Engineering Innovations
ISSN: 2813-1002, Vol. 14, pp 99-115
doi:10.4028/p-wcnTd6
© 2025 The Author(s). Published by Trans Tech Publications Ltd, Switzerland.

Submitted: 2024-08-27
Revised: 2024-12-02
Accepted: 2024-12-04
Online: 2025-04-04

Implementation of Boosting Algorithms in Predicting Air Quality Index of South Indian Cities

Deepasikha Mishra[1,a], Ansuman Sahu[1,b], Sreajan Naman[1,c]

[1]School of Computer Science and Engineering, VIT-AP University, Andhra Pradesh, India

[a]deepasikharesearch@gmail.com, [b]ansumansahu264@gmail.com, [c]namansreajan@gmail.com

Keywords: Air Pollution, Air Quality Index (AQI), XGBoost, LightGBM, Ensemble Stack.

Abstract. This paper addresses the critical issue of atmospheric pollution in India, underscoring the necessity for precise predictive analytics of Air Quality Index (AQI) data for effective pollution control. The study delineates the etiological factors and substantial health hazards correlated with air pollution, encompassing elevated mortality rates, respiratory and cardiovascular diseases, and mental health complications. The AQI is presented as a necessary component for converting complex air quality data into a single, easily understandable metric. The principal aim of this research is to facilitate effective pollution control through real-time AQI monitoring and precise future predictions for timely interventions. To attain this objective, the research employs the use of boosting algorithms, like extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM), and an ensemble stack of XGBoost and LightGBM for AQI prediction of four South Indian cities - Amaravati, Bengaluru, Chennai, and Hyderabad. The results presented in this paper are based on daily data of the aforementioned cities, collected from the Central Pollution Control Board (CPCB) website, Government of India, covering the period from May 31, 2019, to November 22, 2023. The performance of the deep learning models on the data was found noteworthy, with consistently high R^2 scores and low root mean squared error (RMSE), exhibiting their efficacy in providing accurate results. By merging technological innovation with machine learning capabilities, the research aims to equip decision-makers with actionable insights for informed pollution mitigation strategies, promoting a more sustainable environment.

1. Introduction

Atmospheric air constitutes a vital ecological resource essential for the survival of biotic entities on Earth. Humans, animals, and plants depend on atmospheric quality, rendering the preservation of pristine air paramount for life sustenance. According to the Blacksmith Institute's 2008 report, the globe's most severely polluted locales are marred by deteriorating urban air quality and pervasive indoor pollution [1]. Outdoor air pollution leads to around 6.7 million deaths globally each year, while indoor air pollution also results in significant number of premature deaths annually in certain regions. [2,3]. Various sources, including the use of unclean fuels for household purposes in developing countries, emissions from furniture and construction materials, and emissions from microorganisms cause indoor air pollution. Poor ventilation and additional indoor sources make indoor air pollution more complex and often more concentrated than outdoor air pollution [3,4]. Addressing the detrimental impacts of air pollution necessitates a dual focus on urban air quality and indoor pollution abatement. India, as a rapidly developing nation, grapples with escalating air pollution due to swift urbanization and industrial proliferation. The burgeoning population, vehicular emissions, and industrial effluents are accelerating atmospheric degradation. Exposure to polluted air precipitates both acute and chronic health repercussions. Empirical studies have substantiated that air pollutants, such as fine particulate matter (PM2.5) and other noxious substances, elevate mortality rates and induce respiratory, cardiovascular, and psychological disorders. Acute exposure to polluted air correlates with heightened mortality from all causes, including cardiorespiratory conditions, especially among the elderly, while young children exposed to PM2.5 exhibit increased anxiety symptoms [7,8]. Data analytics and machine learning paradigms have emerged as robust methodologies for enhancing air quality prognostications, offering superior accuracy in pollution assessment and forecasting. Researchers have extensively utilized data-driven and machine learning-based models for evaluating the quality of air. The AQI serves as an effective communicative instrument, translating complex air quality data into accessible terms for public comprehension and action. The combination of AQI monitoring, accurate predictions through machine learning, and a deep analysis of air quality patterns positions this study as a significant step toward addressing the air pollution crisis in India. The remaining segments of the paper is organized as follows: the literature review related to the work is covered in Section 2, the research methodology is outlined in Section 3, the results are presented in Section 4, and the conclusions are provided in Section 5.

2. Literature Review

The unprecedented air pollution rise in India, spurred by rapid urbanization and industrialization, mandates an exhaustive examination of contemporary air quality assessment and management research. This literature review aims to synthesize the research findings delving into the detrimental effects of air pollution in Indian cities, highlighting the significant threats to public health and the environment. Integrating machine learning (ML) algorithms emerges as a promising strategy for predicting and mitigating air pollution, specifically focusing on enhancing our understanding of the intricate dynamics of pollutants.

2.1 Health Hazards

Air pollution is emerging as a significant concern which is impacting public health. Studies have shown a correlation between exposure to atmospheric pollutants and the incidence of acute lower respiratory infections, chronic obstructive pulmonary disease, asthma, and pulmonary carcinogenesis [5,8]. PM10 and PM2.5 have been identified as major contributors to deleterious respiratory health outcomes, even at minimal pollutant concentrations [31,32]. In India, the situation is particularly alarming, where high levels of particulate matter pose a grave concern, necessitating policy actions to reduce pollutants and attain immediate health benefits. Controlling outdoor air pollution in India can lead to significant health benefits, including reductions in morbidity and mortality. However, the distribution and extent of these benefits vary widely across geography, time, and populations. Climate change-driven temperature increases exacerbate air quality, but climate solutions, such as clean and renewable energy and cool roofs, can improve air quality and health. Research indicates that higher air quality is intricately linked to an improved quality of life, which offers potential avenues for mitigating the health risks associated with air pollution [33]. Children are especially susceptible to the harmful effects of air pollution on their respiratory health, emphasizing the need to enhance air quality. Studies have shown that prolonged exposure to air pollution can adversely affect lung function development in children, reinforcing the importance of improving air quality [34,35]. Additionally, reducing PM2.5 levels has been associated with decreased medical costs and less loss of working and living time for patients, demonstrating the health benefits of air quality improvement [36]. Thus, it is essential to have efficient air quality management planning in place to mitigate the negative consequences of air pollution on human well-being.

2.2 Monitoring and Analysis of AQI

Monitoring and analyzing the AQI for comprehending pollution trends is of significant importance. A generative time-series model based on a recurrent extension of the variational autoencoder (VAE) has been proposed by researchers to forecast major pollution indicators with high efficiency [42]. The AQI is a metric widely used for recording and understanding pollution trends. Various machine learning algorithms have been proposed for predicting air contamination and analyzing the AQI [9,37,38,39]. Sensor-based networks have been developed to monitor air quality parameters and predict a locality's sustainability. In addition, dispersion models and control systems help understand the distribution and dynamics of pollutants and control air pollutant concentrations. Incorporating satellite data alongside statistical and deep learning techniques for AQI forecasting during COVID-19 lockdowns have substantially reduced air pollutant levels [43]. These studies have demonstrated the efficacy in AQI forecasting by implementing machine learning, achieving high accuracy and outperforming other models. Researchers have investigated the air quality in Bangladesh during lockdown periods, revealing a strong correlation between COVID-19 spread and air pollution reduction [44]. Similarly, the impact of the lockdown on air quality in Henan, China, has been studied, and a decrease in pollutants during the lockdown period has been observed [45]. Furthermore, an analysis of the spatial and temporal variations in air quality measures during lockdown over the Indo-Gangetic Plain has shown a decline in pollutants and improved regional air quality [46]. Analyzing the trend of air pollutant concentrations can be challenging, mainly when dealing with data from monitoring networks of varying lengths. However, advanced monitoring systems incorporating satellite data and machine learning approaches offer a nuanced understanding of pollution trends. This understanding paves the way for designing effective abatement strategies to control air pollution levels.

2.3 Climate Change

Climate change and air pollution have complex and intertwined interactions, with significant environmental and public health consequences. Studies have shown that in the second half of this century, temperature changes and emissions of biogenic volatile organic compounds (VOCs) are projected to increase ozone concentrations, and addressing climate change can directly and indirectly reduce air pollution [47]. However, distinct policy levers are required to tackle both issues. Future climate scenarios show improved pollutant concentrations due to reduced emissions despite higher temperatures and lower precipitation. Climatic variables such as temperature, rainfall, wind, and humidity play a pivotal role in air pollution by affecting pollutant strength, transportation, and dispersion. Researchers have emphasized the need to integrate climate change considerations into future air quality predictions and policy interventions, and they have highlighted the need for improved tools to assess the combined implications of addressing air quality and climate change. Current Integrated Assessment Models (IAMs), which are commonly utilized in policy development, often employ global or regional marginal response factors to assess the air quality impacts of climate scenarios [48]. Still, this approach may lead to inaccurate conclusions. To address this gap, researchers have developed computationally efficient methods to measure the air quality impacts. The convergence of machine learning with air quality research highlights the potential of advanced technologies in facilitating more precise predictions and epidemiological analyses, shaping the trajectory of future research and interventions [9,37,38]. Extreme weather events, which are becoming increasingly common due to climate change, contribute to air pollution through events such as wildfires, releasing large amounts of pollutants into the atmosphere [49,52]. This relationship is not unidirectional, as certain pollutants, such as black carbon, can exacerbate climate change by absorbing sunlight and contributing to atmospheric warming. Recognizing these connections is essential for developing comprehensive plans addressing climate change and air pollution, ensuring effective environmental and public health management.

2.4 Usage of Predictive Analytics

Machine learning models have been extensively studied and utilized to improve air pollution research and prediction. They offer several advantages over conventional methods such as increased accuracy, simplified mathematical and statistical approaches, and improved analysis of air pollution data. Many machine learning techniques, such as Random Forest, SVM, Decision Trees, LS-SVM, and ANN, have been used in air pollution epidemiology [37,38]. These models have been employed to evaluate air quality for different regions and analyze air pollution data. XGBoost and LightGBM are two prominent machine learning algorithms that have been integrated with web modules to provide real-time air quality forecasts [9]. Regression models have been utilized to analyze air pollution data, showing noteworthy performances [37,38]. XGBoost has been shown to outperform deep learning techniques in estimating the AQI [39]. Random forest regression has demonstrated high accuracy and low RMSE in predicting pollution levels [40]. Ensemble voting models, such as the AQP-EDLMRA technique, which combines deep learning classification methods, have shown improved forecasting performance for air pollutant prediction [41].

3. Proposed Methodology

The Air Quality Index (AQI) prediction for the four South Indian cities was performed using pre-processed datasets, which included seven air pollutant attributes: PM10, PM2.5, NO2, NH3, SO2, CO, and O3, along with daily AQI values and their respective AQI standards. Given that air quality datasets often contain missing or incomplete data, regression models are particularly useful because they are robust to such imperfections. These models can effectively manage missing data, providing reliable predictions even when some features are absent. Their flexibility and interpretability make them suitable for accurate and dependable AQI prediction. In this study, the regressor models used were XGBoost, LightGBM, and an Ensemble Stack combining both XGBoost and LightGBM to predict the AQI and evaluate the performance of these boosting algorithms. The overall process of the methodology is depicted in Fig. 1.

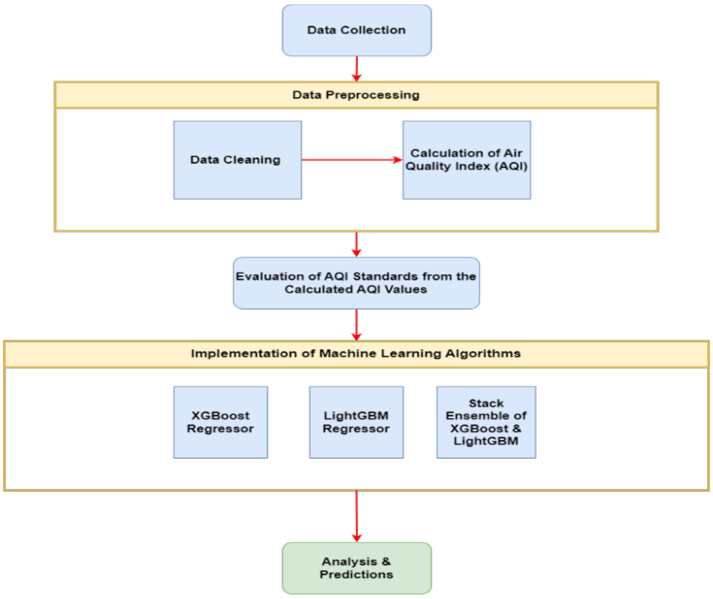
Data Collection
Data Preprocessing
Data Cleaning
Calculation of Air Quality Index (AQI)
Evaluation of AQI Standards from the Calculated AQI Values
Implementation of Machine Learning Algorithms
XGBoost Regressor
LightGBM Regressor
Stack Ensemble of XGBoost & LightGBM
Analysis & Predictions

Fig. 1. Architectural Representation of the Research Work.

3.1 Data Collection

The research pertains to the study of air pollution and involves the use of data that has been downloaded from the CPCB website, under the Department of Environment, Forest, and Climate Change, Government of India [55]. The dataset comprises the air pollutant parameters of four South Indian cities, namely Amaravati, Bengaluru, Chennai, and Hyderabad. It contains seven distinct attributes of Air Pollutants as specified in the Indian System of AQI value calculation. The dataset is thus made ready to offer a comprehensive and detailed analysis of the air quality of these four cities, based on the aforementioned pollutant parameters.

3.2 Data Cleaning

During the data cleansing phase, extensive measures were implemented to address missing values and outliers within the dataset. The use of Excel formulas within the spreadsheet, specifically the IF and ISBLANK functions, allowed for a systematic approach to handling missing values. A data-driven approach was employed by filling missing values with the mean of the respective column values, as determined by Excel formulas such as AVERAGE and IF. Outliers were identified using Excel functions and statistical techniques and removed from the dataset, ensuring the dataset's accuracy and representativeness. Excel formulas were used to calculate the mean of the respective column values for the missing values and potentially inaccurate data points removed, and this mean value was then used to fill in the missing values. The dataset was consistently maintained throughout the data cleansing process to ensure its reliability and suitability for further analysis and implementation. The cleaned dataset was suitable for extracting meaningful insights, conducting statistical analyses, or applying machine learning algorithms, making it a valuable asset for research and decision-making.

3.3 Calculation of AQI and Evaluation of Air Standards Specified by CPCB

The daily AQI levels for the pollutants were calculated using a macro-enabled AQI calculator in Microsoft Excel, based on the pollutant concentrations extracted from the CPCB website data [55].

The formula used for the calculation of the AQI of the pollutant factors is:

$$I_p = \frac{I_{H_i} - I_{L_o}}{BP_{H_i} - BP_{L_o}} \left(C_P - BP_{L_o} \right) + I_{L_o} \quad (1)$$

where I_p denotes the index for pollutant p, C_P refers to the truncated concentration of pollutant p, BP_{H_i} is the concentration breakpoint at or above C_P, BP_{L_o} is the concentration breakpoint at or below C_P, I_{H_i} is the AQI value for BP_{H_i}, and I_{L_o} is the corresponding AQI value for BP_{L_o}.

The Algorithm followed by the calculator for the calculation of the daily AQI is as follows:

STEP 1: Start

STEP 2: Calculate the average AQI for each pollutant over a 24-hour period every day.

STEP 3: Determine the highest daily average AQI value among all pollutants to identify the most impactful pollutant.

STEP 4: Compute the greatest AQI value across all pollutants to assess the overall air quality status.

STEP 5: Determine the final AQI value.

STEP 6: End

The corresponding AQI categories were determined using a Python code that defined a range for each AQI Standard as specified by CPCB [55]. The air quality categories are provided in Table 1.

Table 1. Air Quality Index (AQI) and Standards Specified by CPCB.

Air Quality Index (AQI)	Air Quality Standard
0-50	Good
51-100	Satisfactory
101-200	Moderate
201-300	Poor
301-400	Very Poor
401-500	Severe

Appropriate Standards were assigned based on the calculated AQI values through the code. A new column was added to the Excel sheet to create a comprehensive dataset including the AQI Standards. This additional column provides a quick and precise classification of Air Quality Standards for each day, facilitating straightforward interpretation and analysis of the data. This methodology was used to prepare datasets for all four South Indian cities - Amaravati, Bengaluru, Chennai, and Hyderabad.

3.4 Machine Learning Models

XGBoost is an extremely efficient and scalable implementation of the gradient boosting framework. It leverages parallel processing capabilities, which significantly speeds up the training process, especially on multicore machines. XGBoost incorporates techniques such as L1 (Lasso) and L2 (Ridge) regularization to control overfitting and enhance the model's generalization capability on unseen data. The objective function of XGBoost consists of a loss function and a regularization term. The loss function, often a squared error term, measures the difference between the actual and predicted values, while the regularization term penalizes model complexity to prevent overfitting.

The objective function can be expressed as:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k) \quad (2)$$

$$\text{where, } l(y_i, \hat{y_i}) = (y_i - \hat{y_i})^2 \text{and } \Omega(f) = \gamma T + \frac{1}{2}\lambda |w|^2 \quad (3)$$

Here, γ and λ are regularization parameters, T is the number of leaves in the tree, and w is the vector of leaf weights. LightGBM is designed for high efficiency and scalability. It employs a histogram-based approach for decision tree learning, which reduces the number of split points by grouping continuous features into discrete bins, significantly speeding up the training process. LightGBM grows trees leaf-wise with depth limitation, splitting the leaf with the largest loss reduction among all current leaves. This strategy often results in better accuracy but may lead to overfitting if not properly regulated. The objective function in LightGBM is similar to XGBoost, incorporating loss and regularization terms. The loss function and regularization terms are adapted for the histogram-based and leaf-wise growth approaches, maintaining the balance between model complexity and performance. The ensemble stack of LightGBM and XGBoost involves training individual LightGBM and XGBoost models and then combining their predictions using a metamodel. This approach leverages the strengths of both models, yielding more accurate and robust predictions by capturing diverse patterns in the data. The combined predictions are typically achieved by averaging the predictions of the individual models or using a meta-model to learn the optimal combination of predictions. Experiments with various data split ratios (75-25, 70-30) revealed only marginal variations in R^2 scores and RMSE values for the LightGBM, XGBoost regressors, and their ensemble stack. Optimal performance was achieved with an 80-20 split, where all three models demonstrated consistently high accuracy, emphasizing the significance of data-split ratio selection in influencing model effectiveness. In this work, an 80% training and 20% testing data split was used to train and evaluate the regression-based models. The performance of the regressor algorithms was assessed using R^2 and RMSE scores. The R^2 score, referred as the coefficient of determination, indicates the percentage of variation in the dependent variable that the independent variables explain. This metric ranges from 0 to 1, where 1 represents a perfect fit. The formula for R^2 is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

where, y_i represents the actual values, $\hat{y_i}$ denotes the predicted values, and $\bar{y}$ represents the mean of the actual values. The numerator expresses the sum of squared residuals, measuring the total deviation of the predicted values from the actual values while the denominator is the total sum of squares, measuring the total deviation of the actual values from the mean. A higher R^2 value indicates that a greater proportion of variance is captured by the model. The RMSE measures the average magnitude of prediction errors, with lower values indicating superior model performance. The formula for RMSE is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2} \tag{5}$$

where, y_i represents the actual values and $\hat{y_i}$ denotes the predicted values. RMSE provides a clear metric for comparing the accuracy of different models, as it directly quantifies the typical error magnitude in the predictions. The square root ensures that the error is measured in the same units as the original values for easier interpretation.

4. Results and Discussion

This section provides a detailed examination of the AQI in four Southern Indian cities - Amaravati, Bengaluru, Chennai, and Hyderabad. Southern India has diverse geographical features, including coastal areas, plains, hills, and forests. This diversity can lead to a wide range of pollution sources, such as industrial activities, vehicular emissions, and natural factors. The process of rapid urbanization and industrialization in cities like Chennai, Bangalore, and Hyderabad have significantly contributed to air pollution through vehicle emissions, industrial activities, and construction. Hence, monitoring air quality in urban and industrial areas can provide insights into pollution dynamics in these environments. Southern India experiences diverse climatic conditions, including monsoons, tropical climates, and coastal influences. These climate patterns can affect the dispersion and accumulation of pollutants in the air, and studying how climate variability influences AQI values can enhance our understanding of the complex interactions between meteorological factors and air quality. Furthermore, Southern India has seen improvements in research and monitoring infrastructure, including air quality monitoring stations. Reliable data collection systems facilitate accurate estimation and prediction of AQI values; therefore the study takes into account these four South Indian cities to ensure comprehensive analysis. The results presented are based on the day-wise data available from 31st May, 2019 to 22nd November, 2023 [55].

4.1 Analysis and Discussion

This section provides individual analyses of the AQI for each city, starting with Amaravati. The analysis indicates that Amaravati has an overall good air quality index over four years shown in Fig. 2. The data primarily consisted of good days, with only a few satisfactory days. This is a testament to the fact that Amaravati has a good green cover around the city, and the city's plantation has made the atmosphere and air quality quite good [12]. However, as the city is still developing as Andhra Pradesh's capital, it needs larger industrial endeavours with less pollution-related activities to ensure good air quality.

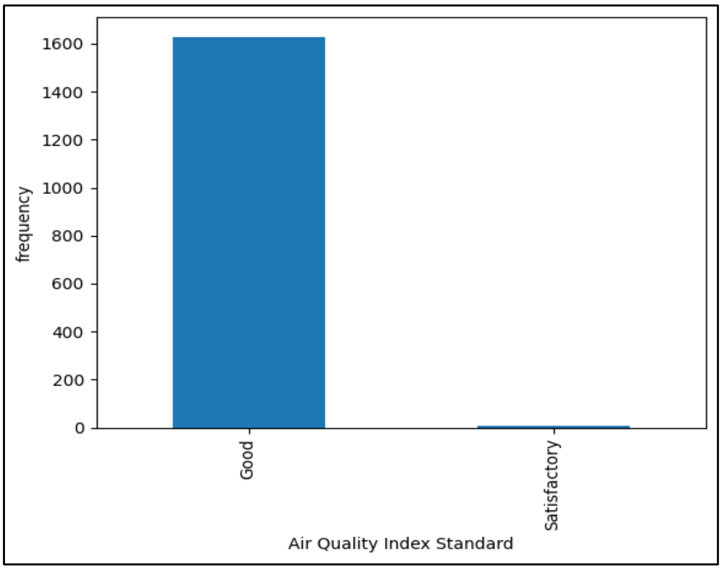

Fig. 2. Comparison of AQI Standard of Days from 2019 to 2023 in Amaravati

The analysis of pollutant concentrations in the air reveals that PM10 concentrations were higher than those of all other pollutants, with ozone being the second-highest pollutant contributor and Carbon monoxide being the least contributing pollutant. The report suggests that the dust generated from construction sites contributes as a major source of particulate matter. As a developing capital city, the region's ongoing construction or infrastructure development could significantly contribute to elevated PM10 levels shown in Fig. 3. Construction sites are known to have high concentrations of suspended particulate matter (PMs), and industrial activities and physical construction are significant sources of PM2.5-10 emissions [14]. Studies have shown that PM10 levels around cement industries and industrialized coastal cities are significantly higher than control areas [14]. The heavy metals and rare-earth elements in atmospheric particulate matter further confirm the impact of industrial emissions [26]. Moreover, Amaravati's warm and sunny climate could contribute to the photochemical reactions leading to elevated ozone levels [15]. Ozone is a secondary pollutant which forms in sunlight through complex photochemical reactions involving precursor pollutants, primarily nitrogen oxides and VOCs [16]. Increased sunlight provides the energy needed to drive these photochemical reactions, producing ozone from the precursors.

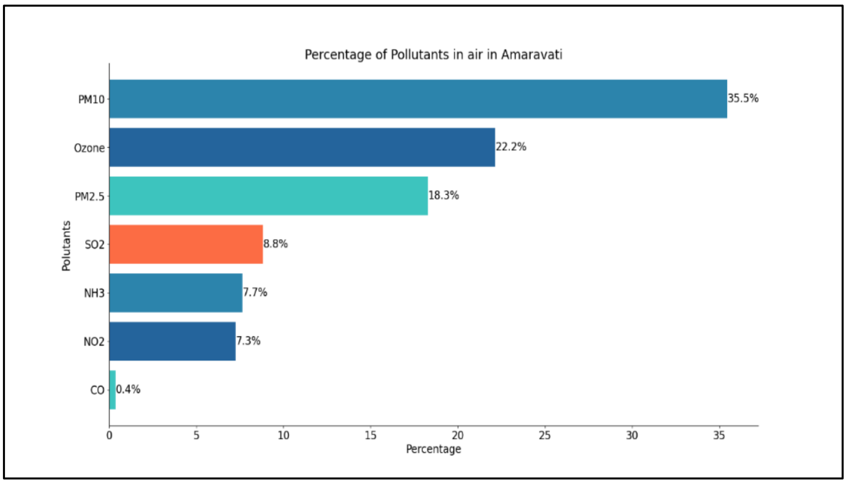

Fig. 3. Comparison of Concentration of Air Pollutants from 2019 to 2023 in Amaravati.

In Bengaluru, the Air Quality Index (AQI) shows that the city experiences a considerable number of good days, with some days being satisfactory and very few being moderately polluted. The city's green cover and numerous parks serve as natural filters, trapping particulate matter and absorbing pollutants, leading to improved air quality [25]. Furthermore, Bengaluru's moderate climate and geographical location may contribute to favourable air quality. The city's elevation, temperature, and weather conditions can impact how pollutants disperse and dilute, leading to cleaner air as shown in Fig. 4.

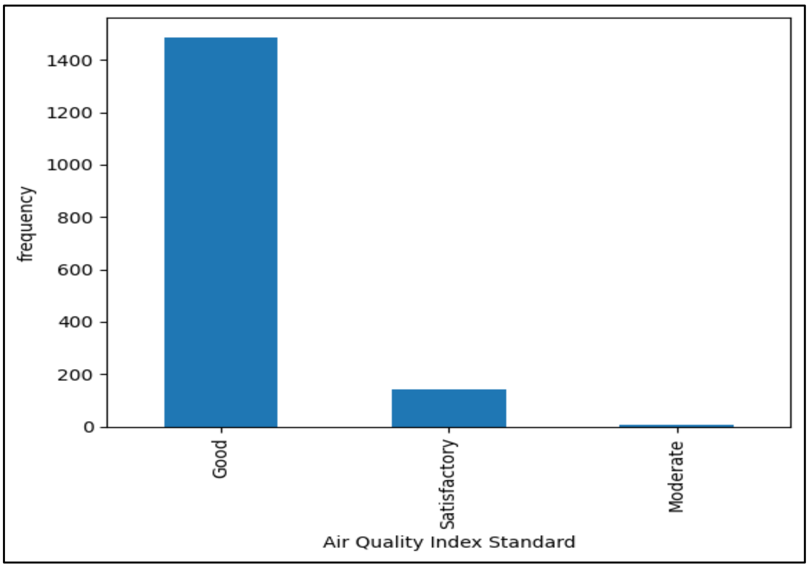

Fig. 4. Comparison of AQI Standard of Days from 2019 to 2023 in Bengaluru.

PM10, PM2.5, and Ozone were found to have a greater contribution to pollution than other pollutants [17,18,20]. Bengaluru, like many megacities, is facing a rise in air pollution issues due to urbanization, industrialization, and economic growth. Prolonged exposure to PM2.5 in Bengaluru has been linked with attributable deaths for chronic respiratory disease, coronary artery disease, stroke, and lung malignancy [17,19]. Exposure to O3 has also been linked to attributable deaths from respiratory diseases [19]. A study also found that air pollution in Bengaluru is increasing due to rising temperatures and annual heavy rainfall [21]. The concentration of air pollutants has been shown in Fig. 5.

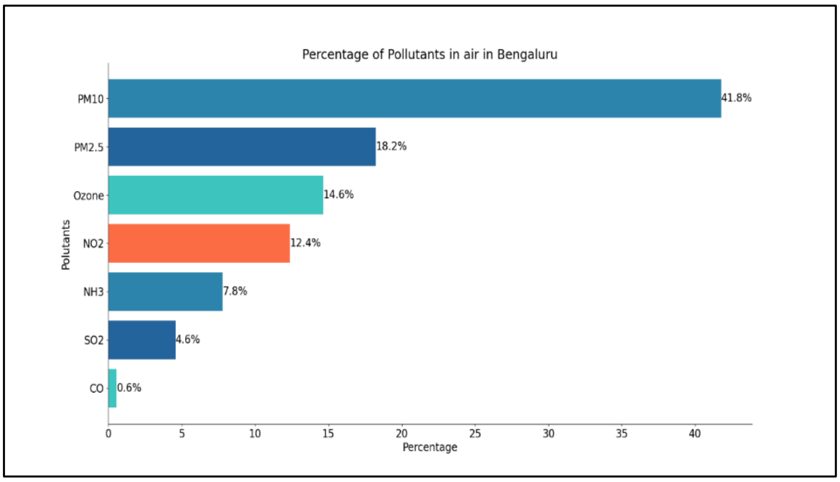

Fig. 5. Comparison of Air Pollutant Concentrations from 2019 to 2023 in Bengaluru.

The Air Quality Assessments conducted for the city of Chennai have revealed air quality that ranges from good to moderate days, as shown in Fig. 6.

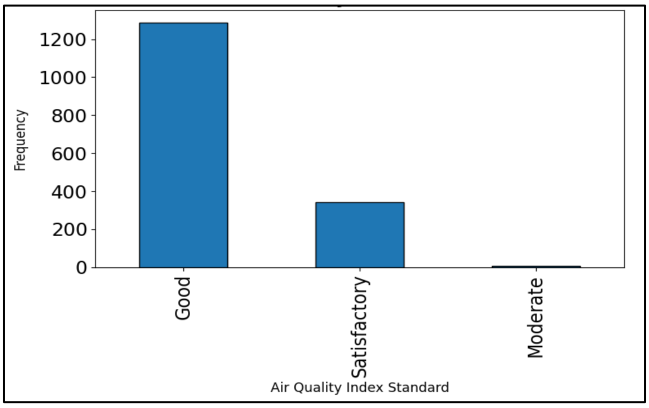

Fig. 6. Comparison of AQI Standard of Days from 2019 to 2023 in Chennai.

The primary contributors to air pollution in Chennai include PM10, PM2.5, NO2, and ground-level ozone as shown in Fig. 7. Particulate matter, both fine and coarse, is primarily caused by vehicular emissions, industrial activities, and construction sites [22,23]. PM10 and PM2.5 are the major contributors to pollution in Chennai, and their sources include vehicular emissions, construction sites, and industrial activities. NO2 is mainly emitted from vehicle exhaust and industrial processes, while ground-level ozone is formed through complex chemical reactions. It is noteworthy that indoor and outdoor pollution are interrelated, with anthropogenic sources such as vehicle emissions and power plant effluents being the primary contributors to both [23,24]. Indoor sources of pollution in Chennai include cooking and heating activities, as well as the use of cleaning agents and personal care products [24].

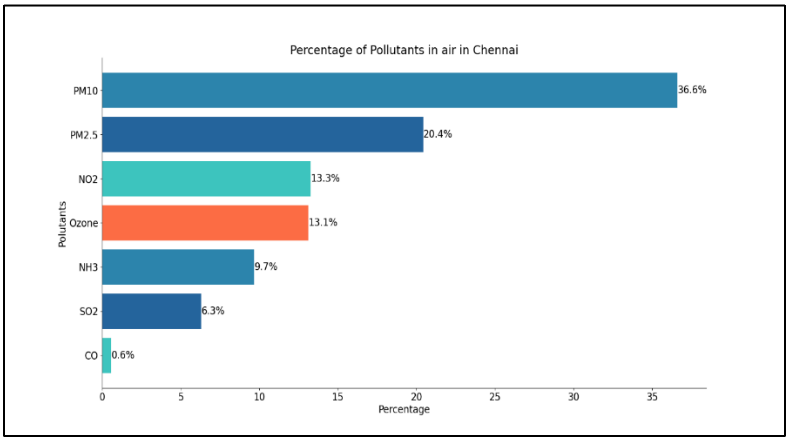

Fig. 7. Comparison of Concentration of Air Pollutants from 2019 to 2023 in Chennai.

Meanwhile, in Hyderabad, the air quality remains a pressing concern due to consistently high levels of PM10 and PM2.5 pollution as shown in Fig. 8. Particulate matter originates from a range of sources, like vehicle emissions, road dust, coal combustion in industries, burning of garbage, and secondary particulate matter. High concentrations of particulate matter pose a significant health risk to the population, with PM2.5 being associated with chronic obstructive pulmonary disease (COPD) [26]. The persistence of elevated particulate matter levels throughout the city emphasizes the need for sustained efforts to mitigate pollution sources and improve air quality management practices.

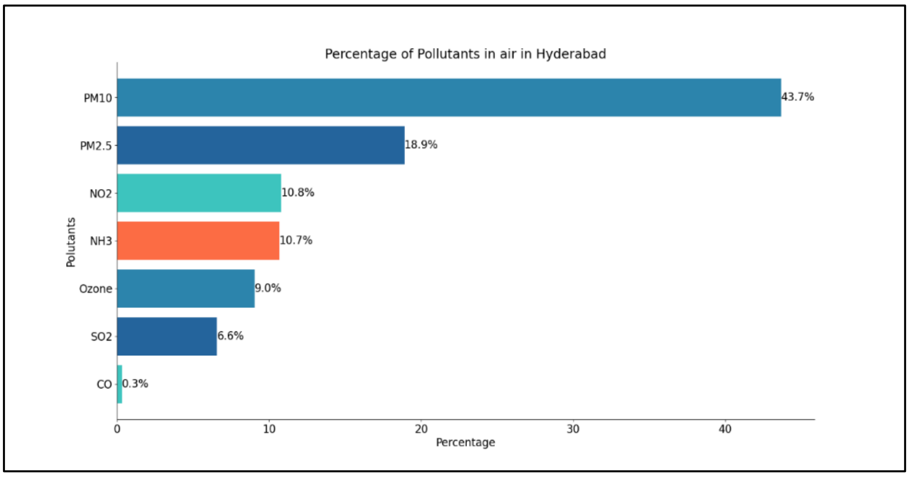

Fig. 8. Comparison of Concentration of Air Pollutants from 2019 to 2023 in Hyderabad.

Despite the city experiencing a mix of good and satisfactory air quality days over the years, there are also instances of moderate, poor and very poor air quality, as illustrated in Fig. 9. This variability underscores the need for sustained efforts to mitigate pollution sources and enhance air quality management practices. Industrial activities prove to be a potential source of air pollution in Hyderabad, and targeted interventions are necessary to reduce their impact [28,29]. Additionally, the use of clean energy sources and the implementation of green transportation options in the city can help reduce the levels of air pollution.

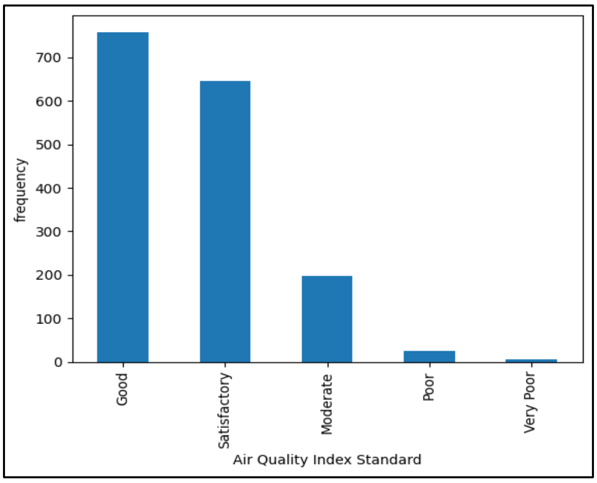

Fig. 9. Comparison of AQI Standard of Days from 2019 to 2023 in Hyderabad.

4.2 Predictions by Machine Learning Models

This study examines day-wise air quality data from the Central Pollution Control Board (CPCB) website [55], covering the period from May 31, 2019, to November 22, 2023, to evaluate the performance of predictive deep learning models. After thorough preprocessing of raw data, comprehensive datasets were created for the four South Indian cities. The datasets were partitioned into training and testing subsets in an 80:20 ratio, with data randomly ordered to ensure that both subsets offered a balanced and representative sample of the overall dataset. This randomized allocation was essential in preventing temporal biases that could occur in time-ordered splits, thus enhancing model robustness and generalizability across different periods. By randomly sampling data points for training and testing, this method ensured a comprehensive reflection of the data's inherent variability, thereby minimizing overfitting and providing a more accurate assessment of model performance. This split in data also demonstrated lower root mean square error (RMSE) and higher R^2 scores, outperforming other split configurations such as 75:25 and 70:30. The RMSE for both the training and testing sets was regularly monitored, and early stopping was applied to cease the training process when performance metrics showed signs of degradation, further reducing the risk of overfitting.

The study evaluates three deep learning architectures—XGBoost, LightGBM, and an ensemble stack combining both—on the prepared air quality datasets of four South Indian cities, using R^2 and RMSE as key performance metrics, as illustrated in Tables 2-5. LightGBM, with its histogram-based algorithm, provides advantages in computational efficiency and scalability, particularly with extensive datasets, whereas XGBoost is noted for its capability in modeling complex patterns within the data. The ensemble model, combining LightGBM and XGBoost, aimed to exploit the strengths of both models, producing a more robust predictor of air quality levels. State-of-the-art research highlights the effectiveness of XGBoost and LightGBM for urban air quality prediction. Notably, LightGBM's histogram-based approach has been shown to improve training speed and accuracy, as demonstrated in studies on air quality data from urban areas such as Beijing [53]. Moreover, the inclusion of key air quality indicators, such as PM2.5, nitrogen dioxide, and sulphur dioxide has shown to contribute to the precision of predictions [54]. XGBoost's proficiency in time series analysis has shown enhanced forecasting abilities by leveraging historical trends to predict future air quality levels [54]. Recent literature also supports the use of hybrid models that integrate regression techniques with XGBoost, effectively addressing overfitting and boosting predictive robustness [54].

Table 2. Comparison of Model Performance for AQI Prediction in Amaravati.

City	Algorithm	R^2 Score	RMSE
Amaravati	XGBoost	0.9960	0.5668
	LightGBM	0.9968	0.5125
	Ensemble Stack	0.9965	0.5353

Table 3. Comparison of Model Performance for AQI Prediction in Bengaluru.

City	Algorithm	R^2 Score	RMSE
Bengaluru	XGBoost	0.9870	2.0229
	LightGBM	0.9848	2.1846
	Ensemble Stack	0.9612	3.4924

Table 4. Comparison of Model Performance for AQI Prediction in Chennai.

City	Algorithm	R^2 Score	RMSE
Chennai	XGBoost	0.9969	0.9540
	LightGBM	0.9972	0.9132
	Ensemble Stack	0.9962	1.0637

Table 5. Comparison of Model Performance for AQI Prediction in Hyderabad.

City	Algorithm	R^2 Score	RMSE
Hyderabad	XGBoost	0.9985	1.6507
	LightGBM	0.9982	1.8373
	Ensemble Stack	0.9935	3.4767

The results indicate that the model performance varied across cities, with LightGBM consistently achieving lower RMSE and higher R^2 scores in cities like Amaravati and Chennai, while XGBoost excelled in Bengaluru and Hyderabad, demonstrating its ability to effectively capture feature relationships in these regions. The Ensemble Stack, however, underperformed, especially in Bengaluru and Hyderabad, likely due to noise introduced by combining models or suboptimal weighting that may have increased prediction error. When models were compared based on RMSE and R² scores, both XGBoost and LightGBM displayed strong predictive capabilities, with LightGBM showing a slight advantage in most instances. The ensemble approach, however, yielded mixed results, often leading to increased RMSE, indicating lower generalizability than individual models.

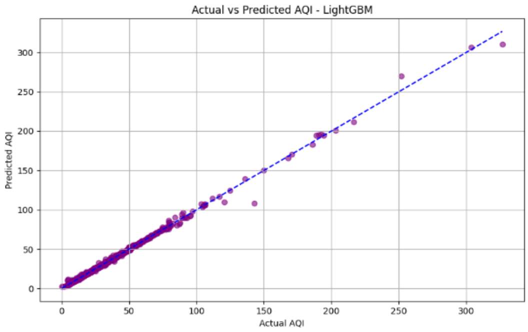

Fig. 10. Visualization of Predictive Performance of LightGBM.

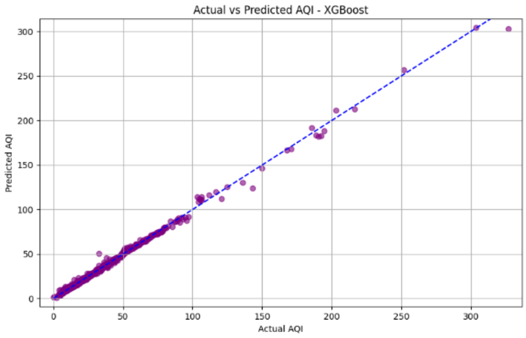

Fig. 11. Visualization of Predictive Performance of XGBoost.

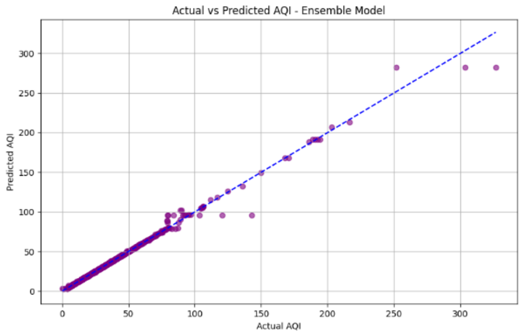

Fig. 12. Visualization of Predictive Performance of Ensemble Model.

For evaluating the predictive performance of the three models, scatter plots of predicted versus actual AQI values provided a clear visualization of each model's accuracy. In each scatter plot, a reference line (y = x) represented ideal predictions, where the model outputs matched actual values accurately. The scatter plot of the LightGBM model, as illustrated in Fig. 10, shows a strong alignment with the reference line, demonstrating a tight clustering of points especially in the low to moderate AQI range (0–150). It also indicates minimal deviation even at higher AQI levels, underscoring the model's high predictive accuracy across the entire AQI spectrum. XGBoost similarly maintained a reasonable degree of accuracy, with data points clustering near the equality line in the lower AQI range as shown in Fig. 11. However, it exhibited a slightly more scatter at elevated AQI levels. The Ensemble model resulted more variability at high AQI levels as illustrated in Fig. 12, where its predictions deviated more significantly from actual values, potentially due to noise or suboptimal weighting in the model integration.

These findings suggest that either XGBoost or LightGBM should be selected based on regional data characteristics, while the ensemble model requires further refinement to improve its utility in air quality forecasting applications. Further exploration into the underlying factors influencing model performance in each city could enhance our understanding of geographic-specific trends, providing a foundation for more tailored and robust predictive frameworks in future

Conclusion and Future Work

Air pollution remains a pressing issue in India, significantly impacting human health and the environment. Despite efforts to mitigate this challenge, further research is essential to identify and implement sustainable, long-term solutions for pollution control. The convergence of COVID-19 with air pollution dynamics in urban settings has created an opportunity for valuable insights into how environmental and health crises may interact, underscoring the need for comprehensive air quality management strategies. Integrating these insights can play a crucial role in reducing health risks and environmental damage, which is further validated by numerous global studies advocating for stringent policy actions to curb pollutants. Within this framework, advanced monitoring systems and machine learning technologies have become vital to Indian urban air quality management, providing predictive accuracy and actionable insights for policymakers. Future research can expand into developing spatiotemporal models for each primary air pollutant across various regions in India. By leveraging advanced geostatistical frameworks, these models could provide detailed analyses of pollution patterns at different spatial and temporal scales, improving the precision of air quality predictions across diverse zones. This approach would enable a targeted understanding of pollution sources and variations, supporting region-specific interventions. Furthermore, spatiotemporal models would enhance predictive capabilities by accounting for both location-based and time-based factors, enabling more adaptive and responsive pollution management strategies. Through such an integrative approach, modeling techniques could greatly contribute to a comprehensive solution for India's escalating air pollution crisis. These findings can guide policymakers in crafting informed, sustainable policies that effectively address pollution levels, fostering a healthier and more sustainable environment for future generations.

References

[1] Outa, J.O., Kowenje, C.O., Plessl, C. *et al.* Distribution of arsenic, silver, cadmium, lead and other trace elements in water, sediment and macrophytes in the Kenyan part of Lake Victoria: spatial, temporal and bioindicative aspects. *Environ Sci Pollut Res* 27, 1485–1498 (2020). https://doi.org/10.1007/s11356-019-06525-9.

[2] Mehndiratta, M. M., & Garg, D. (2023). Beware! We are Skating on a Thin Ice: Air Pollution is a Killer. *The Journal of the Association of Physicians of India*, *71*(7), 11–12.

[3] Addisu, A. (2023). Indoor Air Pollution. IntechOpen. doi: 10.5772/intechopen.110587

[4] Gaikwad, Asha & Shivhare, Niharika. (2020). INDOOR AIR POLLUTION-A Threat.

[5] Kim Y, Radoias V. Severe Air Pollution Exposure and Long-Term Health Outcomes. Int J Environ Res Public Health. 2022 Oct 27;19(21):14019. doi: 10.3390/ijerph192114019. PMID: 36360899; PMCID: PMC9655248.

[6] Faustini, Annunziata. (2021). Air Pollutants Short-Term and Long Term Effects. 10.1016/B978-0-08-102723-3.00181-5.

[7] Cucchi I, Chanel O. Long-term health and economic impacts of air pollution in Greater Geneva. JAPH. 2023;8(2):135-156.

[8] Bostan, Pınar. (2022). Health effects of outdoor air pollutants. World Journal of Environmental Research. 12. 70-81. 10.18844/wjer.v12i2.8095.

[9] A. S. Sofia, S. V. J, S. K, S. K and T. M, "APD - ML: Air Pollution Detection Using Machine Learning Algorithms," *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, Vellore, India, 2023, pp. 1-5, doi: 10.1109/ViTECoN58111.2023.10157131.

[10] S. Rani, P. Kumari and S. K. Singh, "Machine Learning-based Multiclass Classification Model for Effective Air Quality Prediction," *2023 IEEE IAS Global Conference on Emerging Technologies (GlobConET)*, London, United Kingdom, 2023, pp. 1-7, doi: 10.1109/GlobConET56651.2023.10149947.

[11] Sharma, Meghna & Gupta, Eishita & D., Viji. (2023). Air Quality Index (AQI) Prediction using Automated Machine Learning with TPOT-ANN*. 1-9. 10.1109/RAEEUCCI57140.2023.10134166.

[12] Ghadei, Madhusmita. (2018). Amaravati - A City Reborn, Journey Towards a World-Class Smart City. 15-29. 10.1007/978-3-319-61645-2_2.

[13] N., V., Krishna, Prasad., P., Sasikala., N., Madhavi., T., Anil, Babu., Thomaskutty, Mathew., S., Ramesh., M., S., S., R., K., N., Sarma., B., Ramesh, Naik. (2020). Analysis of seasonal variation in particulate matter and relevant pollutants for three stations of andhra pradesh (India) during the period (2018-2020) using multivariate regression analysis. doi: 10.37418/AMSJ.9.12.35

[14] Usharani, Bhimavarapu. "Long-Term Effects of Climate Change on Housing Market analytics in Amaravati, Capital of Andhra Pradesh in India, Using Machine Learning." In *Handbook of Research on Climate Change and the Sustainable Financial Sector,* edited by Odunayo Magret Olarewaju and Idris Olayiwola Ganiyu, 331-352. Hershey, PA: IGI Global, 2021. https://doi.org/10.4018/978-1-7998-7967-1.ch020

[15] Ghosh, Debreka & Sarkar, Ujjaini. (2015). Analysis of the photochemical production of ozone using Tropospheric Ultraviolet-Visible (TUV) Radiation Model in an Asian megacity. Air Quality, Atmosphere & Health. 9. 10.1007/s11869-015-0346-3.

[16] Lupaşcu, A., Otero, N., Minkos, A., and Butler, T.: Attribution of surface ozone to NO_x and volatile organic compound sources during two different high ozone events, Atmos. Chem. Phys., 22, 11675–11699, https://doi.org/10.5194/acp-22-11675-2022, 2022.

[17] Dhanya, G., T. S. Pranesha, K. Nagaraja, D. M. Chate, and G. Beig. 2022. "Variation of Ozone, Carbon Monoxide, and Oxides of Nitrogen at Bengaluru, India". *Journal of Scientific Research* 14 (2):459-70. https://doi.org/10.3329/jsr.v14i2.55626.

[18] Prabhu, V., Singh, P., Kulkarni, P. *et al.* Characteristics and health risk assessment of fine particulate matter and surface ozone: results from Bengaluru, India. *Environ Monit Assess* 194, 211 (2022). https://doi.org/10.1007/s10661-022-09852-6

[19] Peter, Anju & Raj, Monish & Gangadharan, Praveena & Pavizham, Athira & SM, Shiva Nagendra. (2023). Trends, Extreme Events and Long-term Health Impacts of Particulate Matter in a Southern Indian Industrial Area. Water, Air, & Soil Pollution. 234. 10.1007/s11270-023-06302-y.

[20] Yadav, Rahul Kant & Gadhavi, Harish & Arora, Akanksha & Mohbey, K. & Kumar, Sunil & Lal, Shyam & Mallik, Chinmay. (2023). Relation between PM 2.5 and O 3 over Different Urban Environmental Regimes in India. 10.3390/urbansci7010009.

[21] Gupta, Jyothi. (2023). Statistical Assessment of Spatial Autocorrelation on Air Quality in Bengaluru, India. 17. 12. 10.1007/978-3-031-31164-2_21.

[22] P. Sujatha, Jai Shirisha PVS, Krishna Nivash J. and P.V.S. Janardhanam (2023); IMPACT OF URBANIZATION IN CHENNAI *Int. J. of Adv. Res.* 11 (May). 851-871 (ISSN 2320-5407). www.journalijar.com

[23] Nirmala, Muthu & Mallika, M.. (2023). Trend Analysis of Air Quality Index in Chennai, India. 10.9734/bpi/npgees/v6/18578D.

[24] Karuppannan, Shankar & Narasimhan, C. Lakshmi & Hussain, Sajjad & Almohamad, Hussein & Abdullah, Ahmed & Dughairi, Ahmed & Al-Mutiry, Motrih & Alkayyadi, Ibrahim & Abdo, Hazem & Li, Chi & Lu, Xiao & Zhang, Yuqiang & K, Manikanda Bharath & Natesan, Usha. (2022). Multivariate Urban Air Quality Assessment of Indoor and Outdoor Environments at Chennai Metropolis in South India. Atmosphere.

[25] Pillai, Priyadarshini & Taylor, George. (2023). Evaluating Air Pollution Tolerance Index (APTI) of Some Plants Species in Bengaluru City. Advances in Zoology and Botany. 11. 139-149. 10.13189/azb.2023.110206.

[26] Maring, Teshilring & Suman, Dr & Jha, Ajay & Kumar, Naresh & Pandey, Shri. (2023). Airborne Particulate Matter and Associated Heavy Metals: A Review. Macromolecular Symposia. 407. 10.1002/masy.202100487.

[27] Jung, Miyeon, Daegon Cho, and Kwangsoo Shin. 2019. "The Impact of Particulate Matter on Outdoor Activity and Mental Health: A Matching Approach" *International Journal of Environmental Research and Public Health* 16, no. 16: 2983. https://doi.org/10.3390/ijerph16162983

[28] Thanusree, G & Kiran, M & Reddy, A & Rangaswamy, Prem Sudha & Parida, Arati & Aakanksha, Keesagani & Reddy, G. (2023). Air Quality Monitoring at Heavy Traffic Zone in Hyderabad. 2581-9429. 10.48175/IJARSCT-9526.

[29] Kumar, Athur & Kalyan, Kumar & Rama, Krishna. (2023). Assessment of dispersion of pollutants due to Industrial Sources using AERMOD near Hyderabad city, India. Research Journal of Chemistry and Environment. 27. 114-121. 10.25303/2702rjce1140121.

[30] C. Wieland and V. Pankratius, "Regressor-Rater: A Resource-Efficient One-Shot Learner For Estimating Prediction Intervals," *2023 IEEE World AI IoT Congress (AIIoT)*, Seattle, WA, USA, 2023, pp. 0127-0133, doi: 10.1109/AIIoT58121.2023.10174567.

[31] Sukuman, Thanakon & Ueda, Kayo & Sujaritpong, Sarunya & Praekunatham, Hirunwut & Punnasiri, Kornwipa & Wimuktayon, Tuangsit & Prapaspongsa, Trakarn. (2023). Health Impacts from PM2.5 Exposure Using Environmental Epidemiology and Health Risk Assessment: A Review. Applied Environmental Research. 1-14. 10.35762/AER.2023010.

[32] Garcia, Amanda & Helena, Eduarda & De Falco, Anna & Ribeiro, Joaquim & Gioda, Carolina. (2023). Toxicological Effects of Fine Particulate Matter (PM2.5): Health Risks and Associated Systemic Injuries—Systematic Review. Water, Air, & Soil Pollution. 234. 10.1007/s11270-023-06278-9.

[33] Li, Ding & Xiao, Han & Ma, Shuang & Zhang, Jiangxue. (2021). Health Benefits of Air Quality Improvement: Empirical Research Based on Medical Insurance Reimbursement Data. SSRN Electronic Journal. 10.2139/ssrn.3952375.

[34] Aithal, Sathya & Sachdeva, Ishaan & Kurmi, Om. (2023). Air quality and respiratory health in children. Breathe. 19. 230040. 10.1183/20734735.0040-2023.

[35] Yu Z, Merid SK, Bellander T, Bergström A, Eneroth K, Georgelis A, Hallberg J, Kull I, Ljungman P, Klevebro S, Stafoggia M, Wang G, Pershagen G, Gruzieva O, Melén E. Associations of improved air quality with lung function growth from childhood to adulthood: the BAMSE study. Eur Respir J. 2023 May 5;61(5):2201783. doi: 10.1183/13993003.01783-2022. PMID: 36822631; PMCID: PMC10160798.

[36] Álvarez Aldegunde, José & Quiñones-Bolaños, Edgar & Fernández Sánchez, Adrián & Saba, Manuel & Caraballo, Luis. (2023). Environmental and Health Benefits Assessment of Reducing PM 2.5 Concentrations in Urban Areas in Developing Countries: Case Study Cartagena de Indias. Environments. 10. 10.3390/environments10030042.

[37] S. Aggrawal and B. Bhushan, "Machine Learning for Air Pollution," *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, Vellore, India, 2023, pp. 1-6, doi: 10.1109/ViTECoN58111.2023.10157028.

[38] Shivakumar, Sheethal & Shastry, K Aditya & Singh, Simranjith & Pasha, Salman & Vinay, B. & Sushma, V.. (2022). Machine Learning-Based Air Pollution Prediction. 10.1007/978-981-16-3342-3_2.

[39] Ayus, I., Natarajan, N. & Gupta, D. Comparison of machine learning and deep learning techniques for the prediction of air pollution: a case study from China. *Asian J. Atmos. Environ* 17, 4 (2023). https://doi.org/10.1007/s44273-023-00005-w

[40] Vm, Madhuri, Samyama Gunjal Gh and Savitha Kamalapurkar. “Air Pollution Prediction Using Machine Learning Supervised Learning Approach.” *International Journal of Scientific & Technology Research* 9 (2020): 118-123.

[41] Venkatraman, S. and Sivanesh, S. (2023) “Air Quality Prediction using Ensemble Voting based Deep Learning with Mud Ring Algorithm for Intelligent Transportation Systems”, Global NEST Journal, 25(6). Available at: https://doi.org/10.30955/gnj.004810.

[42] Cooper Loughlin, Dimitris Manolakis, Vinay Ingle, "Multivariate air quality time series analysis via a recurrent variational deep learning model," Proc. SPIE 12525, Geospatial Informatics XIII, 125250G (15 June 2023); https://doi.org/10.1117/12.2663201

[43] Singh T, Sharma N, Satakshi, Kumar M. Analysis and forecasting of air quality index based on satellite data. Inhal Toxicol. 2023 Jan-Feb;35(1-2):24-39. doi: 10.1080/08958378.2022.2164388. Epub 2023 Jan 5. PMID: 36602767.

[44] Hossain, Mohammed Tahmid & Hossain, Afra & Meem, Sabrina & Monir, Md Fahad & Miah, Md Saef Ullah & Sarwar, Talha. (2023). Impact of COVID-19 Lockdowns on Air Quality in Bangladesh: Analysis and AQI Forecasting with Support Vector Regression. 1-6. 10.1109/INCET57972.2023.10169997.

[45] Bhatti, M. A., Song, Z., Bhatti, U. A., & Ahmad, N. (2023). Predicting the Impact of Change in Air Quality Patterns Due to COVID-19 Lockdown Policies in Multiple Urban Cities of Henan: A Deep Learning Approach. *Atmosphere*, *14*(5), 902. https://doi.org/10.3390/atmos14050902

[46] Hina, S., Saleem, F., and Ibrahim, M.: COVID-19 Pandemic Hopeful Prospect: Air Quality Improvements over Indo-Gigantic Plain, EGU General Assembly 2023, Vienna, Austria, 23–28 Apr 2023, EGU23-1112, https://doi.org/10.5194/egusphere-egu23-1112, 2023.

[47] Bhattarai, H., Tai, A. P. K., Val Martin, M., & Yung, D. H. Y. (2024). Impacts of changes in climate, land use, and emissions on global ozone air quality by mid-21st century following selected Shared Socioeconomic Pathways. *The Science of the total environment*, *906*, 167759. https://doi.org/10.1016/j.scitotenv.2023.167759

[48] Eastham, S. D., Monier, E., Rothenberg, D., Paltsev, S., & Selin, N. E. (2023). Rapid Estimation of Climate-Air Quality Interactions in Integrated Assessment Using a Response Surface Model. *ACS environmental Au*, *3*(3), 153–163. https://doi.org/10.1021/acsenvironau.2c00054

[49] van Garderen, L., Feser, F., Mindlin, J., and Shepherd, T.: Attributing Extreme Weather Events and Mean Climate Change using Dynamical and Event Storylines, EGU General Assembly 2023, Vienna, Austria, 24–28 Apr 2023, EGU23-17183, https://doi.org/10.5194/egusphere-egu23-17183, 2023.

[50] Turnau, R., Robinson, W. A., Lackmann, G. M., & Michaelis, A. C. (2022). Model projections of increased severity of heat waves in Eastern Europe. *Geophysical Research Letters*, 49, e2022GL100183. https://doi.org/10.1029/2022GL100183

[51] Zeng, G., Williams, J. E., Fisher, J. A., Emmons, L. K., Jones, N. B., Morgenstern, O., Robinson, J., Smale, D., Paton-Walsh, C., and Griffith, D. W. T.: Multi-model simulation of CO and HCHO in the Southern Hemisphere: comparison with observations and impact of biogenic emissions, Atmos. Chem. Phys., 15, 7217–7245, https://doi.org/10.5194/acp-15-7217-2015, 2015.

[52] De Sario, M., Katsouyanni, K., & Michelozzi, P. (2013). Climate change, extreme weather events, air pollution and respiratory health in Europe. *The European respiratory journal*, *42*(3), 826–843. https://doi.org/10.1183/09031936.00074712

[53] Su, Y. (2020). Prediction of air quality based on Gradient Boosting Machine Method. *2020 International Conference on Big Data and Informatization Education (ICBDIE)*, 395-397.

[54] A. A. Varghese, J. Krishnadas and A. M. Antony, "Robust Air Quality Prediction Based on Regression and XGBoost," *2023 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA)*, Ernakulam, India, 2023, pp. 1-6, doi: 10.1109/ACCTHPA57160.2023.10083379.

[55] Central Pollution Control Board (CPCB, 2022). *Air quality data (2019 - 2022)*. https://cpcb.nic.in/