Grenze International Journal of Engineering and Technology, June Issue

AI Tool for Personalised Recommended System on Forecasting Air Pollutants

Ashu[1], Rakesh Raushan[2] and Abhishek Gupta[3]
[1-3]Department of Computer Science and Engineering, Greater Noida Institute of Technology, Greater Noida, Uttar Pradesh, India
Email: ashubaisoya1986@gmail.com, rakeshraushan129@gmail.com, abhishekgupta1864@gmail.com

Abstract*— Air pollution is one of the worst enemies of public health, as it leads to respiratory and cardiovascular diseases while reducing average life expectancy. Most air-quality monitoring systems existing today provide only real-time values of pollution and cannot reliably forecast future concentration, thus failing to give advance warning to the people. This leads to a requirement of a system which assists individuals in predicting risks of ill health due to low air quality. This paper, therefore, introduces an AI-based approach to forecasting air pollution that calculates the levels of various pollutants for a future time by using historical air-quality records and important meteorological variables such as temperature, humidity, and wind speed. Linear regression and random forest algorithms examine existing data to make predictions of pollution levels on the meteorological parameter. The results obtained indicate that accurate predictions are offered with these models and that particulate matter, especially Particulate matter2.5 and Particulate matter10, is an important factor in poor air quality. Based on the levels of predicted pollution, the system provides general health advisories which apprise users of possible risks, including respiratory or cardiovascular discomfort during high-pollution conditions. This approach supports the creation of early awareness and encourages precautionary actions in order to reduce the threat of health risks related to exposure.

***Keywords*— Air Pollution Prediction, Machine Learning, Linear Regression, Random Forest, Meteorological Variables, Particulate matter2.5, Particulate matter10.**

I. INTRODUCTION

Air pollution has proven to be one of the most pressing issues in contemporary times. The current increased pace of industrialization, urbanization, and energy production/consumption has increased the levels of toxic gases such as $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, and O_3 in the atmosphere, which adversely affect air quality globally [1], [10]. Long exposure to these air pollutants leads to respiratory ailments, cardiovascular diseases, early death, and premature death of millions of people each year, which is emphasized by the World Health Organization. Therefore, an efficient air quality monitoring/forecasting tool is the need of the hour.[2], [8]. The conventional technique used for air quality analysis is based on fixed-ground stations with accurate results but poor spatiotemporal representativeness and forecasting potential [6]. Although mathematical and statistical models have been developed to make predictions for future concentrations of pollutants, they face difficulties in handling nonlinear characteristics in atmospheric phenomena and need intensive calculations or domain knowledge inputs [2], [16].

Grenze ID: 01.GIJET.12.2.51_20

AI-based models, especially machine learning and deep learning models such as Random Forest, Linear Regression, LSTM, CNN-LSTM have proved their better potential in dealing with intricate spatiotemporal variations in air pollutants and have a better forecasting capacity than conventional techniques [3], [13].

Although these developments have occurred, most available air quality forecasting systems provide generalized air quality notifications without providing early awareness or substantial insight into health risks [18]. Research such as Patel [18] and Jalali et al. [15] have highlighted air quality coverage, dependency in regions, and insufficient awareness of health risk concerns as core issues. The conception of AI-powered forecasting systems will address these existing issues of inadequate forecasting accuracy for air quality with early awareness of health risks associated with exposure.

II. Literature Studies

Air pollution has recently become a serious public health issue in India. A major share of people is getting affected by improper air, which is resulting in more than 1.1 million premature deaths in a year and is among one of the leading causes of deaths in a country.

Mottahedin et al. [8] presented a systematic review and performance evaluation of machine learning–based air quality prediction techniques. The authors note that more than 90% of the world's population is exposed to unsafe levels of pollution, hence there is growing global concern. Air pollution is directly concern about severe health disorders, such as respiratory and cardiovascular diseases. Current literature classifies the models for forecasting into mechanistic and data-driven techniques. However, the main drawbacks of the mechanistic models are linked with nonlinear behaviour of pollutants; thus, ML and AI methods such as Random Forest, SVR, Gaussian Processes, and Neural Networks are gaining increased attention. Previous studies prove that ANN, LSTM, and hybrid models based on ANFIS-GA have better performance for pollutant gases such as $PM_{2.5}$, CO, and NO_2 [7]. The ANFIS model has also been successfully used in many regions, including Tehran, Jeddah, Karnataka, and Romania. In their comparative study using the dataset of Semnan, Iran, Mottahedin et al. [9] reported that the ANFIS model was the most accurate in all seasons, with an R^2 as high as 0.999 and minimum RMSE and MAE values.

The works of Jalali et al. [15] presented a review of prevailing air quality prediction techniques. This strongly emphasized major concerns in handling aspects of scalability, computational complexities for real-time applications in developing countries like Afghanistan. This paper strongly emphasizes the importance of effective models for real-time air forecasting on the reporting of WHO documents that air pollution has resulted in close to seven million deaths every year. Conventional models like ARIMA, linear regression, and SVM [2], perform very poorly in capturing nonlinear pollutant behavior. Suggestedly relatively more ancient models on ML using Random Forest and XGBoost have shown outstanding performance on accuracy; however, they strongly lack interpretability and regional variations. Deep learning approaches such as CNNs, LSTMs, and ConvLSTM enhanced spatial–temporal learning [9], [12], [14]; however, they remain very computationally intensive and hard to operationalize. Although transformer-based techniques like AirFormer and HEART improved the long-term forecasting, their high hardware demand limits the use of these models in resource-constrained environments. To address these issues, the authors in this paper propose a hybrid approach to combine ML and DL techniques that outperformed the state-of-the-art and are feasible for deployment.

Subramaniam et al. [2] present a comprehensive review on the application of artificial intelligence in forecasting air pollution and mitigating associated health risks. The study outlines that air pollution has remained the significant concern of the world, which has caused nearly seven million deaths each year, recorded with the highest impact in South and East Asia. This paper highlights key sources of this pollution; these include emissions from the transportation sector, the burning of fossil fuels, industries, as well as natural sources like forest fires and volcanic eruptions, which release harmful material like $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, and ozone into the atmosphere. These are further linked with the causation of respiratory illness, cardiovascular disease, lung cancer, and cognitive impairment. The authors report the transition from the traditional statistical models of ARIMA [8], Multiple Linear Regression, and Grey Models that have performed poorly in nonlinear pollution patterns to advanced machine learning and deep learning approaches, including ANN, LSTM, GRU, Random Forest, and SVM [3]. Hybrid models combining decomposition, optimization, and nonlinear learning techniques tend to achieve higher accuracy and include the likes of CEEMDAN-LSSVM [4], Wavelet Neural Networks, and PCA-CS-LSSVM. Other AI applications are found in intelligent monitoring, early warning systems, climate modelling, and disease prediction and include studies on air pollution exposure juxtaposed to COVID-19 vulnerability. Nevertheless, computational requirements, data limitations, and real-time implementation remain challenging tasks well into the future.

TABLE I. SUMMARY OF RELATED WORKS IN AI TOOL FOR PERSONALIZED RECOMMENDED SYSTEM ON FORECASTING AIR POLLUTANTS

Serial No.	Author & Year	Models or Approaches	Dataset Used	Algorithms	Gaps
[1]	Subramaniam 2022	Traditional statistical, Artificial Intelligence, Hybrid, Three dimensional	Statistical Historical data	XGBoost	Lacks of comprehensive data and focus only on PM2.5 and do not integrate with real time data
[2]	Mottahedin 2024	Random forest, M5P, Random Tree	Semnan city dataset	Random forest, Gaussian Processes, CNN, KNN	Focus on only PM2.5 apart from others and single city dataset
[3]	Jalali et al. 2025	Random forest, XGBoost, LSTM, TSMixer, CNN	NEPA, openWeather	XGBoost, LSTM, CNN	Limited on two datasets

Above show a comparative analysis of existing literature on AI-powered air pollution forecasting systems conducted in previous studies. The table illustrates the authors covered in existing studies, datasets used in previous research, algorithms used in existing studies, and research gaps discovered in existing studies. As it can be observed in (Table 1), most existing studies cover a shortlist of air pollutants such as $PM_{2.5}$, have regional or small datasets, and do not integrate real-time datasets, resulting in systems with poor interpretability or health awareness capabilities.

III. PROBLEM STATEMENT

A. Prediction of Air Pollutants

It refers to estimation future concentrations of harmful air pollutants such as PM2.5, PM10, and gases using the analysis of past records of air pollution and meteorological elements.

B. Predication-based Awareness

This will be achieved through utilizing the estimated levels of pollution to inform people of the future air quality and health hazards associated with it.

IV. SOLUTION METHODOLOGY

The methodology used in this project involves a systematic approach to estimate levels of air pollution and achieve awareness based on air pollution. To start with data on air quality and meteorology is collected from reliable sources such as government sites and environmental websites. The air quality record contains information on previous levels of pollutants, responsibilities, and meteorology factors such as temperature, humidity, and windspeed. The collected information is cleaned and pre-processed to handle missing information in the record, remove inconsistencies in records, and achieve uniformity in record formats. Data preprocessing involves securing normalization and scaling to accommodate machine learning models. The analysis of data exploration is performed using statistics and visualization to understand information on nature and relationships of different pollutants and other elements in the atmosphere. After preprocessing, time variables and lag variables are used in this project to address time series in air pollution records. The pre-processed dataset is used for training machine learning and deep learning algorithms with an objective of making predictions related to future concentrations of air pollutants in the atmosphere. The performance of algorithms is tested by using evaluation metrics in order to verify accuracy levels of predictions. Eventually, recommendations related to pollution and health awareness messages are generated with predictions related to concentrations of air pollutants.

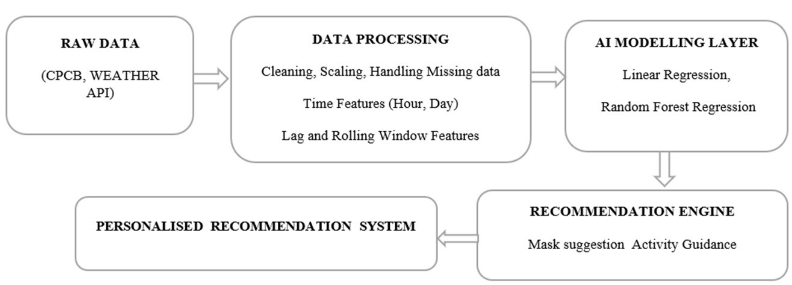

Figure 1. Personalized Air Pollution Forecasting

A. Data Collection

It involves data acquisition as the first step in the methodology to offer accurate and relevant information necessary for predicting air quality. In this case, historical information concerning air pollutants such as PM2.5, PM10, NO_2, CO, and O_3 is openly acquired from CPCB and other environmental datasets offered publicly by the government. Furthermore, information concerning meteorology, such as temperature, humidity, and wind speed, is also collected in this step because meteorology is responsible for the elemental factor resulting in air pollution levels.

B. Data Cleaning and Preparation

Cleaning and processing are major steps in ensuring the quality and validity of the datasets obtained after carrying out data collection. Missing records in a dataset, duplicates, and inconsistencies in a record are all appropriately dealt with. Data is processed in order to make it suitable for machine learning models. Exploration of data is conducted in order to have an insight into the nature of varying pollutants, as well as understanding the existing relationships with meteorological factors. Data cleaning will ensure that meaningful information is used in making a prediction.

C. Forecasting Pollution with the AI Models

Such cleaned datasets will be used to train various machine learning models that can predict future concentrations of air pollutants. In this case, it will employ a variety of learning models, such as Random Forest and Linear Regression, which can learn patterns using historical pollution information coupled with weather information. Such models are supposed to examine historical information with a view to providing a forecast on future concentration expectations either hourly or daily.

D. Recommendations based on Pollution

Based on the projected level of pollution, messages concerning general health awareness and advice are produced. For example, if the projected level of pollution is high, the system will advise people to reduce exposure to outdoor environments, wear face masks, or not to open windows. Through this, it operates based on the level of pollution without considering individual profiles, which is intended to make people aware of any possible health hazard by highlighting preventive measures.

E. Model Evaluation

The system calculates the performance of the forecasting models based on comparisons of both the forecasted level of pollution and the actual level of observed pollution. Some standard metrics will be used in order to assess model performance, which include RMSE, MAE, and R^2 score.

F. Continuous Learning

The system calculates the performance of the predictive models based on comparisons between the predicted results for pollution and the actual observed results recorded in the ground truth. To assess the performance, standard measures will be used, which include RMSE, MAE, and R^2 Score. The assessment will help in judging the level of accuracy and reliability linked to the predictive models.

V. RESULT AND SENSITIVITY ANALYSIS

A. Average Pollutant Levels

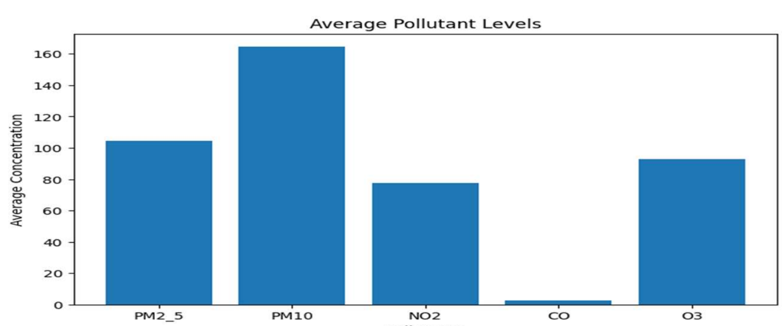

Figure 2. Average Pollutant Level

From the Figure 2, it is evident that PM10 has the highest mean concentration of all the air pollutants; therefore, its contribution to overall air pollution is considerable. PM2.5 and O_3 also have considerable average concentrations of pollution, which may be a source of considerable health hazards. However, the average concentrations of NO_2 are moderate, while concentrations of CO are lower. These findings indicate that particulate matter is a matter of concern in the region and the need to incorporate it into a model related to air pollution predictions.

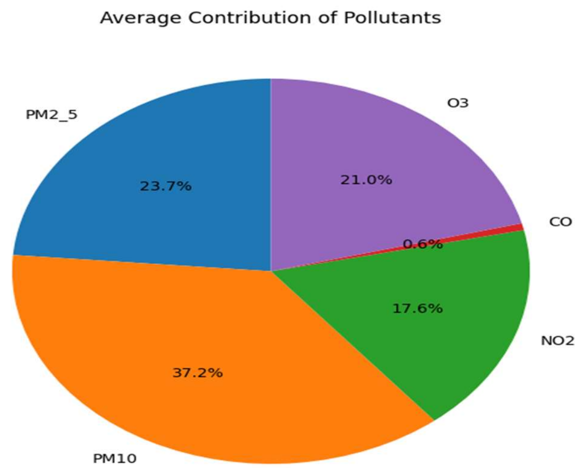

Figure 3. Average Contribution of Pollutants

C. State-Wise Pollution Level

When comparing the concentrations of the pollutants on a state-by-state from the below figure 4, there are considerable variations in the quality of the air in different regions. It has been exhibited from figure 4 that the concentrations of PM10 is the highest in all states on average, and there are high concentrations in the states of Maharashtra, Karnataka, and Tamil Nadu. This is expected due to the high emissions from dust, traffic, and industries. The concentrations of PM2.5 are also high in different regions, while the concentrations of NO2 are lower in states like Delhi and Uttar Pradesh due to their high level of urbanization.

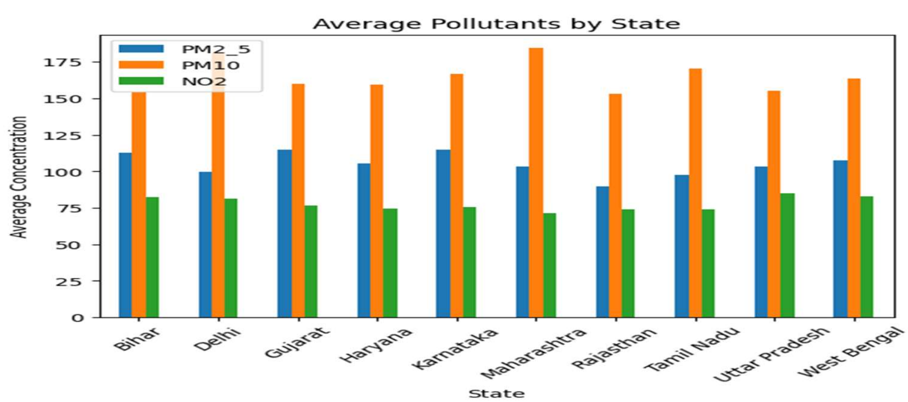

Figure 4. State-wise Average Pollution Level

D. Classification and Guidelines for PM2.5 Air Quality

From the figure 5, it can be observer that the PM2.5 air quality classification directly links the predicted concentration of the air pollutants ranges with corresponding health impact categories, ranging from good to very hazardous. Each health category is linked to the specific suggested messages. This fills the gap between the predicted results and health consciousness since it is used to link the PM2.5 values to messages. The air quality classification enables one to better understand the health hazards and the preventive measures needed to avoid health hazards when air pollution levels are high.

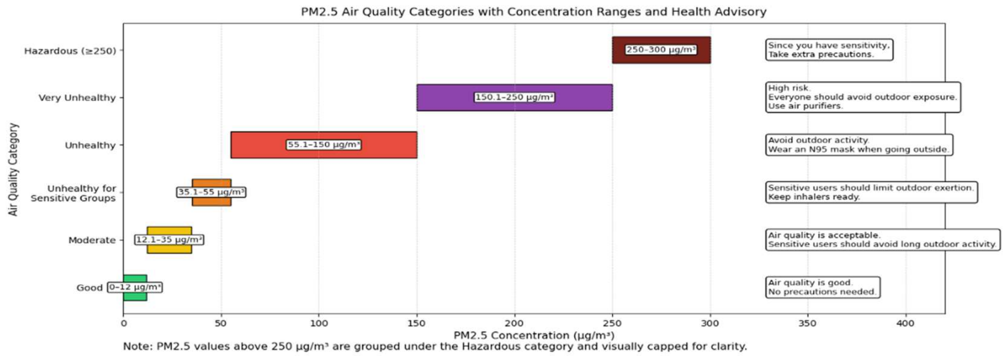

Figure 5. Classification and Guidelines for PM2.5 Air Quality

The results obtained from all the visual analyses gives a clear indication about the intensification of air pollutants in the region of study. Bar charts and pie charts give a clear indication that PM10 and PM2.5 are the major air pollutants that intensify the degradation of the overall air quality, with minimal contribution from CO. Distribution charts give a clear indication that the variables related to the pollutants and meteorological parameters are balanced and hence useful for predictive modelling. The result obtained from the state-wise comparison study reveals the overall regional variations in the intensification of the different air pollutants.

The overall PM2.5 air quality indicator translates the predicted values into significant categories that are related to the intensity of the phenomenon and hence give overall significant indications that are useful for the development of effective awareness and preventive programs related to air pollution through the technical outcomes obtained from the study.

VI. Discussion

It has been out that from this study it finds the complexity of air pollution issues, and emphasized the use of data analysis for air pollution forecasting. From its analysis, it can be seen that particulate matter, specifically PM10 and PM2.5, are major contributors to air pollution, since they can significantly reduce air quality, and this has already been identified in various studies related to the environment. The difference in levels of pollution between states again indicates that regional factors like urbanization, traffic flow, and industrial emission have a significant influence on air quality. Similarly, the conversion of predicted levels of PM2.5 into air quality classes can help reduce the gap between prediction and perception. The obtained results imply that machine learning-based approaches for forecasting can address complicated trends in air pollution effectively. That being said, the tool proves its effectiveness in identifying trends for the improvement of awareness, but at the same time, it underscores the need for refreshing the data on a periodic basis, based on the drastic fluctuations that may occur in the nature of the air pollution level.

VII. Conclusion

It has been shown that the data-driven model is effective in air pollution quality prediction and risk analysis. The proposed system has utilized the existing historical data for air quality, in addition to the essential parameters of temperature, humidity, and wind speed, and achieved the objective of modelling the behaviour of the pollutant. The implementation of Linear Regression and Random Forest helped in the forecasting of the levels of the major air pollution components. The experimental outcomes have shown that particulate matter, especially $PM_{2.5}$ and PM_{10}, are major factor in determining overall air quality, which is in accordance with air quality trends in various regions. The Random Forest algorithm particularly showed its effectiveness in dealing with nonlinear relationships between meteorological parameters and pollutant concentrations. In addition to the capability of the system to make forecasts, the system has the ability to link the forecasted concentration of PM2.5 with air quality indices and related health advice. Based on the link provided by the system between the forecasted concentration of air pollutants, such as PM2.5, and health advice, the proposed solution not only enables early warning but also facilitates health-enhancing decisions by providing quantitative forecasts with a risk category or health advice, such as restricting outdoor activities or the use of a protective mask. Therefore, the proposed solution can assist in health informatics by providing a tool that can reduce health risks related to exposure.

LIMITATION AND FUTURE SCOPE

Although the proposed system performs effectively in predicting air pollution trends, it has certain limitations. The system design predicts air pollution trends effectively but does not measure individual disclosure, which limits exact health risk assessment. It is also dependent on historical data. While the other models learn from past pollution and weather patterns for prediction, they may not capture sudden or unexpected pollution events accurately such as activities related to construction, traffic congestion, or crackers emissions. Additionally, data quality and availability can vary across regions, and missing or inconsistent data obtained from sensors may degrade prediction accuracy.

The system can be enhanced by other machine learning algorithms such as XGBoost, LightGBM, and deep learning models just to improve prediction accuracy and robustness. Through Integration of real-time sensor data and satellite analysis can be helpful in identify vast changes in pollution levels. Alert mechanism can be integrated with prediction of high pollutant and health risk actions by defining them with precautions on different platform based on login on that dashboard. These changes and improvements in system, they can make it more reliable, scalable, and useful for large-scale air quality monitoring and public health awareness.

REFERENCES

[1] B. Liu et al., “Application of an RR–XGBoost combined model for data-driven particulate matter prediction,” Scientific Reports, 2021.

[2] S. Subramaniam, N. Raju, A. Ganesan, N. Rajavel, M. Chenniappan, C. Prakash, A. Pramanik, A. K. Basak, and S. Dixit, “Artificial intelligence technologies for forecasting air pollution and human health: A narrative review,” Sustainability, vol. 14, no. 8, pp. 1–22, 2022.

[3] T. D. Morapedi, M. M. Ramoelo, and A. M. Botai, “Air pollution PM2.5 prediction in South African cities,” Front. Artif. Intell., vol. 6, pp. 1–15, 2023.

[4] Q. Guo et al., “Prediction of Hourly PM2.5 and PM10 Concentrations in Chongqing,” AAQR, 2023.

[5] H. D. Tran et al., “Forecasting hourly PM2.5 concentration with an optimized model,” ScienceDirect / journal, 2023.

[6] A. Setianingrum, “Prophet model performance analysis for Jakarta air quality,” in Proc. Int. Conf. Environ. 2023.

[7] F. Arnaut et al., “Short-term forecasting of PM10 and PM2.5 concentrations,” Conference/PDF, 2023.

[8] S. Mottahedin, M. Ghafouri, M. Allahbakhshi, and N. Ayoobian, “A systematic review and experimental evaluation of machine learning-based air quality prediction techniques,” Environ., 2024

[9] X. Huang, “PM2.5 prediction based on attention mechanism and Bi-LSTM,” SPIE / proceedings, 2024.

[10] Rautela, K. S., & Goyal, M. K. (2024). Transforming air pollution management in India with AI and machine learning technologies. Scientific Reports.

[11] M. M. Rahman, “Random Forest, Logistic Regression, Decision Tree on a global dataset: Comparison with SVM, KNN, MNB, and SVC addressing data latency, missing values, and specific pollutants (CO, O_3, NO_2),” 2024.Q

[12] A. T. Nguyen et al., “Predicting AQI using attention hybrid deep models,” Journal of Big Data, 2024.

[13] S. Bedi et al., “Utilizing LSTM models to predict PM2.5 levels,” ScienceDirect / journal, 2024.

[14] Z. Cui et al., “Air quality prediction and early warning based on Prophet-XGBoost,” SPIE / conference proceedings, 2024.

[15] M. W. Jalali *et al., Scalable AI-based Air Quality Forecasting and Classification for Public Health Purposes, *Discover Atmosphere*

[16] Z. Alfasanah et al., “AQI monitoring using EWMA and hybrid machine-learning models," ScienceDirect,2025

[17] M. Karmoude et al., “Machine-learning techniques for air-quality prediction and data analysis: A review,” ScienceDirect, 2025

[18] P. Patel, “A systematic study on PM2.5 and PM10 concentration modeling,” 2025.