting public health management, this model raises community awareness by providing AQI predictions. When individuals are informed about air quality, they can make better decisions regarding outdoor activities, leading to greater engagement in pollution reduction initiatives.

Scalability and adaptability are key features of our approach. The model can be customized to fit various geographic contexts, ensuring its relevance across urban, suburban, and rural settings. Investing in this multi-modal AQI prediction model yields significant economic benefits. By reducing healthcare costs associated with pollution-related illnesses and improving workforce productivity, communities can achieve long-term economic sustainability. Furthermore, the model supports ongoing research into air quality issues, enabling academic institutions to study the effects of pollution and inform policy formulation.

Thus, this multi-modal approach to AQI prediction represents a transformative solution for combating air pollution. By leveraging machine learning analytics and comprehensive data sources, the project aims to empower stakeholders to take informed, proactive actions to improve air quality and safeguard public health.

6.3 Technical Details

Most of the research experiments conducted in field of AQI only catered the key contributors however there was not much details on the cause of high or low AQI. Almost all previous experiments conducted referred to Random Forest, SVM or other Linear Regression models as best to predict AQI.

The technical framework below outlines the comprehensive approach of the multi-modal AQI prediction model, integrating advanced machine learning technologies with a robust architecture to enhance air quality forecasting. By leveraging diverse data sources and sophisticated modeling techniques, this solution aims to provide accurate and actionable insights for effective air quality management and public health protection.

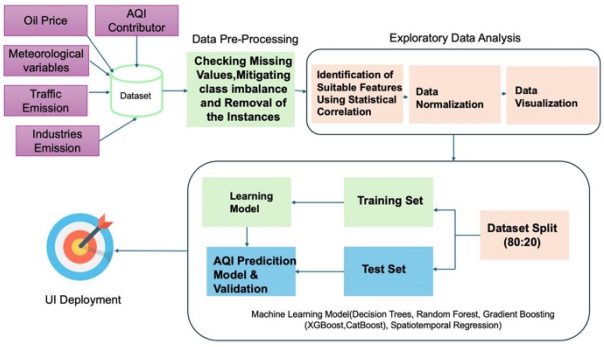

Figure 3: High-level architecture of the solution

A. Data Collection

The data collection process involves gathering information from multiple sources through web crawling. This diverse range of datasets presents challenges when it comes to merging

them due to their varying formats and structures. However, a common feature across these datasets is the date attribute. By leveraging this date feature, we can effectively establish correlations and synchronize samples, facilitating a more streamlined integration of the data for analysis.

B. Data Preprocessing

In this process, we aim to address two critical challenges: handling missing values and mitigating class imbalance. To ensure the integrity of our dataset, we will implement strategies to fill in any missing values appropriately. Additionally, since the dataset only provides values for the key contributors to the Air Quality Index (AQI), we will calculate the AQI levels for all key contributors using the established AQI formula. This calculation will allow us to create a more complete and accurate representation of air quality, which is essential for effective analysis and modeling.

C. Exploratory Data Analysis

In this we perform identification of suitable features using in the statistical correlation, regularize and normalize the data wherever necessary, and visualize the data for finding relations with various factors and features.

D. Data Splitting

In our approach, we will partition the dataset into three distinct subsets: training, validation, and test sets. The training set will be used to train our machine learning models, allowing them to learn patterns and relationships within the data. The validation set will help us fine-tune model hyperparameters and assess performance during training, ensuring that the models do not overfit the training data. Finally, the test set will serve as an independent evaluation of model performance, providing a reliable measure of how well the models can generalize to unseen data. This structured data splitting strategy is crucial for developing robust and reliable predictive models for AQI.

E. Machine Learning Models

a. Decision Tree

Our base-line solution sticks with decision tree, the primary reason for that is the missing pattern, most of the information are missing at not at random, and it is critical to understand the data generative process. Yet with the provided information we found that boosting has innate characteristics to handle such missing values.

b. MICE

Multivariate Imputation by Chained-Equation. Imputation at various level is challenging task with the shallow dataset we encounter. It is due to the fact of accounting features likes industrial production and traffic; can be proximal features (not at exact location) such features can be used in imputation across various levels but we are not sure how our model will behave. In order add more dynamism and less skewness to missing patterns we will be using Multivariate iterative imputation procedure.

c. Model-Ensemble
The short-term dynamics of the changing weather pattern and Air-quality level is really a challenge. For each intra-day changes we see short term spiks. In such cases our modeling process will show effect at various levels that is weekly model, monthly model and daily model. Having said that each model will be stacked and blended together.

d. Cat-Boost
Our problem encounters various features like energy production per capita, population density which are projecting across years and are binned at various locations, Cat-boot will and sublevel of categories capture those features. Moreover, the missing values with test and train split can generalize well across such categories

e. Spatiotemporal Regression
The cross-section of spatial, temporal dynamics of PM2.5 particles, this model accounts for the variations in changing weather and traffic morphology across various destinations (i.e. sensory locations) at nearby location. Furthermore, industrial production also takes for account.

F. UI Deployment
In this project, we plan to deploy the final model through a user-friendly interface (UI) that will facilitate access to our Air Quality Index (AQI) predictions and analyses. The UI will be designed to cater to various stakeholders, including public health officials, environmental agencies, researchers, and the general public.

The deployment process will involve creating a web-based application that provides AQI forecasts, historical data visualizations, and actionable insights based on the predictions. Users will be able to input specific parameters, such as location and date, to receive customized air quality prediction.

To ensure accessibility and ease of use, the UI will feature intuitive navigation, clear data visualizations (e.g., graphs and heatmaps), and informative dashboards that present key metrics and trends related to air quality.

Overall, the deployment of the final model through an interactive UI aims to empower users with timely and accurate air quality information, fostering greater awareness and engagement in air pollution management efforts.

6.4 Mockups
The mockup design for the AQI project comprises two primary sections: Dashboard and Prediction page. Dashboard, serves as the homepage, presenting detailed information about AQI through multiple, visually distinct sections, each designed for intuitive navigation and clarity in data representation. These sections utilize various visualization techniques to make AQI data easy to interpret and actionable. Meanwhile, the Prediction page features a straightforward form where users input specific parameters to generate a predicted AQI score, enhancing user interaction and data personalization.

6.4.1 Dashboard View

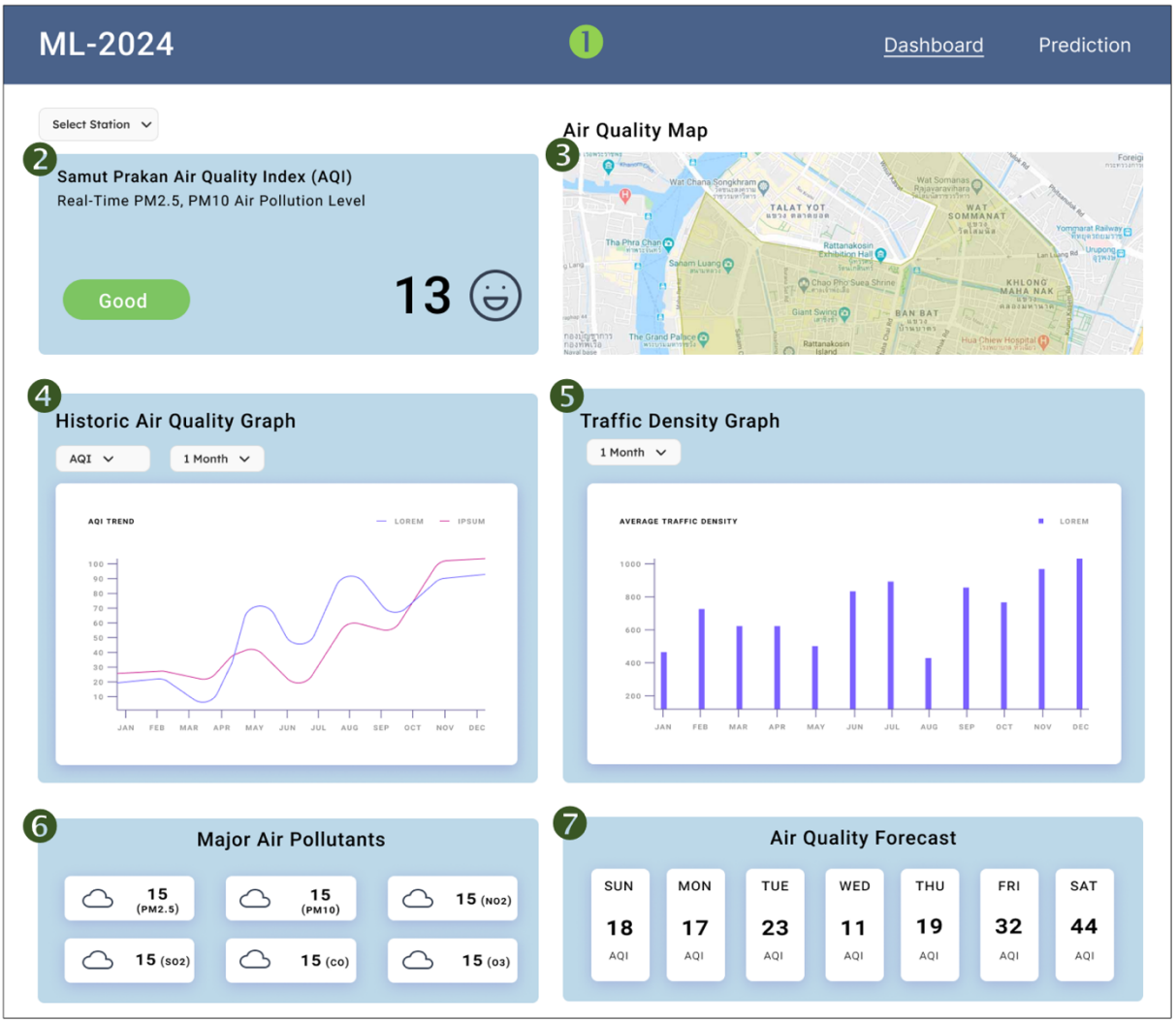

Figure 4: Mockup design – Dashboard

[1] Navbar: The navbar contains two tabs 'Dashboard' and 'Prediction'. It is used for navigating between the two pages.

[2] AQI Indicator: This section contains details about the location and the AQI. It also contains a filter which users can use to check out different stations.

[3] Air Quality Map: The map is used to display the exact location of the substation.

[4] Historic Graph: This graph shows the historic trend of specific elements i.e., NO2, O3, SO, etc. Users can toggle between these elements through the filter and the range of data that they want to view as well.

[5] Traffic Density Graph: This graph shows the density of traffic near the particular station. Users can also filter the time range to view historical data.

[6] Air Pollutants: This section displays the major air pollutants of that area with their scores. This way users can understand which air pollutants is affecting the AQI most in that region.

[7] Forecast: This section forecasts AQI levels for the current week so that users can use this information and plan their activities ahead.

6.4.2 Prediction Page

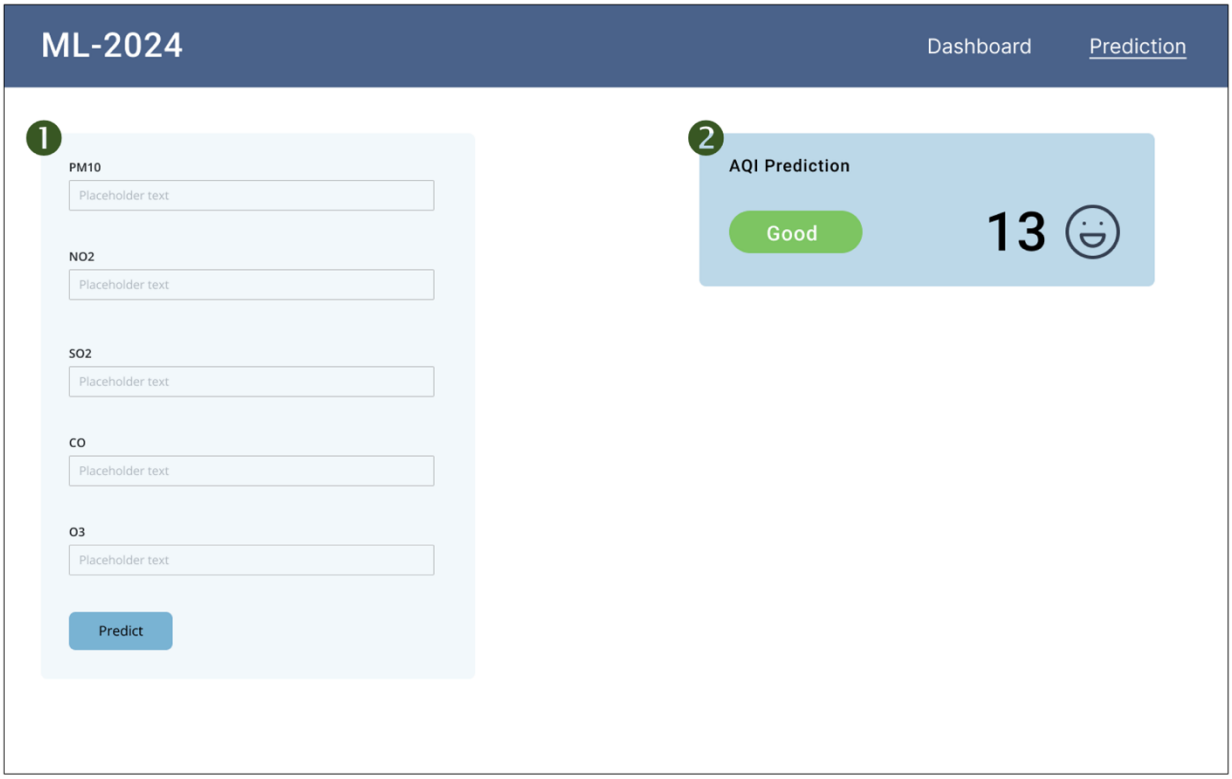

Figure 5: Mockup of Prediction page

[1] **Input Form**: The form contains input fields for feature variables i.e., PM10, NO2, SO2, CO and O3. Validations will also be present to make sure the data is passed in the correct format. After providing the details and clicking the 'Predict' button, the result will be displayed in section 2.

[2] **Prediction Display**: This section contains the output of the prediction based on the input parameters provided

before. It displays the AQI score, label and an avatar to convey the meaning of the score in layman term.

6.5 Use Cases

The identified stakeholders can use the multi-modal AQI prediction model across various use cases. Primarily by the use of this the stakeholders can provide accurate and actionable air quality forecasts, companies can enhance their operations, contribute to public health initiatives, and promote environmental sustainability.

The government aiming to improve air quality and enhance public health can deploy the AQI prediction model to monitor and forecast air quality. By integrating data from various sensors distributed across the city, the model can provide timely alerts to residents regarding high pollution days. Furthermore, city officials can implement traffic management strategies, such as rerouting heavy traffic or temporarily restricting vehicle access in high-pollution areas based on the predictions. This proactive approach not only improves air quality during critical periods but also promotes a healthier living environment for residents.

Similarly, public health agency focused on mitigating health risks associated with air pollution, particularly for vulnerable populations, can leverage the AQI prediction model to issue health advisories based on forecasted pollution levels. For instance, if the model predicts high levels of particulate matter (PM2.5) on certain days, the agency can recommend that sensitive groups, such as children and the elderly, limit outdoor activities. Additionally, the agency can analyze historical data to identify correlations between AQI levels and hospital admissions, helping to allocate healthcare resources more effectively and efficiently.

UN Agencies, I/NGOs and Environmental consulting firms, tasked with assessing and advising on their environmental impact can utilize the AQI prediction model to help governments and companies understand the air quality implications of their operations. The firms can then recommend strategies for emission reduction, such as optimizing production schedules during lower pollution periods or implementing green technologies to mitigate their environmental footprint. This approach not only assists the firm's clients but also contributes to broader environmental goals.

6.6 Experiments

A. Missing value correlation map:

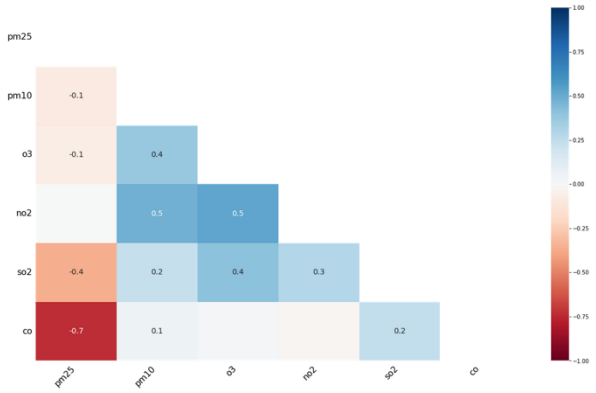

Figure 6: Missing value correlation heatmap

The missing value correlation heatmap visualizes nullity correlations, indicating how strongly the presence or absence of one variable relates to another. Variables that are consistently complete or entirely missing are excluded, as they provide no meaningful correlation.

Nullity correlation values range from -1 to 1:

- **-1**: Exact negative correlation, where the presence of one variable guarantees the absence of the other.
- **0**: No correlation, meaning the presence or absence of one variable has no effect on the other.
- **1**: Exact positive correlation, where the presence of one variable ensures the presence of the other.

This heatmap is valuable for identifying data completeness patterns between variable pairs, though it has limited capability for explaining broader patterns or supporting very large datasets.

In our dataset, a missing correlation of -0.7 for CO (carbon monoxide) suggests a missing-not-at-random pattern, where CO data gaps could indicate low PM2.5 levels or insufficient sensor data to capture particles. Our predictive model will incorporate this correlation to account for missing data patterns.

B. Industrial emission production

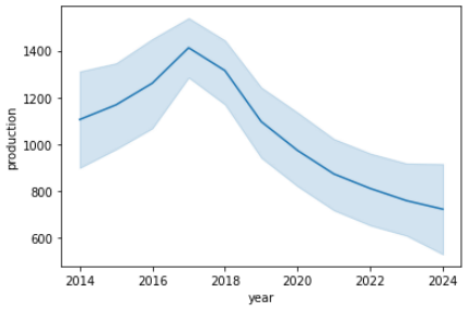

Figure 7: Annual cumulative industrial production

Evaluating the cause of air-pollution is challenging tasks, various test is performed to find the clear pattern, yet it is hard to trust those statistical tests because of high non-linearity and sentiments. For now, we have included all the features as we perform the model interpretation and find the behavioral pattern and importance of such features. The above figure shows the accumulated change in the volatility of industrial production in Thailand. Post 2018 show the down impact across the years. Such observable behaviors are tested below in order to check their predictive power.

C. Lag effect of industrial production with PM2.5

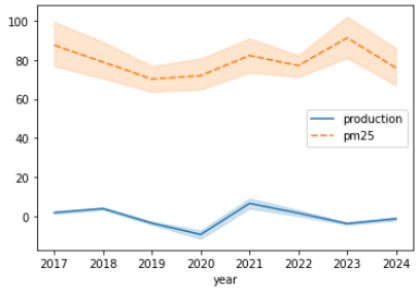

Figure 8: Lag effect of industrial production with PM2.5

While observing the average lag effects across industrial and pm25 dataset in monthly patterns as with changing spatial relation we are not able to encounter it we are using various locations across Thailand near Bangkok.

In our further analysis morphological maps at various levels will be applied. Coming back to industrial dataset, let's check the dependency power to PM2.5.

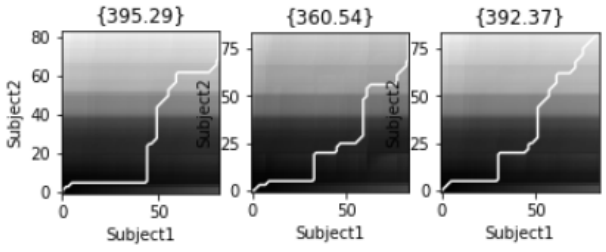

Figure 9: DTW (Discrete time warping) for Industrial production at 3 locations

We perform DTW test across three prominent locations (chulalongkorn-hospital, samut-

sakhon, thonburi-power), we can see that the initial impact is close to zero as we increase the lag factor the growing relation is shown, still we cant be sure unless the ww perform more tests and do model interpretability.

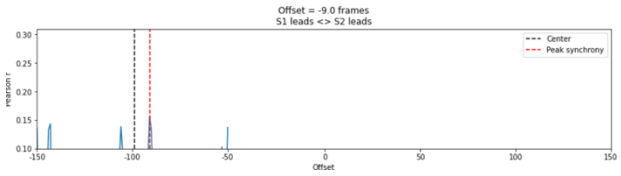

Figure 10: Changing time frame and peak-correlation

Another test to check the global relation we perform correlation plot across various frames in time, at most 15% correlation is at the negative offset i.e that lags are shown.

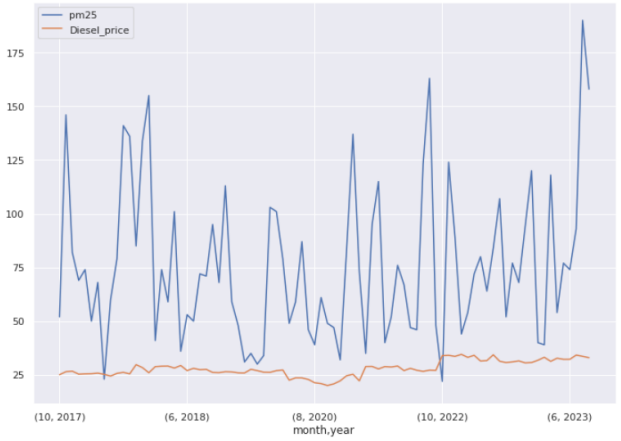

Figure 11: Diesel Price with PM2.5

Further we added more exogenous features , the dataset was crawled from Bangchak co-operation, one of the prominent fuel industries in Thailand, again by eye-balling we can observe a few patterns. To verify the effect we perform GARCH test.

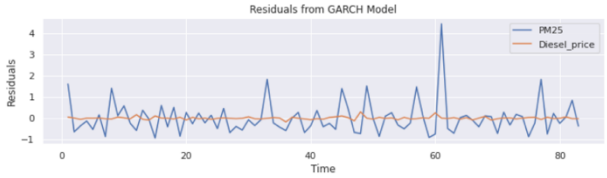

Figure 12: Showing the lag effect

```
Granger Causality
number of lags (no zero) 8
ssr based F test:         F=3.9827  , p=0.0008  , df_denom=58, df_num=8
ssr based chi2 test:   chi2=41.2006 , p=0.0000  , df=8
likelihood ratio test: chi2=32.8373 , p=0.0001  , df=8
parameter F test:         F=3.9827  , p=0.0008  , df_denom=58, df_num=8

Granger Causality
```

Performing GARCH (Generalised AutoRegressive Conditional Heteroskedasticity) test gave p-value 0.01 below , that means the lag effect can be 8 months though the effect is very slow in growth the partly reason is that fuel companies in thailand are using low-emission fuel. These features will encounter more randomness and uncertainty in our predictive variable that is the reason for understanding the behaviour of such dynamics in production.

Furthermore, we perform more research on the causes of air-pollution . More the consumptions of resources and the requirement for the nation is more the resource utilisation. Although , we are not able to find the dataset at such granular level. Energy consumption is one of the key factors.

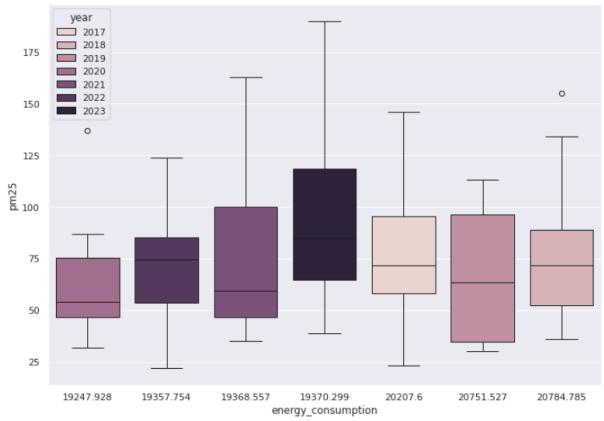

Figure 13: Energy consumption morphology per PM2.5

Ranging from 2107 to 2023 we can see the lastly in 2023 Thailand has started to rely on more energy utilisation, which can be a prominent feature for further development of our solution Although the changing electronic vehicle has the mitigating impact, but industrial production and consumption still cannot resolve the root cause of air-pollution.

Our future prospect is on:

- Geo-spacial behavioural pattern with
- Encouraging more local traffic
- Model building and evaluation
- Deployment and production testing.

References

[1] Ravindra, K., Singh, T., Pandey, V., and Mor, S. 2020. Air pollution trend in Chandigarh city situated in Indo-Gangetic Plains: Understanding seasonality and impact of mitigation strategies. *Science of The Total Environment*. 729, 138717. https://doi.org/10.1016/j.scitotenv.2020.138717.

[2] Expert Market Research. 2023. *Global Air Quality Monitoring Market Report and Forecast 2023-2028*. [Online]. Available: https://www.expertmarketresearch.com/reports/air-quality-monitoring-system-market

[3] IMARC Group. 2023. *Air Quality Monitoring Market: Global Industry Trends, Share, Size, Growth, Opportunity and Forecast 2023-2028*. [Online]. Available: https://www.imarcgroup.com/air-quality-monitoring-market

[4] Markets and Markets. 2023. *Air Quality Monitoring System Market Growth, Drivers, and Opportunities*. [Online]. Available: https://www.marketsandmarkets.com/Market-Reports/air-quality-monitoring-equipment-market-183784537.html