International Journal of Engineering Applied Sciences and Technology, 2024
Vol. 9, Issue 01, ISSN No. 2455-2143, Pages 197-201
Published Online May 2024 in IJEAST (http://www.ijeast.com)

MACHINE LEARNING-BASED FORECASTING OF AIR QUALITY INDEX

Rashmitha
Dept. of Master of Computer Application
Vivekananda College of Engineering and Technology Puttur, India

Shamitha N Shetty
Dept. of Master of Computer Application
Vivekananda College of Engineering and Technology Puttur, India

Neema H
Dept. of Master of Computer Application
Vivekananda College of Engineering and Technology Puttur, India

Abstract—**Air pollution increases the risk of various illnesses that harm people. Therefore, it is important to establish fore casting methods for authorities. This research developed a solid ML models for predicting the Air Quality Index (AQI) using a Random Forest Regressor. By leveraging a comprehensive dataset from the Central Pollution Control Board (CPCB), this comprises information from major Indian cities like Bangalore, Mangalore, Delhi, Ahmedabad, Mumbai, and Chennai, the model successfully identified key pollutants, PM10 and PM2.5, as significant predictors of AQI. The findings highlight the model's reliability and potential applications in environmental monitoring and public health, providing valuable insights for timely interventions and policy-making aimed at improving air quality.**

Keywords: **Air Quality Prediction, Machine Learning, Random Forest Regressor, PM10, PM2.5, Environmental Monitoring.**

I. INTRODUCTION

Air pollution affects ecosystems and human life, making it a serious environmental hazard. Air pollution presents serious problems on a number of fronts. It endangers people's health by exposing them to dangerous chemicals and tiny particles potential of causing cancer, heart problems and other respiratory conditions. The result is environmental deterioration, emissions cause acid rain, which damages ecosystems and erodes infrastructure. Furthermore, air pollutants that trap heat, such as carbon dioxide and methane cause global warming and weather pattern disruptions by trapping heat in the atmosphere. Air pollution has a significant financial impact due to medical bills, lost production, and damage to buildings and crops. Pollution has a disproportionate impact on vulnerable com munities' health, underscoring disparities in exposure to and availability of clean air. Emission requirements are difficult for regulatory systems to successfully enforce, particularly in areas that are industrializing quickly.

For the intension of developing successful policies and putting into action prompt measures, accurate forecast of air quality is essential. Conventional monitoring techniques may have expensive prices and poor coverage. Conversely ML offers a strong substitute by utilizing extensive datasets to efficiently predict levels of pollution occurred. The primary goal of this is to build a ML models for predicting the air quality. It provides thorough and accurate forecasts by combining data from several sources. Regression and neural networks are two ML approaches that the model uses to improve precision and range of air quality forecasts. This method offers proactive environmental management techniques in addition to addressing current monitoring issues.

II. RELATED WORK

Air Pollution Forecasting has seen extensive research using various ML models. LSTM-based models effectively capture temporal dependencies, leading to improved forecasting ac curacy [1], [2].Regression models used to forecast PM2.5 levels, showcasing their capability in predicting levels of particulate particles [3]. Spatiotemporal models, such as regional spatiotemporal collaborative prediction, leverage spatial correlations to improve prediction accuracy

[4].Spatiotemporal models, such as regional spatiotemporal collaborative pre diction, leverage spatial correlations to improve prediction accuracy [7]. Comparative analyses provide a comprehensive comparisons of different ML techniques for predicting the AQI [3]. Novel implementations, like a genetic algorithm improved extreme learning machine, demonstrate innovative approaches to enhance forecasting performance [6]. Practical applications in specific regions, like Indian cities, offer insights into real world scenarios and the effectiveness of these models [5].

III. METHODOLOGY

A. Data Collection

Quality and completeness of the available data form the basis of every prediction model. Acquiring accurate and thorough data regarding air quality and meteorology guarantees the ability to learn from a range of scenarios and produce accurate forecasts.

To build a reliable air quality forecast model, large amount of data was collected from 2020 to 2024 from six major Indian cities. PM10, PM2.5, SO2, Ozone and NO2 were among the important pollutants included in the data, which came from the Central Pollution Control Board (CPCB). Furthermore, meteorological information was obtained, including temperature, humidity, wind direction (WD), and wind speed (WS), for Giving a complete image of the variables affecting air quality. This large dataset served as a reliable foundation for the analysis and air quality indicators predictions.

B. Data preprocessing

Preprocessing ensures that the models can train efficiently by converting unprocessed input into an organized and useable structure. Important actions to improve model performance include handling missing values, eliminating unnecessary data, and normalizing the features.

Strict preparation procedures were followed to guarantee data relevancy and quality. To preserve the integrity of the dataset, missing values were filled in using median values. High null value columns and those judged unnecessary for air quality indicators predictions were carefully eliminated. To enable efficient model training and comparison, all characteristics were then normalized using Min-Max scaling to standardize their ranges.

C. Feature Selection

By choosing the most essential characteristics, the model’s efficiency and accuracy are enhanced. The model is able to generate forecasts more accuratly by concentrating on the characteristics that significantly affect AQI.

To find predictors that strongly linked with AQI, feature selection was done. Considering characteristics with correlation coefficients greater than 0.4, a correlation matrix was employed. Important pollutants having substantial positive correlations with AQI, namely PM2.5, PM10, and NO2, were included. Furthermore, because of their substantial influence on air quality, climatic factors including air temperature were kept in place. For their possible impact on pollution dispersion and concentration, additional climatic parameters such as humidity, wind direction (WD), and wind speed (WS) were also taken into consideration.

D. Data Splitting

Analysing the model’s performance requires categorizing the data collected into training and testing sets. It guarantees that the model is tested using hypothetical data, giving an accurate assessment of its prediction power.

To precisely evaluate the performance of the model, the dataset was divided into training and testing sets. Data from 2024 to April has been reserved for testing, while data from 2021 to 2023 was reserved for model training. By evaluating the models on fresh, untested data, this temporal split improved the models’ robustness and generalizability.

E. Model Training

It is possible to compare and choose the best-performing algorithm by training numerous models. Every model has certain advantages, and comparing the models’ results enables one to select the best method for AQI prediction.

The prepared datasets were utilized to learn four different models: temporal relationships in sequential material led to its selection. Selected as a foundational paradigm to comprehend linear connections is linear regression. The feedforward neural network is a tool for identifying intricate nonlinear patterns in data. Random Forest is a collective learning approach that uses many decision trees to achieve high accuracy. To improve parameters and boost prediction performance, each model underwent extensive training based on the training dataset.

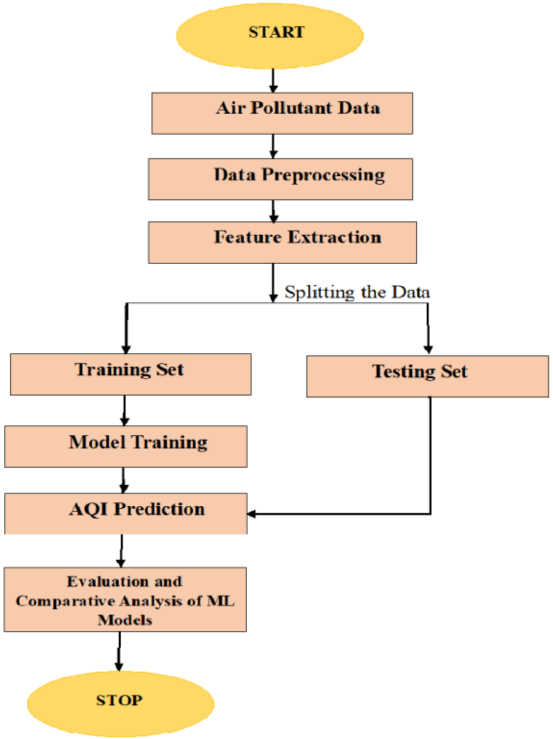

Fig. 1: Flowchart of the proposed model

F. AQI Prediction

Accurately predicting AQI is the model's ultimate aim. The test dataset is utilized to test the trained models' usability and dependability in real-world situations.

AQI values were predicted using each model on the assigned test dataset (2024 to April) after the models were trained. The models' accuracy in forecasting air quality was then examined by comparing these forecasts with the actual AQI readings.

G. Model Evaluation

Models are reviewed using relevant criteria to guarantee the reliability and accuracy of the selected model. A comparative study may be utilized to determine the advantages and disadvantages of each model.

The model's performance was assessed using metrics like MAE, R^2 value, and RMSE. It was made easier to understand which model was better for AQI prediction by comparing these measures. The outcome of this comprehensive examination were reliable in regard to the generated models' ability to anticipate air quality index (AQI) in a variety of urban and environmental circumstances. Figure 1 shows the methodological phases for the adopted approach.

IV. EXPLORATORY DATA ANALYSIS

The dynamic aspect of air quality is demonstrated with graphs that display the fluctuation of various pollutants over time, including PM2.5, PM10, NO2 and others. The aforementioned figure 2 demonstrate that all pollutants undergo annual variations, with levels rising and declining. The pollutant levels fluctuate shows the effect of a various variables, including industrial operations, vehicle emissions, seasonal variations, and governmental regulations. We may get important insights for efficient air quality management and policy creation by looking at these trends, which help us understand how various pollutants react to interventions and environmental circum stances.

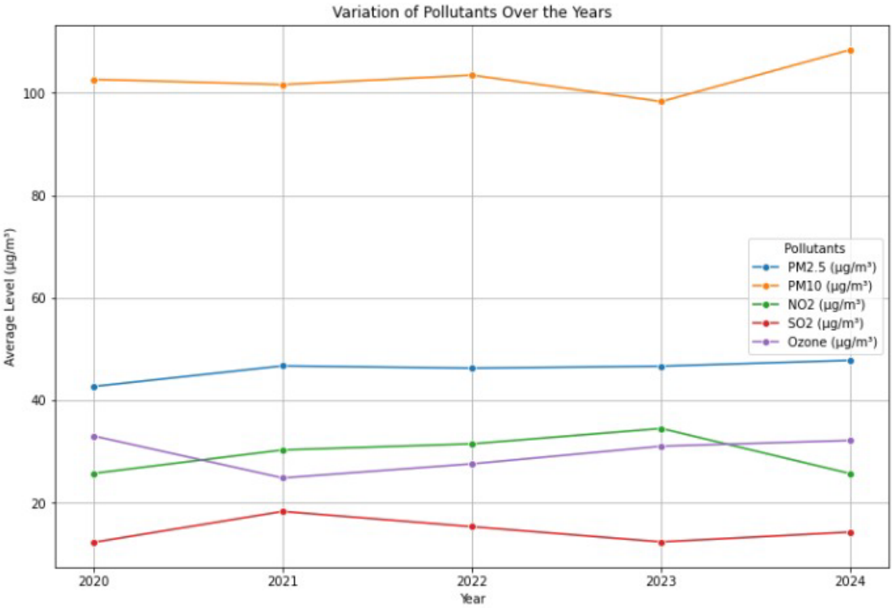

Fig. 2: Variation of Different Pollutants over the Year

V. RESULTS AND DISCUSSION

In this study, assessed many ML models, such as Random Forest, LSTM, Linear Regression and Neural Network. The evaluation of each model was conducted by utilizing historical data from six Indian cities to forecast AQI. R² scores and RMSE were used to gauge the outcomes.

Linear Regression exhibited good performance with Train RMSE 0.0295, Test RMSE 0.0214 and R² 0.9517, adept at capturing fundamental data trends despite potential complexity in more intricate patterns. Neural networks achieved Train RMSE 0.011, Test RMSE 0.012 and R² 0.98, indicating highly accurate predictions by effectively capturing complex nonlinear interactions within the dataset. LSTM networks achieved Train RMSE 0.045, Test RMSE 0.049 and R² 0.78, showing capability in sequential data prediction but with lower consistency, possibly due to complex temporal dependencies. Random Forest achieved a remarkable R² score of 0.9988, Train RMSE 0.023, Test RMSE 0.026, excelling in air quality forecasting by effectively managing feature interactions and delivering reliable predictions, outperforming other models.

After comparing various models, we may infer that Random Forest offers the most trustworthy and accurate air quality forecasts. This model's robustness to overfitting and ability to identify complex interactions within the data. Metrics of different models are shown in table I .

TABLE I: Performance Metrics of Various Models

Model	Train RMSE	Test RMSE	R²
Linear Regression	0.0295	0.0214	0.9517
Neural Network	0.011	0.012	0.98
LSTM	0.045	0.049	0.78
Random Forest	0.023	0.026	0.9988

VI. CONCLUSION

This effort on air quality forecasting using ML algorithms yielded important insights into the dynamics of air pollution. We produced accurate predictions over a broad range of contaminants because of thorough model training, validation, and assessment. The results show the necessity for sophisticated analytical techniques to understanding and predicting air quality, which is critical for successful public health and environmental management. Moving forward, incorporating our prediction models into decision-making frameworks can improve proactive steps for enhancing the air quality and protecting community well-being. This study shows a robust ML-based approach to air quality forecasting, demonstrating significant improvements over traditional methods. By integrating many data sources and employing advanced algorithms, the suggested technology provides accurate results. These predictions can help guide public health actions and environmental policies, contributing to improved air condition management and public well-being.for broader applicability.

REFERENCES

[1]. Chang, Y.-S., Chiao, H.-T., Abimannan, S., Huang, Y.-P., Tsai, Y.-T., and Lin, K.-M. (2020). An LSTM-based aggregated model for air pollution forecasting, Atmospheric Pollution Research, 11(8), 1451–1463.

[2]. Chen, H., Guan, M., and Li, H. (2021). Air quality prediction based on integrated dual LSTM model, IEEE Access, 9, 93285–93297.

[3]. Harishkumar, K.S., Yogesh, K.M., and Gad, I. (2020). Forecasting air pollution particulate matter (PM2.5) using machine learning regression models, Procedia Computer Science, 171, 2057–2066.

[4]. Zhao, G., Huang, G., He, H., He, H., and Ren, J. (2019). Regional spatiotemporal collaborative prediction model for air quality, IEEE Access, 7, 134903–134919.

[5]. Gupta, N.S., Mohta, Y., Heda, K., Armaan, R., Valarmathi, B., and Arulkumaran, G. (2023). Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis, Journal of Environmen tal and Public Health, Article ID 2023.

[6]. Liu, C., Pan, G., Song, D., and Wei, H. (2023). Air Quality Index Forecasting Via Genetic Algorithm-Based Improved Extreme Learning Machine, IEEE Access, Article ID 2023.

[7]. Kumar, K. and Pande, B.P. (2023). Air pollution prediction with ma chine learning: a case study of Indian cities, International Journal of Environmental Science and Technology, 20(5), 5333–5348.

[8]. Bhalgat, P., Pitale, S., and Bhoite, S. (2019). Air quality prediction using machine learning algorithms, International Journal of Computer Applications Technology and Research, 8(9), 367–370.

[9]. Castelli, M., Clemente, F.M., Popovic, A., Silva, S., and Vanneschi, L. ´ (2020). A machine learning approach to predict air quality in California, Complexity, Article ID 2020.

[10]. Sanjeev, D. (2021). Implementation of machine learning algorithms for analysis and prediction of air quality, International Journal of Engineering Research & Technology (IJERT), 10(3), 533–538.

[11]. Soundari, A.G., Jeslin, J.G., and Akshaya, A.C. (2019). Indian air quality prediction and analysis using machine learning, International Journal of Applied Engineering Research, 14(11), 181–186.

[12]. Mendez, M., Merayo, M.G., and N ´ u´nez, M. (2023). Machine learning ˜ algorithms to forecast air quality: a survey, Artificial Intelligence Review, 1–36.