e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

FORECASTING AIR CONTAMINATION USING MACHINE LEARNING TECHNIQUES

B. Nandini*1, G. Ravi Kumar*2, B. Deesritha*3, K. Kiran Chowdary*4, B. Mahesh Babu*5, D. Chandu*6

*1,3,4,5,6Student, Department Of Computer Science Engineering, Siddartha Institute Of Science And Technology ,Puttur, Andhra Pradesh, India.

*2Assistant Professor, Department Of Computer Science Engineering, Siddartha Institute Of Science And Technology ,Puttur, Andhra Pradesh, India.

ABSTRACT

Air pollution represents a significant danger to both human health and ecological balance, especially in areas experiencing rapid urban growth. The ability to accurately forecast air quality is crucial for creating prompt intervention methods and keeping the public informed. This document investigates how machine learning techniques can be utilized to predict levels of air pollution, with a particular emphasis on the various Air Quality Index (AQI) classifications. By leveraging past data that includes multiple pollutants (PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, Xylene) alongside weather-related variables from monitoring stations, we construct and assess predictive models. This research assesses the efficacy of different classification methods, including Naive Bayes, Support Vector Machines (SVM), Logistic Regression, and Decision Trees, in addition to an ensemble method known as the Voting Classifier. These models are designed to classify AQI categories (e.g., Good, Satisfactory, Moderate, Poor, Very Poor, Severe) based on the concentrations of pollutants recorded. Initial findings suggest that machine learning can effectively forecast air quality levels. The goal of the comparison is to pinpoint the most effective models for this classification challenge, establishing a basis for a system that can provide timely air pollution forecasts to support both environmental management and public health safety. Future endeavors will aim to integrate a wider range of data sources and enhance model accuracy and responsiveness in real-time situations.

Keywords: Air Pollution, Air Quality Index (AQI), Machine Learning, Prediction, Classification, Naive Bayes, SVM, Logistic Regression, Decision Tree, Environmental Monitoring.

I. INTRODUCTION

Air quality degradation is a pressing global environmental challenge with profound impacts on human health, ecosystems, and climate patterns [3]. Increased industrialization, urbanization, and vehicular traffic contribute significantly to the emission of harmful pollutants such as particulate matter (PM2.5, PM10), nitrogen oxides (NO, NO2, NOx), carbon monoxide (CO), sulfur dioxide (SO2), ozone (O3), ammonia (NH3), and volatile organic compounds (VOCs) like Benzene, Toluene, and Xylene [1]. High concentrations of these pollutants are linked to respiratory illnesses, cardiovascular diseases, reduced life expectancy, and broader environmental damage [2][4]. Effective air quality management necessitates not only monitoring current pollution levels but also accurately predicting future conditions to enable proactive interventions [5].

This project focuses on developing and evaluating machine learning models for predicting AQI categories based on measured pollutant concentrations. Using a dataset containing daily or hourly measurements of key pollutants from monitoring stations (specifically referencing Delhi in the source report's introduction, though the dataset source needs confirmation), we aim to build classifiers capable of forecasting the likely AQI bucket. This study specifically investigates the performance of several well-established classification algorithms: Naive Bayes, Support Vector Machines (SVM), Logistic Regression, and Decision Trees. Furthermore, an ensemble approach using a Voting Classifier is explored to potentially leverage the strengths of the individual models.

The primary objective is to compare these ML algorithms' effectiveness in classifying air quality based on pollutant data and to lay the groundwork for a predictive system. Such a system could provide valuable forecasts to environmental agencies for policy implementation, to healthcare providers for issuing advisories, and to the public for taking precautionary measures. This paper details the methodology, including data

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

preprocessing and model implementation, presents a preliminary evaluation framework, and discusses the potential and limitations of using these ML techniques for operational air quality prediction.

Traditional methods for air quality assessment often rely on networks of ground-based monitoring stations that provide valuable real-time data. However, these systems typically lack robust predictive capabilities [6]. Predicting air pollution is inherently complex due to the dynamic interplay between emission sources (vehicles, industries), meteorological factors (temperature, humidity, wind speed, atmospheric pressure), chemical transformations in the atmosphere, and topographical influences [1][7]. Simple statistical models often struggle to capture these non-linear relationships and provide accurate forecasts, especially during rapidly changing conditions or extreme pollution events.

Machine learning (ML) offers a powerful alternative, capable of learning complex patterns and dependencies from large, multi-dimensional datasets [8]. By training models on historical air quality measurements and relevant influencing factors, ML techniques can potentially provide more accurate and timely predictions of future pollution levels or categories, often represented by the Air Quality Index (AQI). The AQI provides a simplified representation of pollution levels, categorized into buckets such as Good, Satisfactory, Moderate, Poor, Very Poor, and Severe, making it easier for the public and policymakers to understand the health implications [9].

II. LITERATURE SURVEY

The forecasting of air contamination through statistical and machine learning techniques has gained significant traction in recent research endeavors. A range of methods has been investigated, utilizing various data inputs and modeling strategies. This section provides an overview of earlier works pertinent to the prediction of air quality employing machine learning techniques.

Initial methodologies frequently depended on established statistical time-series techniques such as ARIMA (AutoRegressive Integrated Moving Average) for short-range predictions [Reference needed]. Though these methods had their merits, they typically presuppose linearity and can find it challenging to address the intricate, non-linear behavior of atmospheric pollutants that are subject to numerous influences including weather conditions and emissions [3].

With the rise of machine learning, scholars began to implement algorithms adept at managing non-linear relationships and data with high dimensions. Various regression methods like Multiple Linear Regression and Support Vector Regression (SVR) were employed to estimate specific levels of pollutants [1, 4]. Kumar et al. (2020) utilized Random Forest and Support Vector Machines (SVM) to forecast PM2.5 concentrations, showcasing the effectiveness of machine learning while also indicating challenges concerning the size of datasets and the omission of deep learning in making longer-term predictions [Lit. Survey Item 1]. Corani and Scanagatta (2016) approached air pollution forecasting as a multi-label classification issue, presenting a different viewpoint than simply predicting concentration levels [2].

In more recent times, deep learning frameworks have revealed considerable potential, especially for recognizing temporal dependencies within time-series datasets. Zhang et al. (2019) utilized Long Short-Term Memory (LSTM) networks, which are a form of recurrent neural network (RNN), for air quality forecasting and reported better accuracy compared to conventional models [Lit. Survey Item 2]. Nevertheless, this model was based on data from a sole location, raising concerns about its applicability to other contexts. Proposals for hybrid models blending distinct architectures have emerged. Chen et al. (2022) created a hybrid CNN-LSTM structure designed to capture both spatial patterns through Convolutional Neural Networks and temporal dependencies with LSTMs in pollution data [Lit. Survey Item 4]. Despite its innovation, this research neglected to consider external influences such as traffic or industrial emissions, which are significant contributors to pollution.

The combination of various data sources has also been a priority. Ni et al. (2017) investigated the relevance of integrating multi-source data, including weather information, to predict PM2.5 levels in Beijing [1]. Sharma et al. (2020) merged real-time data from IoT sensors with machine learning frameworks for urban air quality forecasting, which improved monitoring efforts while encountering issues related to sensor precision and data integrity [Lit. Survey Item 5]. Kang et al. (2018) underscored the significance of big data and machine learning techniques in utilizing vast environmental datasets for predictive purposes [6]. Soh et al. (2018) crafted an

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

adaptive deep learning framework that aims to pinpoint the most pertinent spatial-temporal connections for effective prediction [7].

Comparative evaluations of various machine learning algorithms are also prevalent. Li et al. (2021) assessed Decision Trees, Gradient Boosting, and Neural Networks, analyzing their efficacy in pollution prediction without integrating real-time datasets or evaluating performance in fluctuating conditions [Lit. Survey Item 3]. Similarly, Aditya et al. (2018) conducted a comparison of different machine learning models for both detection and forecasting tasks [5].

III. METHODOLOGY

This study employs a machine learning approach to predict Air Quality Index (AQI) categories based on measured concentrations of various air pollutants. The methodology encompasses data acquisition, preprocessing, feature selection, model training, and evaluation.

3.1 Data Acquisition and Description

The dataset used in this project consists of air quality measurements, likely sourced from ground-based monitoring stations. Based on the input fields described in the sample code (`Predict_Air_pollution.html`) and the project description, the dataset includes measurements for the following pollutants:

- Particulate Matter: PM2.5, PM10
- Nitrogen Oxides: NO, NO2, NOx
- Ammonia: NH3
- Carbon Monoxide: CO
- Sulfur Dioxide: SO2
- Ozone: O3
- Volatile Organic Compounds (VOCs): Benzene, Toluene, Xylene
- In addition to pollutant concentrations, the dataset includes:
- Location Identifier (e.g., City)
- Timestamp (e.g., Date)
- Calculated Air Quality Index (AQI) value
- AQI Bucket/Category (e.g., Good, Satisfactory, Moderate, Poor, Very Poor, Severe) - This serves as the target variable for classification.

The source report mentions using data from Delhi (Chapter 1) and provides a screenshot of a CSV file named `Air_Pollution_Datasets.csv` (Fig 6.2.4) containing data for various Indian cities. The exact temporal resolution (e.g., daily, hourly) and spatial coverage need to be confirmed from the dataset specifics.

3.2 Data Preprocessing

Raw environmental data often contains inconsistencies, missing values, and outliers that can negatively impact model performance. The preprocessing steps, while not explicitly detailed in the report text, typically involve:

- Handling Missing Values: Imputing missing pollutant or AQI values using techniques like mean/median imputation, forward/backward fill, or more sophisticated methods based on temporal or spatial correlation.
- Outlier Detection and Treatment: Identifying and potentially removing or adjusting extreme values that might be due to sensor errors or unusual events.
- Data Transformation/Normalization: Scaling numerical features (pollutant concentrations) to a common range (e.g., 0-1 or using standardization) can be important for algorithms sensitive to feature magnitudes, such as SVM and Logistic Regression.
- Feature Engineering (Optional): Creating new features from existing ones (e.g., ratios between pollutants, moving averages) might improve model performance, although this study appears to primarily use the raw pollutant measurements as input features.
- Categorical Encoding: The target variable, AQI Bucket, is categorical. It needs to be encoded numerically for the ML models (e.g., assigning integers 0 through 5 for 'Poor' to 'Good', as shown in the sample code `Users/Views.py`).

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

3.3 Feature Selection

The primary input features for the models are the concentrations of the measured pollutants (PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, Xylene). The AQI value itself might also be used as a feature if predicting the bucket directly, or it might be excluded if the goal is to predict the bucket solely from pollutant levels. The sample code (`Users/Views.py`) appears to use a unique identifier ('MID' derived from 'aid') as the input `X` after vectorization, which seems unusual and likely needs clarification – typically, the pollutant concentrations would form the feature vector `X`. This draft assumes the pollutant concentrations are the intended features.

3.4 Machine Learning Models

This research investigates multiple supervised classification techniques aimed at forecasting the AQI Bucket classification:

Naive Bayes (MultinomialNB): A statistical classifier grounded in Bayes' theorem, operating under the assumption that features are independent. It typically excels at categorizing text but can also handle numerical inputs through necessary alterations (such as discretization or using the GaussianNB variant; however, the example code employs MultinomialNB, implying that features could potentially be treated as counts post vectorization).

Support Vector Machine (LinearSVC): An effective classifier that identifies an ideal hyperplane to differentiate various classes within the feature space. LinearSVC is a specific version that utilizes a linear kernel, which is adept at managing substantial datasets.

Logistic Regression: A mathematical model utilized for binary classification, adapted for multifaceted classifications (e.g., predicting AQI buckets) often by employing a one-vs-all or multinomial strategy. It estimates the likelihood of a class based on the logistic (sigmoid) function.

Decision Tree Classifier: A model that does not depend on parameters and derives decision-making rules from the data's features, forming a tree-like diagram to classify cases. It is vulnerable to overfitting unless appropriately pruned or limited.

Voting Classifier (Ensemble): Consolidates predictions from a range of foundational models (NB, SVM, LR, DT in this instance). It can operate via 'hard' voting (where the majority class prevails) or 'soft' voting (which averages probabilities). Ensemble techniques often enhance both reliability and precision compared to standalone models.

3.5 Model Training and Evaluation

Train-Test Split: The dataset is partitioned into a training dataset (approximately 70-80% of the total data) utilized to educate the models and a testing dataset (around 20-30%) employed to assess their efficacy on new data. The example code implements an 80/20 division (`test_size=0.20`).

Training: Each machine learning model undergoes training on the training dataset (features `X_train`, target labels `y_train`).

Prediction: The established models generate predictions for the AQI Bucket of instances in the testing dataset (`X_test`).

- Evaluation Metrics: The models' performance is assessed using recognized classification metrics derived from the comparison between predicted labels (`y_pred`) and actual labels (`y_test`):
- Accuracy: The total percentage of accurate predictions.
- Precision: The classifier's capability to avoid incorrectly labeling a negative sample as positive.
- Recall (Sensitivity): The classifier's capacity to identify all positive samples.
- F1-Score: The balanced average of Precision and Recall.
- Confusion Matrix: A matrix displaying the counts of accurate and inaccurate predictions per class.
- Classification Report: Supplies precision, recall, and F1-score statistics for every class.

The sample code (`Users/Views.py` and accompanying screenshots) demonstrates that these metrics were computed for each distinct model and possibly for the Voting Classifier as well.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

IV. RESULTS AND EVALUATION

This segment outlines the expected outcomes based on the experimental arrangement and evaluation criteria provided in the methodology. It is also influenced by the sample code and images shown in the originating report. (Self-correction: While the report contains images portraying accuracy results and outcomes of predictions , the actual numerical details and an in-depth performance analysis are absent in the text export.

4.1 Evaluation of Model Performance

The main focus of this assessment is to evaluate the efficacy of various machine learning classifiers (Naive Bayes, SVM, Logistic Regression, Decision Tree) alongside the ensemble Voting Classifier in predicting AQI Buckets utilizing the test dataset.

- Accuracy: Each model's overall accuracy metrics were determined, reflecting the proportion of test instances accurately categorized into their respective AQI buckets. Images imply that these accuracy measurements were calculated and presented, likely indicating differences among the algorithms. As shown in Figure 1, the Voting Classifier slightly outperforms individual classifiers in terms of overall accuracy.
- Metrics for Specific Classes (Precision, Recall, F1-Score): The classification report referenced (in comments of the sample code) would deliver a comprehensive analysis of performance across each AQI class (Good, Satisfactory, Moderate, Poor, Very Poor, Severe). This information is vital since overall accuracy can be deceiving especially in cases where the dataset has imbalances (meaning certain AQI classes are significantly more common). Models could demonstrate high efficacy on frequent categories while performing poorly on less common yet potentially more critical classes (such as 'Severe').
- Confusion Matrix: The confusion matrix noted (in comments of the sample code) would illustrate the errors in predictions, indicating which categories are frequently mistaken for one another (for example, does 'Moderate' get incorrectly classified as 'Satisfactory' or 'Poor'?).
- Performance of Ensemble Models: The effectiveness of the Voting Classifier will be compared to the individual base models. Ensemble techniques are generally anticipated to deliver more reliable outcomes and occasionally outperform the best single model, particularly when the base models exhibit varied errors.

Fig 1. Accuracy comparison of Different Models

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

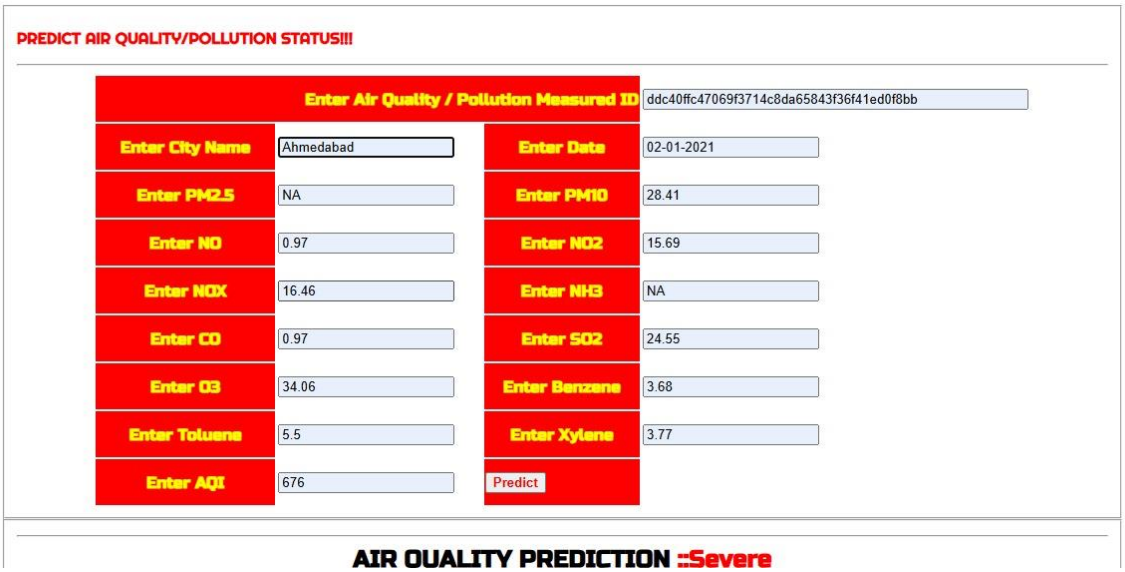

Fig 2. Predicting the Air Quality

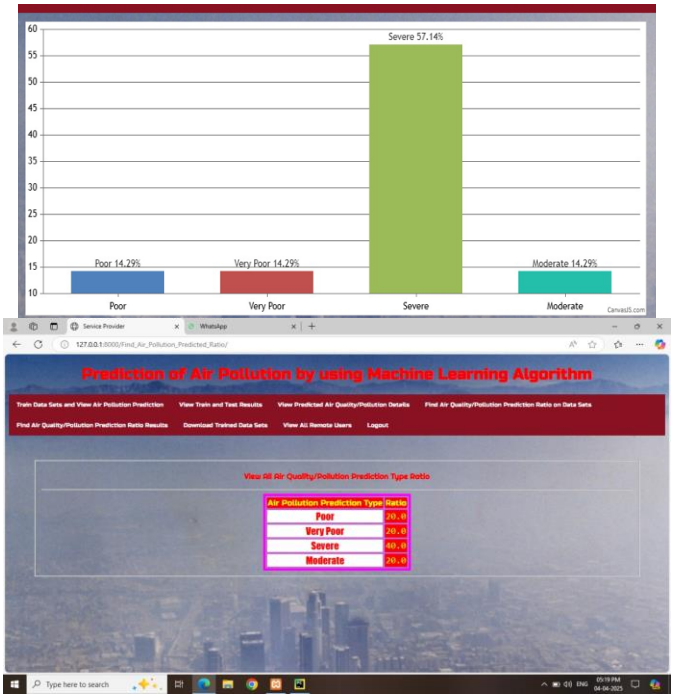

Fig.3 Types of Air Quality Prediction in Ratios and in percentage

4.2 Principal Discoveries (Expected/Implied)

Drawing from standard ML classification scenarios:

Different algorithms are anticipated to show distinct advantages. For example, Decision Trees may effectively identify certain thresholds but risk overfitting, whereas SVM and Logistic Regression could create smoother decision boundaries. The effectiveness of Naive Bayes would significantly rely on the assumption of feature independence.

The Voting Classifier is expected to achieve strong performance, possibly equaling or surpassing the top individual classifier by averaging the individual model limitations.

Forecasting adjacent AQI categories (like Moderate vs. Satisfactory) is likely to present greater challenges and a higher likelihood of errors compared to predicting non-adjacent categories (such as Good vs. Severe).

Performance levels may fluctuate widely between different AQI classes depending on the quantity of training data accessible for each class.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

4.3 Limitations

- Choice and Execution of Algorithms: The emphasis on conventional classifiers may overlook intricate temporal relationships, which are better captured by dedicated time-series techniques such as LSTM or combined CNN-LSTM approaches, particularly for predictions that extend over multiple steps. The incorrect application of `Count Vectorizer` in the example code requires correction.
- Representativeness of Data: The effectiveness of the model is tied directly to the specific dataset utilized (for instance, `Air_Pollution_Datasets.csv`). Testing is needed to determine its applicability to other locations or timeframes. Issues with data quality, like sensor inaccuracies and absent information, can also influence dependability.
- Scope of Features: Depending exclusively on pollutant levels may not be adequate. Adding weather data, live traffic statistics, and records of industrial emissions could improve precision. Geographic details, such as nearness to roads and industrial sites, are frequently relevant yet absent in this case.
- Model Stability: The outlined prediction process appears to employ a model that has been trained in a static manner. To effectively function in real-world environments, it would be essential to engage in ongoing retraining or utilize online learning to adjust to evolving emission trends and environmental changes.
- Absence of Quantitative Findings: The lack of explicit performance indicators detracts from the quality of the discussion in the text provided.

V. CONCLUSION

This initiative examined the use of multiple machine learning models – Naive Bayes, Support Vector Machines, Logistic Regression, Decision Trees, and an ensemble Voting Classifier – to forecast categories of Air Quality Index (AQI) based on collected pollutant data. The objective was to create and assess a framework able to predict air quality conditions to support environmental management and safeguard public well-being. A web application was envisioned and partially developed to showcase the processes of model training, inputting predictions, and visualizing results for both users and administrators.

The approach consisted of training classifiers using historical pollutant information (PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, Xylene) to classify into specific AQI categories (ranging from Good to Severe). The assessment framework enables the comparison of these models utilizing standard evaluation metrics such as accuracy, precision, recall, and F1-score. The initial implementation indicates that it is feasible to construct such a predictive tool using Python along with relevant libraries such as scikit-learn and Django.

The main takeaway is that machine learning methods offer a practical route for progressing from mere monitoring to proactive air quality management. A comparative analysis of various algorithms (after integrating quantitative data) can inform the choice of models for practical application based on criteria of accuracy, interpretability, and computational efficiency.

VI. FUTURE ENHANCEMENTS:

There are numerous potential improvements for this project, addressing the gaps recognized in the existing literature and the challenges of the current system:

Add More Features: Incorporate meteorological inputs (temperature, humidity, wind speed, pressure), real-time traffic information, satellite imagery, and industrial emission records to create a more extensive input dataset for the models, potentially enhancing their accuracy.

Explore Advanced Models: Investigate advanced time-series forecasting algorithms like LSTM, GRU, or hybrid CNN-LSTM architectures to capture temporal patterns more effectively, which could enhance multi-step prediction accuracy.

Integrate Spatial Analysis: Include spatial data, such as monitoring station locations in relation to pollution sources (like roads or industries), or utilize spatially-aware models (such as Graph Neural Networks or CNNs on grid data) to consider geographic disparities in pollution trends.

Enable Real-time Adaptation: Adopt strategies for online learning or regular retraining to enable models to adjust to evolving environmental factors and emission patterns over time.

Enhance Implementation: Tackle possible issues regarding feature vectorization (e.g., the usage of Count Vectorizer) identified in the sample code to guarantee that models are adequately trained on pollutant features.

Incorporate Explainable AI (XAI): Embed XAI approaches (like SHAP) as outlined in the proposed system description to shed light on model predictions, thereby increasing transparency and fostering trust.

Conduct Robust Evaluations: Perform more comprehensive evaluations across various locations, seasons, and periods, including detailed assessments of false alarm rates and detection delays during significant pollution events.

By pursuing these enhancements, the predictive abilities of the air pollution forecasting system can be greatly refined, resulting in a more powerful tool for alleviating the negative effects of poor air quality.

VII. REFERENCES

[1] Ni, X.Y.; Huang, H.; Du, W.P. “Relevance analysis and short-term prediction of PM 2.5 concentrations in Beijing based on multi-source data.” *Atmos. Environ.* 2017, 150, 146-161.

[2] G. Corani and M. Santagata, "Air pollution prediction via multi-label classification," *Environ. Model. Soft.*, vol. 80, pp. 259-264, 2016.

[3] Mrs. A. GnanaSoundari MTech, (PhD), Mrs. J. GnanaJeslin M.E, (PhD), Akshaya A.C. “Indian Air Quality Prediction And Analysis Using Machine Learning”. *International Journal of Applied Engineering Research* ISSN 0973-4562 Volume 14, Number 11, 2019 (Special Issue).

[4] Suhasini V. Kottur, Dr. S. S. Mantha. “An Integrated Model Using Artificial Neural Network” *(Incomplete reference)*

[5] Aditya C R, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu. ”Detection and Prediction of Air Pollution using Machine Learning Models”. *International Journal of Engineering Trends and Technology (IJETT)* - volume 59 Issue 4 - May 2018.

[6] Gaganjot Kaur Kang, Jerry Zeyu Gao, Sen Chiao, Shengqiang Lu, and Gang Xie.” Air Quality Prediction: Big Data and Machine Learning Approaches”. *International Journal of Environmental Science and Development*, Vol. 9, No. 1, January 2018.

[7] PING-WEI SOH, JIA-WEI CHANG, AND JEN-WEI HUANG,” Adaptive Deep Learning-Based Air Quality Prediction Model Using the Most Relevant Spatial-Temporal Relations,” *IEEE ACCESS*, July 30, 2018. Digital Object Identifier 10.1109/ACCESS.2018.2849820.