Air Quality Analysis of Madurai City using Machine learning approach

Mugesh S[1], Gayathri M[2]
[1]*PG Student, Department of Civil Engineering, Alagappa Chettiar Government College of Engineering and Technology, Karaikudi*
[2] *Assistant Professor, Department of Civil Engineering, Alagappa Chettiar Government College of Engineering and Technology, Karaikudi*

***Abstract*—The deteriorating air quality in urban centres like Madurai necessitates a comprehensive understanding of its contributing factors and potential mitigation strategies. This study presents an in-depth analysis and modelling of air quality parameters in Madurai, focusing on particulate matter (PM2.5 and PM10). Utilizing a combination of ground-based monitoring data, satellite imagery, and meteorological observations, this research investigates the spatial and temporal variations of air pollutants across different regions of Madurai. Statistical analyses and machine learning techniques are employed to identify key sources and factors influencing air pollution levels. Several types of regression analysis such as Linear regression, Random Forest regression, Gradient boosting regression, ridge regression, lasso regression, MLP regression, K Neighbours (KNN) regression, and Decision tree regression are performed and the analysis best suited for predicting the air quality of Madurai is depicted. Additionally, predictive models are developed to forecast future air quality scenarios under varying socio-economic and environmental conditions. The findings of this study contribute valuable insights for policymakers, urban planners, and environmental agencies to formulate effective strategies for improving air quality and public health in Madurai.**

Keywords: PM2.5, PM10, Pollutants, Analysis, AQI, Regression

I. INTRODUCTION

Air pollution has emerged as a significant environmental and public health challenge in India[1]. The primary causes include industrial emissions from factories and power plants, vehicular emissions [2], agricultural practices like stubble burning, construction activities, and domestic sources such as biomass burning[3] . Natural events like dust storms and forest fires also contribute. This pollution has severe health impacts, causing respiratory and cardiovascular diseases and increasing the risk of premature death[4]. The environment suffers as well, with air pollution degrading soil, water quality, and ecosystems[5]. The economic costs are significant, encompassing healthcare expenses and productivity losses[6]. In response, the Indian government has implemented measures like the National Clean Air Programme (NCAP)[7], stricter vehicle emission standards [8] through the Bharat Stage norms, and initiatives promoting public transport and electric vehicles[9]. The National Clean Air Programme (NCAP) in India is a strategic initiative launched in 2019 to combat the country's severe air pollution[10]. Aiming to reduce particulate matter pollution (PM10 and PM2.5) by 20-30% by 2024 [11], the program targets 102 cities that consistently fail to meet national air quality standards. NCAP's approach includes developing city-specific action plans, enhancing the air quality monitoring network, and enforcing stricter emission standards for industries and vehicles. NCAP classified 132 cities as non-attainment cities[12] that do not meet the National Ambient Air Quality Standards (NAAQS) [13] set by regulatory authorities for various pollutants, including particulate matter (PM10 and PM2.5), nitrogen oxides (NOx), sulfur dioxide (SO2), carbon monoxide (CO), and ozone (O3)[14]. Madurai is one among the nonattainment cities choosen by NCAP to mitigate ambient air quality[15]. Madurai, a bustling city in the southern part of India, is also grappling with significant challenges in terms of air quality[16]. The city has witnessed rapid urbanization and industrialization in recent decades, leading to a rise in pollution levels.[5] Understanding the dynamics of air quality in Madurai is crucial for addressing the health and environmental concerns faced by its residents[17]. The Particulate

matter(PM2.5 and PM10) are of major concern. PM2.5 refers to fine particulate matter with a diameter of 2.5 micrometers or smaller, and PM10 includes particles with a diameter of 10 micrometers or smaller[18]. These particles get into the lungs causing respiratory diseases. This analysis delves into various aspects of air quality in Madurai, including its sources, pollutants of concern, influence of meteorological parameters such as temperature, wind speed, humidity in air quality.[19] The study involves conducting a comprehensive analysis of air quality in Madurai city, focusing on key pollutants (PM 2.5 and PM 10) and their spatial and temporal distribution[20]. Various machine learning techniques [21]such as Linear regression, Random Forest regression, Gradient boosting regression, ridge regression, lasso regression, MLP regression, K Neighbours regression, and Decision tree regression were performed and their predictions is further studied[22]. This diverse set of models provides a comprehensive framework for analyzing air quality data[23], identifying key predictors of pollution levels, and understanding temporal and spatial variations in air quality across Madurai[24]. By examining these factors, we can develop insights into the state of air quality in Madurai and explore strategies for improving it sustainably.

II. COLLECTION OF DATASET

A. Dataset

The dataset is the input parameter required for performing analysis. The dataset for two years namely 2022-2023 were collected. The concentration of key pollutants were collected . The meteorological data namely temperature, humidity , windspeed and it's direction were collected. The data collected were added as input.

B. Pollutant Concentration

The raw data of key pollutants namely PM 2.5, PM10, CO2, SO, NO and its AQI value are collected from Tamil Nadu Pollution Control Board, Kappalur, Madurai. The pollutant dataset for two years (2022-2023) is collected and compared with predicted value obtained from regression analysis.

C. Meterological data

Meteorological factors such as wind speed, direction, temperature,and humidity, exert profound influences on air pollution levels and patterns. The metereological datas are collected from Weather Research and Forecasting web portal. By influencing the transport, dispersion, and transformation of pollutants, meteorological conditions shape the spatial and temporal distribution of air pollution.

D. Topographical data

The topographical data of Madurai is generated using QGIS software. The air quality monitoring station in Madurai is in four stations namely TNPCB Kappalur, Birla Vishram, Pichai Pillai Chavadi, Hotel Tamil Nadu. The Kappalur is continuous ambient air quality monitoring station and other three are manual monitoring stations. The site coordinates are input in spreadsheet and topographical data is generated by QGIS software[25].

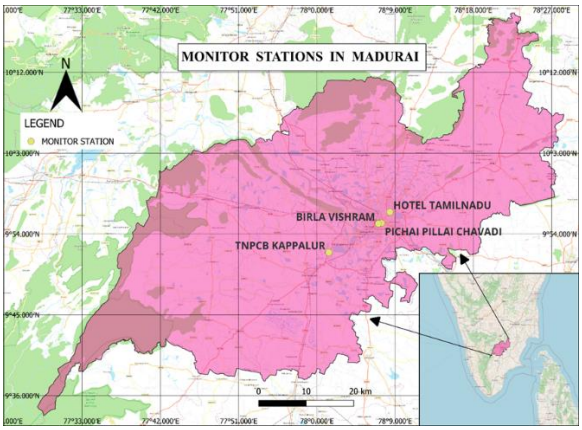

Fig.I Topography of Madurai City

III. METHODOLOGY

A. Data Refinement

To prepare the data for analysis, instances with missing values in the input parameters were removed. The interpolation is performed and the missing value are depicted with the mean value of that input parameter[26].

B. Data Interpretation

All parameters were transformed and the highest value for each instance was chosen for analysis. Since the input have multiple parameters, the data has to be normalized. It has to be scaled within a particular range so that all parameters get equal weightage.

C. Feature selection

Feature selection involves choosing a subset of the initial features that provide relevant information for predicting the output[27].

D. Training the model

The regression techniques namely Linear regression, Random Forest regression, Gradient boosting regression, lasso regression, MLP regression, K Neighbours regression, and Decision tree regression were implemented using Python programming which is as on open source machine learning library[28]. The google collab, which is an open-source Python data science platform was used for accessing Python Notebook (an open-source Python editor) for programming in Python. There were three cases of output - first case for PM2.5, second case for PM10 and last case of AQI of gases. Hence, there are total three sets of training data where each was trained using eight regression models[29]. The comparison of actual values and the estimated values using the eight regression models for normalized PM2.5, PM 10 and AQI at Madurai city is analyzed.

a) Training Phase: During this phase, the model is trained using the dataset, fitting a line or curve based on the chosen algorithm.

b) Testing Phase: The model is then tested with input data to evaluate its performance and its accuracy is checked.

The algorithm will learn the relationships between the input features (pollutant concentrations, meteorological variables) and the target variable (air quality) [30]. The performance of the trained model is evaluated using the testing dataset. Common evaluation metrics for regression tasks include Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) score[31]. These metrics quantify the accuracy and goodness-of-fit of the model's predictions. The trained model is utilized to make predictions on unseen or future air quality data. The predicted air quality trends is visualized over time and compared them with actual observations to assess the model's performance and gain insights into air pollution dynamics.

IV. RESULTS AND DISCUSSION

A. Results

(1) PM10 prediction

The PM10 value is predicted using different regressive techniques and the following results are obtained. Off these, Random forest regression shows a better R^2 value of 0.95916(train) and 0.82101(test), followed by Gradient Boosting of R^2 value of 0.86828 (train) and 0.79132(test). The Error also minimized in Random forest followed by Gradient boosting & decision tree. The Decision Tree and KNN models perform well on training data but with a significant drop in performance on the testing set. The Gradient Boosting and Random Forest are the most effective models in this prediction of PM10, showing high R^2 values and low error metrics, indicating strong predictive power and generalization.

Table.I PM10 Prediction results

	PM10(TRAIN)				PM10(TEST)			
	MSE	RMSE	MAE	R2	MSE	RMSE	MAE	R2
Linear regression	1.17321	0.07643	0.23238	0.71545	1.84428	0.85108	0.74321	0.68208
Decision tree	0.29853	0.35721	0.09849	0.81402	1.12703	1.06229	0.23001	0.61712
MLP regression	0.43842	0.68414	0.11328	0.74728	0.02827	0.16818	0.11235	0.72123
KNN regression	0.33741	1.65453	1.04749	0.79224	2.63038	1.62239	1.19222	0.76043
Ridge	0.49043	0.38126	0.07523	0.67228	0.58143	0.67327	0.51907	0.62113
Lasso	0.30222	0.45715	0.09715	0.65338	0.41118	0.52616	0.27406	0.63902
Random forest	0.11804	0.24428	0.04515	0.95916	0.26129	0.37223	0.09643	0.82101
Gradient Boosting	0.20444	0.35637	0.03001	0.86828	0.31235	0.59507	0.16521	0.79132

(2) PM2.5 Prediction

Various regression models predicting PM2.5 levels, including both training and testing data are generated. Random Forest and Gradient Boosting are the best performing models, with high R^2 values of 0.96414 and R^2 values of 0.91417 respectively consistently across training datasets, with high R^2 values of 0.88302 and R^2 values of 0.82542 respectively consistently across training and testing datasets reasonable error metrics and low error metrics, indicating robust predictive power and good generalization. Lasso regression shows consistent performance with modest error metrics and decent R^2 values. This model balances bias and variance well, likely due to its regularization.

Table.II PM 2.5 Prediction results

	PM2.5(TRAIN)				PM2.5(TEST)			
	MSE	RMSE	MAE	R2	MSE	RMSE	MAE	R2
Linear regression	0.03703	0.07623	0.03221	0.83332	0.05508	0.68123	0.04301	0.79211
Decision tree	0.03543	0.07232	0.03110	0.89223	0.04505	0.59247	0.05723	0.85905
MLP regression	0.04821	0.08654	0.01901	0.69409	0.05308	0.36651	0.07201	0.61617
KNN regression	0.05613	0.09514	0.13623	0.72816	0.19405	0.21521	0.09112	0.70225
Ridge	0.06534	0.18441	0.08135	0.67214	0.05502	0.43908	1.37918	0.66438
Lasso	0.04903	0.09203	0.04538	0.74818	0.06817	0.34021	0.07807	0.71404
Random Forest	0.01611	0.12504	0.02419	0.96414	0.02934	0.16607	0.05509	0.88302
Gradient Boosting	0.02221	0.03400	0.02728	0.91417	0.05703	0.12123	0.03945	0.82542

(3) AQI Prediction

The Air Quality Index (AQI) was predicted using various regression techniques and its results are generated. The decision tree model shows good performance on the training data with a high R^2 value, but the performance drops on the testing data, as indicated by the increase in error metrics and lower R^2. The MLP regressor shows decent performance on both training and testing datasets, with consistent R^2 values and moderate error metrics. KNN, Ridge, and Lasso models show less effective performance, with higher error metrics and lower R^2 values, suggesting they may not capture the complexity of the data as well as other models Random Forest and Gradient Boosting stand out as the most effective models for AQI prediction in this analysis, offering strong generalization capabilities and accurate predictions. Random Forest is particularly noteworthy for its lower errors and higher R^2 values of 0.95732 at train and 0.81811 at test phase, making it the best choice in the prediction of AQI of Madurai city.

Table.III AQI Prediction results

	AQI (TRAIN)				AQI (TEST)			
	MSE	RMSE	MAE	R2	MSE	RMSE	MAE	R2
Linear regression	1.23702	0.07112	0.13200	0.62405	1.97422	0.97321	0.84321	0.67221
Decision tree	0.43623	0.51832	0.39807	0.87122	0.65523	0.80913	0.78258	0.65111
MLP regression	0.33831	0.48833	0.18705	0.70531	0.04608	0.21432	0.57152	0.75303
KNN regression	1.52321	1.71224	1.06824	0.79225	2.62728	1.62124	1.19246	0.76021
Ridge	0.41208	0.45357	0.06115	0.64428	0.61349	0.88407	0.44107	0.67222
Lasso	0.50504	0.45365	0.06219	0.62827	0.76825	0.05402	0.84306	0.59154
Random Forest	0.19502	0.24227	0.05724	0.95732	0.32508	0.39303	0.11711	0.81811
Gradient Boosting	0.24508	0.31323	0.06621	0.86407	0.55421	0.79544	0.16521	0.78710

B.DISCUSSION

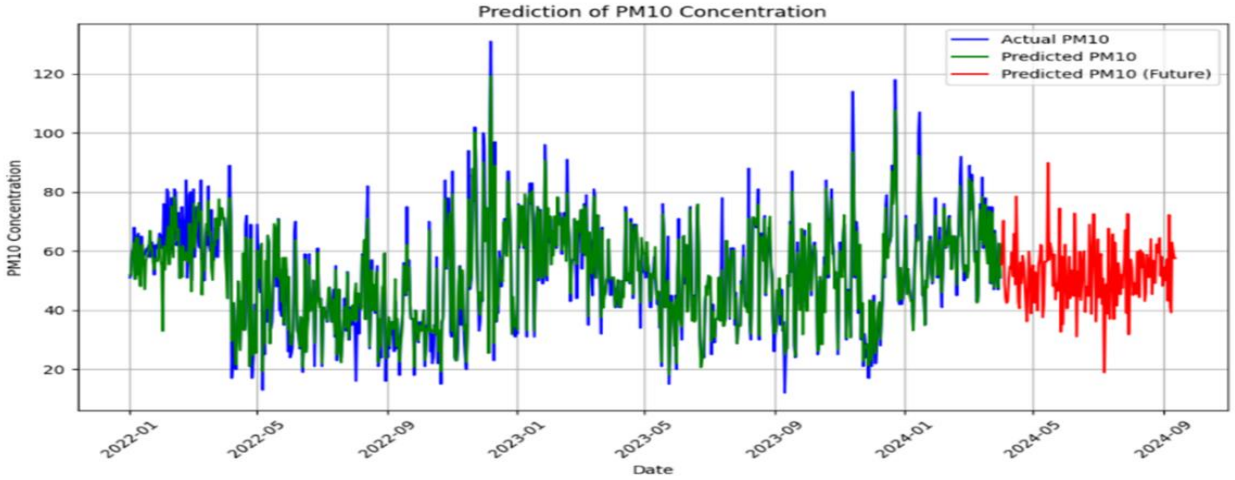

Fig.II Random Forest regression(PM10 prediction)

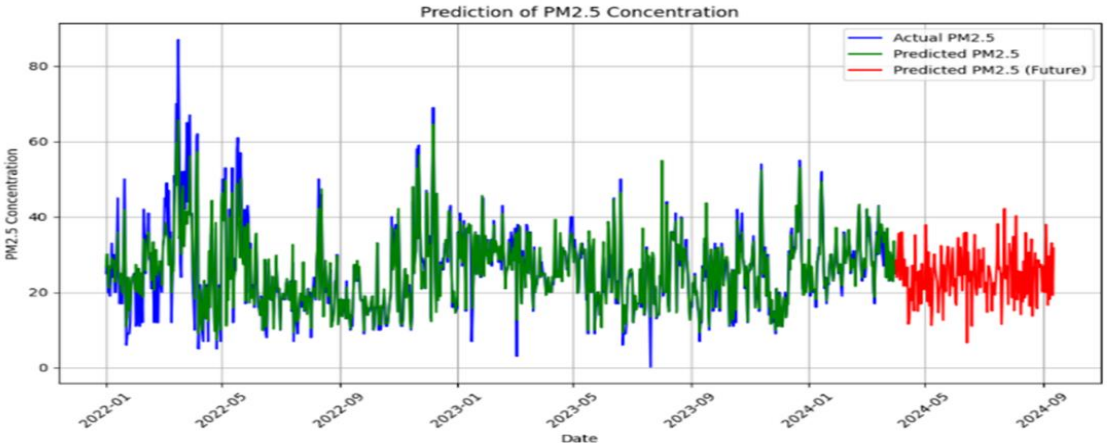

Fig.III Random Forest regression(PM2.5 prediction)

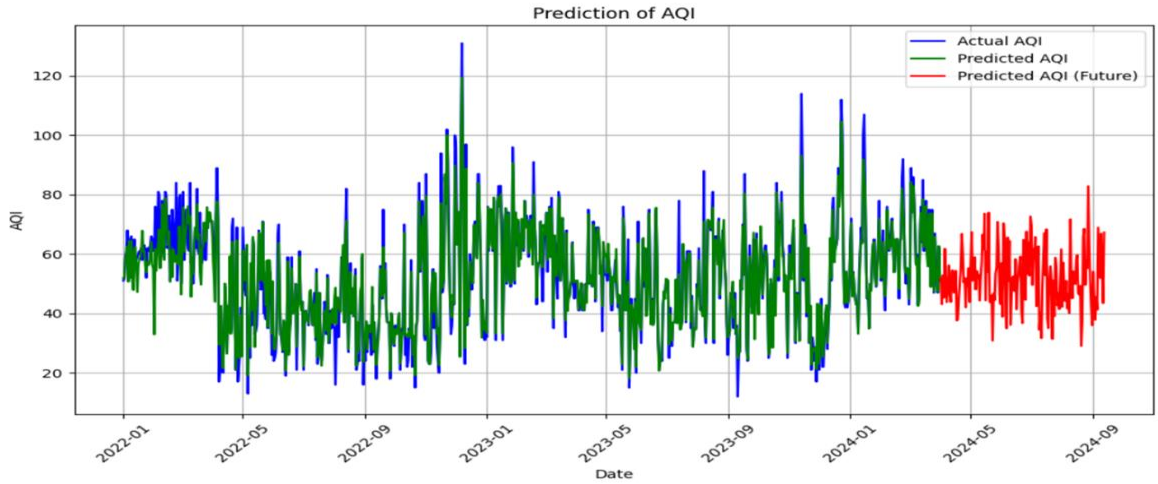

Fig.IV Random Forest regression(AQI prediction)

The R^2, RMSE, MSE, MAE value obtained for different regression analysis is tabulated. The graph is plotted for PM2.5, PM10, and AQI for the varying features and its plotted against the date. The variation between actual pollutant vs pollutant predicted using regression algorithm for the year 2022 – 2023 is plotted. Random Forest regression shows better accuracy. Among all the results obtained for PM2.5, PM10 and AQI, the Random Forest regression depicts more accuracy and minimized error. Random Forest for PM10 achieved an R^2 of 0.95916 on training and 0.82101 on testing, indicating strong predictive capability. For PM2.5, the model performed exceptionally well with an R^2 of 0.96414 on training and 0.88302 on testing, demonstrating its reliability in predicting fine particulate matter levels. An R^2 of 0.95732 on training and 0.81811 on AQI prediction shows good predictive performance, crucial for overall air quality assessment Random Forest consistently provided the best performance across all three targets (PM10, PM2.5, AQI), indicating its robustness and ability to capture complex relationships in the data.

Table.IV Random Forest regression results

Random Forest Regression								
	(TRAIN)				(TEST)			
	MSE	RMSE	MAE	R2	MSE	RMSE	MAE	R2
PM10	0.11804	0.24428	0.04515	0.95916	0.26129	0.37223	0.09643	0.82101
PM2.5	0.01611	0.12504	0.02419	0.96414	0.02934	0.16607	0.05509	0.88302
AQI	0.19502	0.24227	0.05724	0.95732	0.32508	0.39303	0.11711	0.81811

V. CONCLUSION

The ambient air quality in Madurai city is primarily influenced by PM10. In the analysis of air quality prediction for Madurai city using data from 2022-2023, Random Forest Regression emerged as the most effective modeling approach. It consistently provided high accuracy across PM10, PM2.5, and AQI metrics, demonstrating robust predictive performance with R^2 values exceeding 0.95 for training data and maintaining strong results on testing data. This indicates its strong capability to capture and generalize complex patterns within the dataset. Other models, including Decision Trees, Gradient boosting, and simpler approaches like Linear Regression, showed varying degrees of effectiveness, but none matched the predictive power of Random Forest. The superior performance of Random Forest highlights the importance of using advanced machine learning techniques for environmental data analysis, particularly in capturing non-linear relationships and interactions among variables. For Madurai city, these findings underscore the potential of data-driven approaches in air quality management. Accurate predictions of pollutants like PM10 and PM2.5 are crucial for public health interventions and policy-making. The insights gained from such models can guide timely public advisories and strategic planning to mitigate pollution effects, ultimately contributing to better environmental and public health outcomes.

ACKNOWLEDGMENT

The authors wish to acknowledge Alagappa Chettiar Government College of Engineering and Technology, Karaikudi, Tamil Nadu, India for providing lab facilities. The authors further wish to acknowledge Tamil Nadu Police Control Board, Kappalur, Madurai for providing the necessary pollutant concentration & AQI dataset.

REFERENCES

[1] S. K. Guttikunda, R. Goel, and P. Pant, "Nature of air pollution, emission sources, and management in the Indian cities," *Atmos. Environ.*, vol. 95, pp. 501–510, 2014, doi: 10.1016/j.atmosenv.2014.07.006.

[2] J. Salva, M. Vanek, M. Schwarz, M. Gajtanska, P. Tonhauzer, and A. Ďuricová, "An assessment of the on-road mobile sources contribution to particulate matter air pollution by aermod dispersion model," *Sustain.*, vol. 13, no. 22, 2021, doi: 10.3390/su132212748.

[3] A. Mukherjee and M. Agrawal, "World air

particulate matter: sources, distribution and health effects," *Environ. Chem. Lett.*, vol. 15, no. 2, pp. 283–309, 2017, doi: 10.1007/s10311-017-0611-9.

[4] A. Pandey *et al.*, "Health and economic impact of air pollution in the states of India: the Global Burden of Disease Study 2019," *Lancet Planet. Heal.*, vol. 5, no. 1, pp. e25–e38, 2021, doi: 10.1016/S2542-5196(20)30298-9.

[5] P. Pant, S. K. Guttikunda, and R. E. Peltier, "Exposure to particulate matter in India: A synthesis of findings and future directions," *Environ. Res.*, vol. 147, pp. 480–496, 2016, doi: 10.1016/j.envres.2016.03.011.

[6] P. Pant *et al.*, "Monitoring particulate matter in India: recent trends and future outlook," *Air Qual. Atmos. Heal.*, vol. 12, no. 1, pp. 45–58, 2019, doi: 10.1007/s11869-018-0629-6.

[7] Ministry of Environment Forests and Climate Change, "National Clean Air Programme (NCAP) India," *Gov. India*, 2018, [Online]. Available: http://www.moef.gov.in/sites/default/files/NCAP with annex-ilovepdf-compressed.pdf

[8] A. Rastogi, A. V. Rajan, and M. Mukherjee, "A Review of Vehicular Pollution and Control Measures in India," pp. 237–245, 2018, doi: 10.1007/978-981-10-7122-5_24.

[9] C. Khandar and S. Kosankar, "A review of vehicular pollution in urban India and its effects on human health," *Jalrb) J. Adv. Lab. Res. Biol.*, vol. 5, no. 3, pp. 54–61, 2014, [Online]. Available: https://e-journal.sospublication.co.in

[10] S. Kumari, "National Clean Air Programme , 2019 : Critical Analysis INTRODUCTION :," vol. II, no. I, pp. 1–17, 2021.

[11] P. Revell, "Reference division," *Libr. Rev.*, vol. 32, no. 1, pp. 33–44, 1983, doi: 10.1108/eb012744.

[12] CPCB, "Non attainment cities," *Cent. Pollut. Control Board*, no. 1, pp. 2–5, 2021.

[13] CPCB, "National Ambient Air Quality Status & Trends 2019," *Cent. Pollut. Control Board*, vol. 53, no. 9, pp. 1689–1699, 2020.

[14] A. Roychowdhury and A. Somvanshi, "Breathing Space: How to track and report air pollution under the National Clean Air Programme," *Cent. Sci. Environ. New Delhi*, 2020.

[15] T. Ganguly, K. L. Selvaraj, and S. K. Guttikunda, "National Clean Air Programme (NCAP) for Indian cities: Review and outlook of clean air action plans," *Atmos. Environ. X*, vol. 8, p. 100096, 2020, doi: 10.1016/j.aeaoa.2020.100096.

[16] S. K. Guttikunda, K. A. Nishadh, and P. Jawahar, "Air pollution knowledge assessments (APnA) for 20 Indian cities," *Urban Clim.*, vol. 27, no. November 2018, pp. 124–141, 2019, doi: 10.1016/j.uclim.2018.11.005.

[17] "Current Science Association Emission inventory – a preliminary approach to primary pollutants Author (s): Seshapriya Venkitasamy and B . Vijay Bhaskar Published by : Current Science Association Stable URL : https://www.jstor.org/stable/24911545 REFERENC," vol. 111, no. 11, pp. 1831–1835, 2021, doi: 10.18520/cs/vl.

[18] F. Mazzei *et al.*, "Characterization of particulate matter sources in an urban environment," *Sci. Total Environ.*, vol. 401, no. 1–3, pp. 81–89, 2008, doi: 10.1016/j.scitotenv.2008.03.008.

[19] S. Bali, "Indian Air Quality Prediction and," vol. 14, no. 11, pp. 181–186, 2019.

[20] D. Iskandaryan, F. Ramos, and S. Trilles, "Air quality prediction in smart cities using machine learning technologies based on sensor data: A review," *Appl. Sci.*, vol. 10, no. 7, 2020, doi: 10.3390/app10072401.

[21] V. Stojov, N. Koteli, and P. Lameski, "Application of machine learning and time-series analysis for air pollution prediction," no. April, pp. 2–7, 2018.

[22] G. Pandey, B. Zhang, and L. Jian, "Predicting submicron air pollution indicators: A machine learning approach," *Environ. Sci. Process. Impacts*, vol. 15, no. 5, pp. 996–1005, 2013, doi: 10.1039/c3em30890a.

[23] X. Xi *et al.*, "A comprehensive evaluation of air pollution prediction improvement by a machine learning method," *10th IEEE Int. Conf. Serv. Oper. Logist. Informatics, SOLI 2015 - conjunction with ICT4ALL 2015*, pp. 176–181, 2015, doi:

10.1109/SOLI.2015.7367615.

[24] M. R. Delavar *et al.*, "A novel method for improving air pollution prediction based on machine learning approaches: A case study applied to the capital city of Tehran," *ISPRS Int. J. Geo-Information*, vol. 8, no. 2, 2019, doi: 10.3390/ijgi8020099.

[25] S. K. Guttikunda and G. Calori, "A GIS based emissions inventory at 1 km × 1 km spatial resolution for air pollution analysis in Delhi, India," *Atmos. Environ.*, vol. 67, pp. 101–111, 2013, doi: 10.1016/j.atmosenv.2012.10.040.

[26] M. Castelli, F. M. Clemente, A. Popovič, S. Silva, and L. Vanneschi, "A Machine Learning Approach to Predict Air Quality in California," *Complexity*, vol. 2020, no. Ml, 2020, doi: 10.1155/2020/8049504.

[27] C. Bellinger, M. S. Mohomed Jabbar, O. Zaïane, and A. Osornio-Vargas, "A systematic review of data mining and machine learning for air pollution epidemiology," *BMC Public Health*, vol. 17, no. 1, pp. 1–19, 2017, doi: 10.1186/s12889-017-4914-3.

[28] C. Srivastava, S. Singh, and A. P. Singh, "Estimation of air pollution in Delhi using machine learning techniques," *2018 Int. Conf. Comput. Power Commun. Technol. GUCON 2018*, pp. 304–309, 2019, doi: 10.1109/GUCON.2018.8675022.

[29] A. C R, C. R. Deshmukh, N. D K, P. Gandhi, and V. astu, "Detection and Prediction of Air Pollution using Machine Learning Models," *Int. J. Eng. Trends Technol.*, vol. 59, no. 4, pp. 204–207, 2018, doi: 10.14445/22315381/ijett-v59p238.

[30] Doreswamy, K. S. Harishkumar, Y. Km, and I. Gad, "Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models," *Procedia Comput. Sci.*, vol. 171, no. 2019, pp. 2057–2066, 2020, doi: 10.1016/j.procs.2020.04.221.

[31] S. Ameer *et al.*, "Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities," *IEEE Access*, vol. 7, pp. 128325–128338, 2019, doi: 10.1109/ACCESS.2019.2925082.