Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Air Quality Modeling and Prediction: An Investigation of Machine Learning Approaches and Exploratory Data Analysis

Anil Kumar Bisht[1], S.S. Bedi[1], Ashwani[1], Gupta, Iram Naim[1], Atul Sarojwal[2]

[1]Department of CS&IT, FET, MJPRU, Bareilly, India
[2]Department of EE, FET, MJPRU, Bareilly, India
Email: anil.bisht@mjpru.ac.in

Environmental science is one of the scientific disciplines that has recently changed as a result of machine learning. In the modern era, where human health and the environment are the most important components for healthy living, both are greatly impacted by air quality. As pollution levels rise due to urbanization, industry, and other human-caused factors, forecasting and monitoring air quality is essential to minimizing negative effects. Traditional air quality monitoring approaches have limited geographic and temporal coverage because they frequently rely on manual data collection. Machine learning, a subfield of artificial intelligence, is primarily responsible for developing algorithms that are able to forecast and learn from data. The literature review of machine learning techniques used for air quality modeling and prediction has been the main focus of this article. Various standard methods are reviewed along with some enhanced techniques which makes the prediction more efficient. Different research works have been analyzed in terms of various algorithms and technologies. A thorough Exploratory Data Analysis (EDA) has also been conducted to get insights into the distribution patterns of important variables and how they interacted with one another throughout the dataset.

Keywords: Air Quality Prediction, Minimum Support Price, Decision Tree Algorithm, Clustering, Decision Tree Classifier, Random Forest Classifier, KNN (K nearest Neighbor), SVM (Support Vector Machine).

1. Introduction

High amounts of air pollution can have a variety of negative health effects. Increased health

risks like respiratory infections, heart disease, stroke, and lung cancer can have a negative effect on people who are already ill, such as children, the elderly, and those living in poverty. Furthermore, poor air quality increases the risk of neurological conditions like dementia and cognitive impairment, as well as miscarriage and stillbirth. In 2019, air pollution claimed the lives of 6.7 million people. Noncommunicable diseases (NCDs), such as diabetes, ischemic heart disease, stroke, lung cancer, asthma, and chronic obstructive pulmonary disease (COPD), account for over 85% of these cases. As a result, air pollution is currently the second leading cause of NCDs globally, after tobacco. [1].

Cities are growing economically and technologically, which is leading to problems with environmental pollution [2]. Currently, deforestation produces significant climate change, which in turn leads to air pollution [3]. Air pollution may have a major effect on public health [4]. Predicting air quality is a difficult experiment because to the dynamic nature, unpredictability, and significant geographical and temporal variability of pollutants and particles. Simultaneously, there is a need to model, forecast, and monitor air quality so that appropriate actions can be made to avoid its negative effects on society and the environment. Environmental science has seen a revolution in recent years due to technology breakthroughs, particularly in the area of artificial intelligence. Machine learning has emerged as a potent tool to improve the efficacy and efficiency of air quality modeling. [4,5,6].

Deep learning is one type of machine learning technology that has recently garnered a lot of interest from the research community [6]. The purpose of this work is to review the literature on machine learning methods for air quality modeling and forecasting. In addition to some enhanced techniques based on the use of machine learning and deep learning that improve forecast accuracy, several approaches are investigated. To learn more about the distribution patterns, a comprehensive Exploratory Data Analysis (EDA) has also been carried out.

2. BACKGROUND AND MOTIVATION

Air pollution is a serious problem in India, where the quality of the air is frequently bad in many places. Any physical, chemical, or biological alteration to the air that contaminates it with dangerous gases, dust, or smoke is referred to as air pollution. Humans, animals, and plants are all negatively impacted by this contamination. The relationship between air pollution and health issues has been the subject of numerous studies. A thorough evaluation of the health risks associated with air pollution was conducted in a study [7]. Nearly all people on the earth (99%) breathe air that is highly polluted and exceeds WHO guideline limits, according to WHO data; low- and middle-income nations are especially vulnerable [8]. Two types of air pollution can be distinguished:

1. Primary Pollutants: These include particulate matter (PM), nitrogen oxides (NOx), sulfur dioxide (SO2), and carbon monoxide (CO), all of which directly contribute to air pollution.

2. Secondary Pollutants: These result from the mixing and reaction of the primary pollutants, and include haze. Fossil fuel combustion, automobile emissions, agricultural activities, factories, and other industries are the main causes of air pollution. The effects of air pollution on the environment and human health are severe. It has been linked to a number of heart problems, lung cancer, and respiratory ailments.

Numerous monitoring sites located throughout India provide data on air quality. The availability of the data can aid in the development and enhancement of our prediction models. At the moment, the government takes action when the air quality reaches hazardous levels. The air quality index (AQI) displays the concentrations of the criteria pollutants in the air. The overall AQI is the highest recorded AQI for each of the individual criteria pollutants. AQI tests also highlight the health risks associated with exposure to air quality. If there is a way to anticipate when the air quality would reach unsafe levels, the government may be able to prevent further degradation of the air quality by implementing limitations like these early on. The proactive approach of creating a forecasting system is definitely helpful in predicting pollution levels, allowing people to be informed about air pollution in advance.

3. LITERATURE SURVEY

Various researchers have focused on leveraging ML techniques to forecast the air quality.

Deters et al. [3] suggested a machine learning method to forecast PM2.5 concentrations from wind speed and wind direction, based on six years of weather and pollution data analysis. The authors came to the conclusion that using machine learning-based statistical models to forecast PM2.5 concentrations using meteorological data is pertinent.

Zhang and Ding [5] tried to determine the amounts of air pollutants in Hong Kong using an Extreme Learning Machine (ELM) algorithm based on information gathered from two monitoring stations' eight air quality indicators. The Extreme Learning Machine (ELM)-based model the scientists suggested performs better than conventional statistical techniques, they concluded.

Zhu et al. [9] created enhanced models that predict, using meteorological data, the hourly concentration of air pollution. According to the authors, advanced optimization techniques speed up huge data training and greatly improve model convergence.

Bhalgat et al. [10] outlined some ML-based projects that used a dataset of 60383 records with 13 variables from Kaggle to forecast the amount of SO2 in the Maharashtra, India, environment. Since their model was unable to produce the desired results, the authors came to the conclusion that it was inadequate.

Castelli et al. [2] used support vector regression (SVR), a machine learning technique, to forecast California's air quality. According to the authors, hourly pollutant concentrations could be accurately predicted by their suggested model, which was based on the radial basis function (RBF).

Arnaudo et al. [11] conducted a thorough investigation into the estimation of air quality in the Milan metropolitan region, putting forward various machine learning techniques that integrate traffic and meteorological data. The authors discovered an intriguing fact: even with more basic linear models, an accurate estimate of the Air Quality Index, or AQI, may be obtained.

Akiladevi R et al. [12] developed a prediction model for air quality forecasting based on machine learning. They have employed a number of machine learning methods, including Decision Tree, K Nearest Neighbor, Random Forest, Support Vector Machine, Naive Bayes Classification, and Logistic Regression. According to the authors, the outcomes of their

suggested decision tree-based model were superior.

Dobrea at al. [13] used machine learning techniques to predict air pollution levels. On the time-series datasets for PM10 and PM2.5 particles, they have contrasted models based on Support Vector Regression (SVR), Autoregressive Integrated Moving Average (ARIMA), and Long Short-Term Memory (LSTM) techniques. Based on the correlation values of 0.966 and 0.921 for PM10, respectively, it was determined that SVR and ARIMA are the most effective algorithms for forecasting air pollutant concentrations.

Liang et al. [14] used an 11-year dataset gathered by Taiwan's Environmental Protection Administration to construct a variety of machine learning (ML) based air quality prediction models. With high R^2 values, AdaBoost and the stacking ensemble perform best for target predictions.

Bhan and Niranjana Murthy [15] examined India's air pollution situation and the use of machine learning techniques for air quality prediction estimation. The authors have come to the conclusion that machine learning approaches to air quality pollution predictions produce promising outcomes.

Badrakh et al. [16] developed a machine learning model to forecast Ulaanbaatar, Mongolia's air quality. Temperature, humidity, wind direction, air pressure, PM2.5 and PM10, NO2, CO, and SO2 were among the features they used. They used the long short-term memory (LSTM) model to forecast the parameters of air pollution. The results showed that as training time increases, so do the standard error and the discrepancy between test and measurement values.

Samayan Bhattacharya and S.K. Shahnawaz [17] claimed to have created a model based on Support Vector Regression (SVR) to predict the levels of several pollutants in New Delhi using the CPCB-provided dataset. Out of all the methods examined, a Radial Basis Function (RBF) kernel produced the best results with SVR. The model predicts the Air Quality Index (AQI) with an accuracy rate of 93.4%.

Kumar and Pande [18] carried out a thorough examination of 23 Indian cities' air pollution data in order to forecast the quality of the air. Exploratory data analysis was carried out in order to comprehend the various hidden patterns in the dataset as well as the substances that have a direct impact on the air quality index. The Gaussian Naive Bayes model achieved the highest accuracy, while the Support Vector Machine model achieved the lowest.

Gladkova and Saychenko [19] conducted a comparative analysis of air quality prediction using machine learning techniques. Numerous machine learning techniques are employed, including Facebook Prophet, Autoregressive Integrated Moving Average (ARIMA), and long short-term memory (LSTM). When considering the MSE and RMSE measures, the predictive models ARIMA and Prophet both generated forecasts with similar accuracy; nevertheless, LSTM performs better than both models in this area.

Deepu et al. [20] used the K-Nearest Neighbor (KNN) machine learning approach to forecast the AQI. The authors' prediction of air pollution was 99.1071% accurate.

Jain et al. [21] We out a thorough analysis of 900 peer-reviewed publications published between 1990 and 2022 that focused on the issue of air pollution and machine learning-based modeling for it. The Web of Science database was used as a source. Using the VOS Viewer

and biblioshiny tools, they have assessed these publications by finding and visualizing key trends, nations, research papers, and journals that are concerned with these subjects. The authors highlighted the relevant subjects for more research and concluded with the most recent contributions.

Thanongsak Xayasauk and Hwamin Lee [7] suggested a deep learning method for predicting South Korea's air pollution. The stacked autoencoder model served as the basis for the experiments. The outcomes are satisfactory.

Bekkar et al. [22] created a CNN-LSTM-based hybrid deep learning system to predict Beijing, China's hourly PM2.5 concentration. The authors' solution-based model beats all of the stated traditional models in terms of predictive performance, according to experimental results.

Sadhana et al. [24] developed a deep learning model to predict Beijing's PM2.5 pollutant concentration. They have merged LSTM and CNN. With an accuracy rating of 90.03 percent, their model outperformed many of the earlier models that were created.

4. METHODOLOGY USED IN REVIEWED PAPERS

Various authors adopted almost the same research methodology. Fig. 1 represents methodology adopted by papers reviewed.

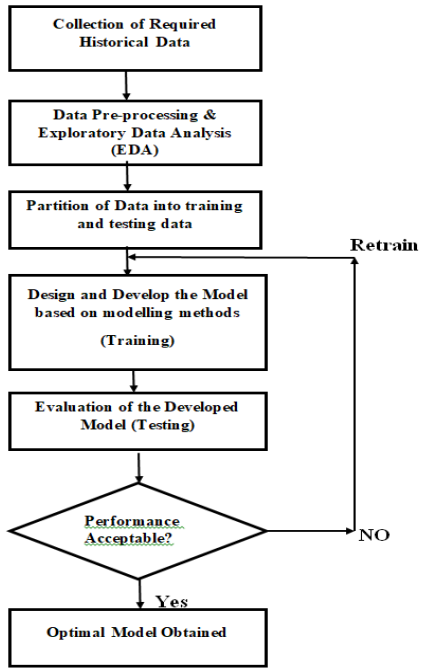

Figure 1: Basic workflow of machine learning model

5. Data Collection and Preprocessing: Exploratory Data analysis (EDA)

5.1 Dataset

The CPCB website provided a dataset for Lalbagh, Lucknow district, Uttar Pradesh, which was used to conduct the study. The dataset began on January 1, 2018, and ended on December 31, 2023, spanning six years. It contained data on the AQI value, PM2.5, NO2, SO2, and ozone.

5.2 EXPLORATORY DATA ANALYSIS (EDA)

An extensive Exploratory Data Analysis (EDA) was the first step in creating the suggested model. Complex insights on the distribution patterns of significant variables and their interactions with one another across the dataset were revealed to us. Strong visuals are used to provide a profound understanding, and the EDA insights served as a lighthouse that influenced further stages of the model development process. Fig. 2 presents the overall statistics summary of collected dataset.

	PM2.5	NO2	SO2	OZONE	AQI
count	2191.000000	2191.000000	2191.000000	2191.000000	2191.000000
mean	181.636011	48.491100	10.450023	32.776814	190.183021
std	111.829639	29.544731	7.925056	21.348552	104.710101
min	9.000000	2.000000	1.000000	1.000000	31.000000
25%	76.000000	26.000000	6.000000	19.000000	93.000000
50%	175.500000	44.000000	9.000000	29.000000	180.000000
75%	273.000000	65.000000	12.000000	42.000000	274.500000
max	459.000000	268.000000	69.000000	207.000000	475.000000

Figure 2: Basic Statistics Summary:

5.2.1 Correlation Matrix:

Fig. 3 represents the correlation matrix. The degree and direction of correlations between variable pairs in the Data Frame are displayed in the correlation matrix that is displayed below. The correlation coefficient between two variables, which ranges from -1 to 1, is represented by each cell in the matrix. A strong positive correlation, where one variable tends to increase as the other does, is indicated by a value near 1. A high negative correlation is shown by a value near -1, which means that when one variable rises, the other tends to fall. There appears to be no linear relationship between the variables when the values are close to 0.

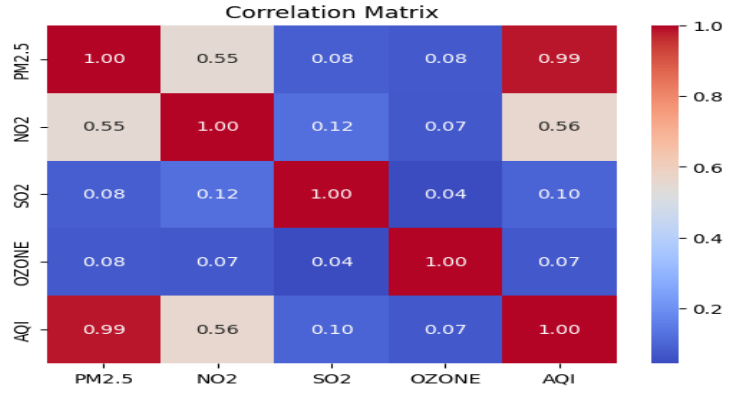

Figure 3: Correlation Matrix

5.2.2. Boxplot for Outliers:

Fig. 4 displays the dataset's boxplot. It displays the numerical data distribution for each of the following parameters: PM2.5, AQI, NO2, SO2, and OZONE. It can also be used to identify the outlier. Outliers are data points that significantly depart from the overall trend. They can significantly affect the results of statistical analyses and make it challenging to draw trustworthy conclusions.

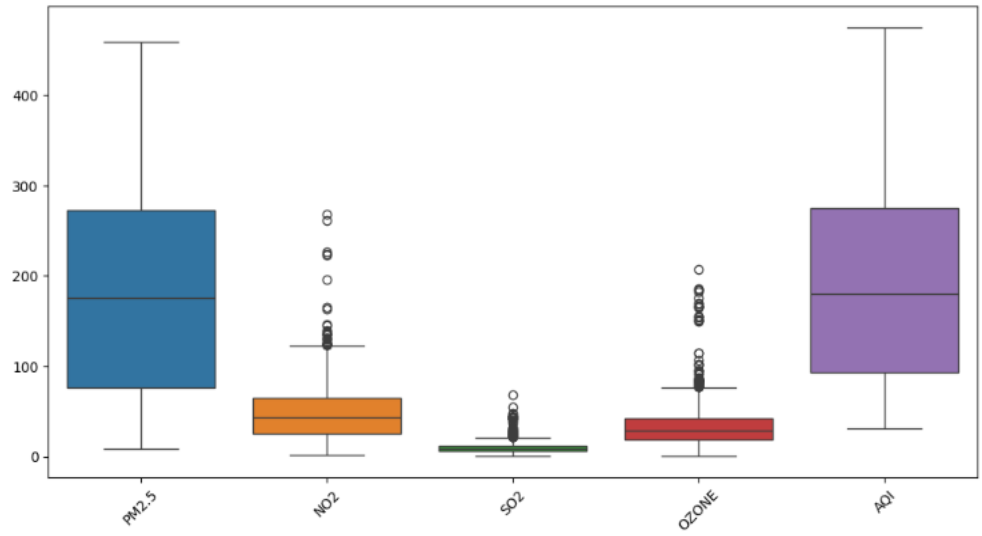

Figure 4: Boxplot Visualization for Outliers

5.2.3. Jointplot: The jointplot is shown in fig. 5. It displays the dataset's correlation between PM2.5 and AQI.

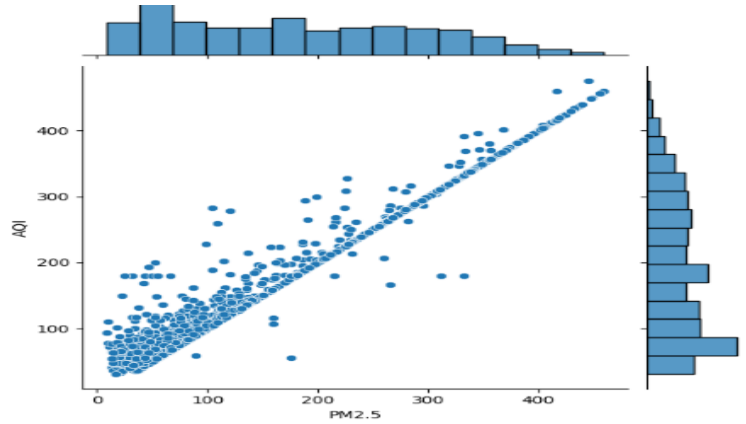

Figure 5: Joint plot Visualization

It offers a thorough perspective by combining histograms with a scatterplot. The scatterplot illustrates the relationship between PM2.5 levels and AQI scores by displaying individual data

points. The distribution of PM2.5 and AQI is shown by the histograms on the upper and lower margins, respectively. By highlighting trends, correlations, and possible outliers between the two variables, this figure provides a better understanding of how they interact.

6. CONCLUSION

Reviewing the literature on machine learning methods for air quality modeling and prediction was the main focus of this article. Several traditional methods are reviewed, as well as several enhanced methods that improve the prediction's performance. This article compares a number of research papers in detail using various algorithms and technologies. Additionally, a thorough Exploratory Data Analysis (EDA) has been conducted to gain a better understanding of the distribution patterns of important variables and their interactions throughout the dataset. PM2.5 and AQI have a substantial positive association, according to the correlation matrix. The boxplot for outliers lets us comprehend the data points known as outliers, which significantly diverge from the overall trend, and allows us to compare the distribution of the complete data set with regard to each parameter. Individual data points are displayed in a joint plot, which provides a greater understanding of the relationship between PM2.5 levels and AQI values. Following a thorough analysis of the available literature and the experimental dataset, the next phase of study will involve creating prediction models based on machine learning methods.

ACKNOWLEDGEMENT: I (Dr. Anil Kumar Bisht), Assistant Professor, CSIT, MJP Rohilkhand University Bareilly" am thankful to MJP Rohilkhand University Bareilly for the financial support in the form of "Innovative Research Grant (IRG)" with File No. "IRG/MJPRU/DOR/2022/04".

References

1. https://www.who.int/news/item/25-06-2024-what-are-health-consequences-of-air-pollution-on-populations
2. Castelli, M., Clemente, F. M., Popovic, A., Silva, S., & Vanneschi, L., "Machine learning approach to predict air quality in California", Hindwai, 2020
3. Jan Kleine Deters, Rasa Zalakeviciute, Mario Gonzalez, and Yves Rybarczyk, "Modeling PM2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters", Journal of Electrical and Computer Engineering, Hindawi, Volume 2017
4. Luca Cagliero, Tania Cerquitelli, Silvia Chiusano, Paolo Garza, Giuseppe Ricupero, Xin Xiao, "Modeling correlations among air pollution-related data through generalized association rules", 2016.
5. Jiangshe Zhang, and Weifu Ding, "Prediction of Air Pollutants Concentration Based on an Extreme Learning Machine: The Case of HongKong", Int. Journal Environ. Res. Public Health, MDPI, 2017.
6. THANONGSAK XAYASOUK & HWAMIN LEE, "AIR POLLUTION PREDICTION SYSTEM USING DEEP LEARNING", WIT Transactions on Ecology and the Environment, Vol 230, © 2018 WIT Press.
7. Izabela Sówka, Dominik Kobus, Krzysztof Skotak, Maciej Zathey, Beata Merenda, Małgorzata Paciorek, "Assessment of the Health Risk Related to Air Pollution in Selected Polish Health Resorts", Journal of Ecological Engineering Vol. 20(10), pages 132–145, 2019

8. https://www.who.int/health-topics/air-pollution#tab=tab_1
9. Dixian Zhu, Changjie Cai , Tianbao Yang and Xun Zhou, "A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization", Big Data Cognitive Computing, MDPI, 2018
10. Bhalgat, P., Pitale, S., & Bhoite, S., "Air quality prediction using machine learning algorithms", International Journal of Computer Applications Technology and Research Volume 8–Issue 09, 367-370, 2019
11. Arnaudo, E., Farasin, A., & Rossi, C. , "A comparative analysis for air quality estimation from traffic and meteorological data", Applied Sciences, doi:10.3390/app10134587, 2020
12. R, Akiladevi., B, N. D., V, N. K., & P, N., "Prediction and analysis of pollutants using supervised machine learning", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878 (Online), Volume-9 Issue-2, July 2020
13. Marius Dobrea et al., " Machine learning algorithms for air pollutants forecasting", IEEE 26th International Symposium for Design and Technology in Electronic Packaging (SIITME), 2020
14. Yun-Chia Liang et al., "Machine Learning-Based Prediction of Air Quality", Applied Sciences, MDPI, 2020
15. Madhu Bhan, Niranjanamurthy, "Control Of Air Pollution Using Machine Learning", Natural. Volatiles & Essential Oils, 8(4): 11773-11785, 2021
16. Otgonsuvd Badrakh, Lodoiravsal Choimaa, "Air quality predictions of Ulaanbaatar using machine learning approach", International Symposium on Grids & Clouds, ISGC2021 22-26 March 2021 Academia Sinica, Taipei, Taiwan (online), 2021
17. Samayan Bhattacharya, Sk Shahnawaz, "Using Machine Learning to Predict Air Quality Index in New Delhi", Cornell University, https://arxiv.org/abs/2112.05753, https://doi.org/10.48550/arXiv.2112.05753, https://arxiv.org/pdf/2112.05753, 10 Dec 2021
18. K. Kumar, B. P. Pande, "Air pollution prediction with machine learning: a case study of Indian cities", International Journal of Environmental Science and Technology, https://doi.org/10.1007/s13762-022-04241-5, 2022
19. Ekaterina Gladkovaa, Liliya Saychenko, "Applying machine learning techniques in air quality prediction", Transportation Research Procedia 63, Elsevier, 2022
20. Deepu B P, Dr. Ravindra P Rajput, "Air Pollution Prediction using Machine Learning", International Research Journal of Engineering and Technology (IRJET), Volume: 09 Issue: 07 | July 2022
21. Shikha Jain et al., "Use of Machine Learning in Air Pollution Research: A Bibliographic Perspective", Electronics, MDPI, 11, 3621. https://doi.org/10.3390/electronics11213621, 2022
22. Abdellatif Bekkar et al., "Air pollution prediction in smart city, deep learning approach", Journal of Big Data, https://doi.org/10.1186/s40537 021 00548 1, 2021
23. Konduri Sai Sadhana et al., "Air Pollution Prediction Using Deep Learning", IEEE, 2nd Mysore Sub Section International Conference (MysuruCon), DOI: 10.1109/MysuruCon55714.2022.9972528, 2022