Machine Learning-Based Air Quality Monitoring in Indian Metropolitan Cities: A Comparative Study

Khushbu Chauhan[1], Kruti Sutaria[2]
Dept. of Computer Science and Engineering, Parul University, Vadodara, 391760, India[1]
Dept. of CSE Cyber Security, Parul University, Vadodara, 391760, India[2]

***Abstract*—Pure and clean air is essential to make the ecosystem healthy. Air pollution is becoming a critical global concern for both the environment and human health. Presence of harmful pollutants such as $PM_{2.5}$, PM_{10}, CO_2, NO_2, SO_2, and O_3 continuously degrades air quality and influences climatic conditions. This study aims to present a comprehensive air quality monitoring between traditional and advanced ensemble-based machine learning models. To monitor air quality, data collected from major metropolitan cities of India from 2015 to 2023 (Three phases- Pre-COVID, during COVID-19, and post-COVID). After pre-processing the data, a baseline supervised machine learning method, Support Vector Machine (SVM), was applied for ease of implementation. Later, to train weak learner features, ensemble-based machine learning techniques include Gradient Boosting Machine (GBM) and Extreme Gradient Boosting Machine (XGBM), evaluated to get better prediction analysis. The systematic analysis is inspected using different performance parameters: R^2, Mean Squared Error, Root Mean Squared Error, and Mean Absolute Error. The outcome indicates XGBM achieves superior predictive accuracy and robustness across most cities and time periods, and achieves better variability in spatial and temporal features in performance. The key findings highlight the importance of location-based specific modelling strategies and demonstrate the potential of ensemble learning models for reliable urban air quality monitoring.**

***Keywords*—AQI; COVID; SVM; Gradient Boosting Machine (GBM); Extreme Gradient Boosting (XGBoost)**

I. Introduction

Rapid urbanisation and industrialisation, vehicular emissions, and deforestation in the 21st century ruin the entire environment and nature [1]. Major air pollutants, such as particulate matter ($PM_{2.5}$ and PM_{10}), nitrogen dioxide (NO_2), sulphur dioxide (SO_2), carbon monoxide (CO), carbon dioxide (CO_2), and ozone (O_3), [2] cause hazardous impacts on human health. The Air Quality Index (AQI) is mainly used to measure and monitor air quality continuously based on different pollution levels, categorizing it from good (0-50) to hazardous (above 300) using a colour scale ranging from green to maroon [3]-[4]. According to the World Health Organization (WHO), urban air pollution increased approximately 56.15% in 2020, which may rise to 68% by 2050. This concerning increase in pollution poses significant threats not only to human well-being but also to the entire ecosystem [5].

Traditional statistical and machine learning models, such as linear regression, decision trees, random forest, and support vector machines (SVMs), have been widely applied for air quality prediction due to their conceptual simplicity and relatively low computational cost. However, these methods often struggle to capture the non-linear, non-stationary, and highly heterogeneous nature of urban air pollution, particularly when applied across diverse climatic zones or during periods of abrupt socio-economic change. Moreover, many existing studies emphasise predictive accuracy as the primary evaluation criterion, with limited attention to model robustness, uncertainty, and interpretability in the context of air quality science and decision-making.

Recent advances in ensemble machine learning techniques, particularly gradient boosting–based methods, have demonstrated strong potential for improving air quality forecasting by combining multiple weak learners to capture complex feature interactions. Models such as Gradient Boosting Machine (GBM) and Extreme Gradient Boosting Machine (XGBoost) offer enhanced generalisation capabilities, built-in regularisation, and robustness to noisy environmental data. Despite these advantages, their application in air quality research has largely focused on short-term forecasting or single-city case studies, with limited exploration of spatial heterogeneity, long-term temporal dynamics, and real-world disruptions such as the COVID-19 pandemic. This work focuses on air quality monitoring by closely observing the Air Quality Index (AQI) of major air pollutants: PM2.5, PM_{10}, O_3, CO, NO_2, and SO_2 from collected data on Indian metropolitan cities, including Ahmedabad, Bengaluru, Chennai, Delhi, Kolkata, and Mumbai. These cities represent diverse climatic zones and industrial profiles, which makes them more ideal for air pollution analysis. To capture complex, non-linear, and weak learner features, ensemble machine learning techniques such as Gradient Boosting Machine and Extreme Gradient Boosting Machine were introduced. The study focuses on three phases: Pre-COVID, during the COVID-19 pandemic, and Post-COVID recovery periods, which provide a clear view of pollution trends under economic conditions. By comparing baseline machine learning models with advanced models to improve the prediction accuracy by calculating different performance parameters, R^2 score, Mean Squared Error, Mean Absolute Error, and Root Mean Squared Error.

Despite significant advancements in machine learning-based air quality monitoring, the existing studies largely focus on short-term temporal prediction and on single-city analysis, with limited attention on long-term temporal variability, spatial heterogeneity, and model robustness, such as COVID-

19 pandemic disruptions. These create a gap in understanding how different models perform across diverse urban environments. Therefore, a strong requirement for a systematic evaluation and analysis of machine learning methods for air quality monitoring across multiple cities with different temporal phases.

The key contribution of this work lies in moving beyond accuracy-centric evaluation toward a more comprehensive understanding of how machine learning models reflect underlying air pollution dynamics in heterogeneous urban environments. By identifying cities and periods where models perform reliably and those where prediction uncertainty and overfitting emerge, the findings provide actionable insights for air quality management, early warning systems, and policy-driven decision-making. This integrated perspective supports the development of adaptive, location-specific air quality monitoring frameworks suitable for complex and rapidly evolving urban settings. The key objective of the research study is to evaluate and analyse long-term air quality trends and to compare the performance between traditional and ensemble machine learning methods.

II. Related Work

The study proposed by Pan et al. [6] was a cluster-based under-sampling strategy combined with a transformer model to improve the prediction of extreme PM2.5 concentration. They optimized model performance for a high-pollution scenario and developed data augmentation for air quality forecasting. Due to the rarity of training datasets, traditional forecasting approaches perform poorly on high-concentration episodes. Alrasheedi et al. [7] emphasized the use of ensemble and hybrid machine learning models to enhance robustness and handle spatial heterogeneity for accurate air quality prediction. The authors used GBM and XGBM strategies to improve the performance of air quality prediction. The review persists challenges of high computational costs, limited model generalizability across regions, and data quality issues. Muhammad et al. [8]; Kumari et al. [9] integrated a long-term climate assessment framework that combines the Mann-Kendall trends with machine learning models to adopt a climate-smart agriculture system in Nigeria. Satish et al. [10] demonstrated a comparative survey on four major sources: industrial activities, vehicular traffic, natural phenomena, and combustion processes to evaluate machine learning-based air pollution prediction. The reviewed studies reveal challenges such as the complexity and variability of pollution sources and the difficulty in accurately modelling emissions from diverse and dynamic processes. A linear regression-based machine learning model to forecast AQI using an AI-driven, data-centric air quality prediction framework proposed by Waghmare et al. [11] for Pune city.

Ansari et al. [12] proposed a comprehensive machine learning-based framework for hourly AQI prediction and demonstrated that Extreme Boost (XGBoost) performed with superior accuracy and computational efficiency in urban air quality forecasting. Dawar et al. [13] proposed a Lasso regressor-based machine learning model that can accurately predict air quality from the Himalayan city of Dehradun. It supported proactive air pollution mitigation aligned with SDGs 3 and 11. Due to limited location-specific AQI prediction studies for Himalayan towns, the influence of complex geographical and climatic conditions on air quality. The authors concluded with the need for reliable, high-accuracy models to enable timely and effective urban air quality management. T. Desai et al. [14] used SVM and KNN-based machine learning models to combine multiple algorithms to make hybrid models to achieve better forecast accuracy, especially when trained on large datasets. They faced difficulty in achieving consistent AQI prediction accuracy across diverse urban industrial environments. Sharma et al. [15]; Ahmad et al. [16] proposed two different frameworks, ARIMAX/SARIMAX-ML and a hybrid BiGRU-1DCNN framework, to integrate temporal and spatial feature learning to accurately predict particulate matter levels across multiple monitoring stations in Delhi. The BiGRU-1DCNN model demonstrates strong predictive performance by effectively capturing spatiotemporal patterns. However, the ARIMAX/SARIMAX-SVM framework provides more interpretable outcomes by explicitly analysing trends and incorporating exogenous factors, which enhances its usefulness for policy formulation. J. Saini et al. [17] presented a systematic review of various machine learning models to predict indoor air quality by identifying Random Forest as the most commonly adopted model, and PM2.5 was identified as a key indicator for critical indoor air conditions. Difficulties in timely forecasting of hazardous indoor air scenarios and a lack of training data face issues in forecasting indoor air quality. Naveed et al. [18] proposed an innovative digital twin–based framework for AQI forecasting in smart cities, integrating deep learning models with a 3D urban digital twin to enable accurate prediction, visualization, and real-time monitoring of air quality. Challenges such as limited integration of AI models with digital twin platforms, and difficulty in real-time visualization and decision support issues have been identified for effective forecasting.

Although extensive research has applied machine learning and ensemble-based techniques for air quality prediction, several limitations remain in the current body of literature. Many existing studies focus on short-term forecasting, single-city case studies, or isolated pollutants, which restricts their ability to capture long-term temporal dynamics and spatial heterogeneity across diverse urban environments. Moreover, while advanced models such as XGBoost, hybrid deep learning architectures, and spatiotemporal frameworks have demonstrated improved predictive accuracy, limited attention has been given to model robustness, uncertainty, and performance degradation under abrupt socio-economic disruptions, such as the COVID-19 pandemic. Most studies primarily evaluate models based on numerical accuracy metrics, with minimal interpretation of how model behaviour reflects underlying pollution complexity or emission stability.

Furthermore, most existing works emphasise accuracy metrics without providing insights. In contrast, the present study extends existing research by conducting a multi-city, long-term analysis across major Indian metropolitan regions representing diverse climatic conditions and emission profiles. By systematically evaluating traditional and ensemble machine learning models across pre-COVID, during COVID,

and post-COVID phases, this work provides insights into spatial variability, temporal stability, and overfitting behaviour under distributional shifts. Furthermore, the study emphasises the practical relevance of model performance for urban air quality management, highlighting where data-driven approaches are reliable and where adaptive or hybrid strategies may be required. This comparative and interpretive perspective positions the work beyond accuracy-centric forecasting and contributes to a more policy-relevant understanding of machine-learning–based air quality monitoring.

III. Materials and Methods

A. Data Collection

Table I shows datasets collected from major metropolitan cities of India, including Ahmedabad, Bengaluru, Chennai, Delhi, Kolkata, and Mumbai. From Mumbai, the data was collected from different highly polluted locations, including Borivali West, BKC, Chakala Andheri East, CSMT Airport, Deonar, Bhandup West, Kurla, Malad West, Mulund West, Vile Parle West, Colaba, Sion, and Worli. These cities were selected to represent a diverse range of climatic zones and to exhibit rapid population growth and pollution sources, including traffic emissions, industrial activities, stubble burning, and deforestation, making them suitable for a comprehensive analysis of air quality. The air quality data were obtained from publicly accessible resources, the Central Pollution Control Board (CPCB) and the State Pollution Control Board (SPCB). The dataset covers the years from January 2015 to July 2023 and consists of hourly observations, including approximately 78,000 records with 24 different attributes, making it a large multivariate time-series dataset. It includes time indicators (From Date and To Date), key air pollutants such as $PM_{2.5}$ ($\mu g\, m^{-3}$), PM_{10} ($\mu g\, m^{-3}$), NO_2 ($\mu g\, m^{-3}$), SO_2 ($\mu g\, m^{-3}$), CO ($mg\, m^{-3}$),, Ozone ($\mu g\, m^{-3}$), and selected Volatile Organic Compounds (VOCs) components like Xylene ($\mu g\, m^{-3}$), along with meteorological parameters including Temperature (^{0}C), Relative Humidity (%), Wind Speed ($m\, s^{-1}$) and Wind Direction (^{0}), Barometric Pressure (hPa), Solar Radiation ($W\, m^{-2}$), and Rainfall (mm). Data validation procedures included consistency checks, removal of unrealistic and outlier values in accordance with the CPCB and SPCB data quality guidelines. As is typical of real-world environmental monitoring data, some attributes contain missing or irregular values. This highlights the need for careful pre-processing before applying machine-learning models.

B. Data Pre-processing

Data Pre-processing is used to enhance data quality for better model performance. From the above section, datasets were collected from 2015 to 2019 on an hourly basis by different key pollutants along with measurements. Techniques such as data cleaning and outlier removal are important before applying models. To handle missing values and outlier removal, various pre-processing approaches were used to ensure data integrity and model performance. A variety of non-arbitrary values without adding any arbitrary data were used to fill missing values [19].

TABLE I. Data Collection from Major Metropolitan Cities

Sr. NO.	Indian metropolitan cities	Sr. NO.	Mumbai Locations
1	Ahmedabad	6.1	Borivali East
2	Bengaluru	6.2	BKC
3	Chennai	6.3	Chakala Andheri East
4	Delhi	6.4	CSMT Airport
5	Kolkata	6.5	Deonar
6	Mumbai	6.6	Bhandup West
		6.7	Kurla
		6.8	Malad West
		6.9	Mulund West
		6.10	Vile Parle West
		6.11	Colaba
		6.12	Sion
		6.13	Worli

$$x_m = x_P + \left(\frac{x_n - x_P}{t_n - t_p}\right) t_m - t_p \quad (1)$$

where, x_m is used to fill missing values. X_p and Xn are the previous and next index, respectively [see Eq. (1)].

Missing values were treated using a time-series interpolation method. Linear interpolation, which works particularly well with time-series datasets, was used to estimate missing values between consecutive observations. Furthermore, polynomial interpolation was used for more intricate data patterns, fitting higher-order curves to more precisely monitor missing inputs.

Attributes such as NH_3, Xylene, Wind Speed, Wind Direction, and Rainfall (RF) were removed due to a high proportion of missing values to maintain data quality and minimise noise. The pre-processed datasets concentrate on the most reliable and pertinent air quality indicators. For short gaps occurring at the beginning or end of a time series, forward fill (FFill) and backward fill (BFill) were used on datasets for quick processing. In order to fill the gaps, Backwards Fill uses the next known value, whereas Forward Fill uses the more recent known value ahead to address missing elements [see Eq. (2) and Eq. (3)]:

$$Xt = Xt - 1 \text{ (Forward Fill)} \quad (2)$$

$$Xt = Xt + 1 \text{ (Backward Fill)} \quad (3)$$

These techniques store the time-series related data structure. Furthermore, for regression imputation methods based on models were used to estimate missing values, K-Nearest Neighbours (KNN) imputation evaluates missing values based on the similarity among neighbouring data points.

Outlier detection and treatment were performed to mitigate the influence of extreme values that occurred in the statistical analysis. To remove outliers, two primary approaches were used, which have the potential to skew statistical research and impair model accuracy. First, outliers were found by

computing boundary values based on standard deviations (values exceeding ± 3) using the Z-score approach [20]. Plotting distributions, identifying borders, identifying outliers, and applying trimming or capping were all included in the outlier-removing process. To reduce skewness while preserving the data's range, capping replaced the present threshold values with extreme values. For verification of changes, statistical conclusions were produced using the describe () function. Moreover, a percentile-based Winsorization approach was used to manage the skewed distribution commonly observed in pollutant concentration data. The 1st and 99th percentile thresholds were selected to cap extreme observations. Winsorization minimises information loss and maintains dataset size, which is particularly useful for time-series modelling. Across major key pollutant features such as $PM_{2.5}$, PM_{10}, SO_2, NO_2, and CO, approximately 7-10 % observations per feature were identified as outliers and subsequently capped. Visual inspection by box plots and distribution plots, along with observations before and after outlier removal, confirmed that these procedures reduced skewness and variance without altering temporal patterns, which were used to evaluate the data spread of selected features [21].

The final pre-processed dataset retained the most reliable air quality and meteorological variables, with improved stable distribution and noise reduction, providing a robust base for subsequent machine learning models development. The cleaned dataset was divided into training (80%) and testing (20%) sets using a time-aware split to preserve temporal dependencies. Feature sets include key pollutants $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, O_3 as input features, and selected meteorological parameters include temperature, relative humidity. AQI is treated as a target feature. Since AQI is derived from pollutant concentrations, it was ensured that AQI was not used as an input feature when predicting AQI to avoid data leakage. Additionally, a validation set (10% training data) was used for hyperparameter tuning. Hyperparameters for machine learning algorithms represented in the section below, specifically for SVM, GBM, and XGBoost, were optimised using the grid search technique.

C. Machine Learning Models

Support Vector Machine (SVM) has emerged as a reliable data-driven approach for air quality prediction, particularly in situations where conventional physical and statistical models face limitations [22]. SVM, and its regression variant Support Vector Regressor (SVR), are well-suited for handling the non-linear and non-stationary nature of atmospheric pollution data. SVM requires lower computational effort for short-term forecasting applications to get superior accuracy. This will make the SVM model more practical and provide stable responses to variations in training data. SVM provides a balanced combination of accuracy, robustness, and computational efficiency, establishing it as a suitable and dependable tool for air quality monitoring across different climatic and urban environments. Parameters such as kernel function are tuned based on validation performance [23]-[24].

An Ensemble learning technique, such as a Gradient Boosting Machine (GBM) that builds weak learners in sequence, specifically decision trees. With optimization of gradient descent, each new model needs to correct the mistakes of its predecessors with exceptional prediction skills and flexibility in handling different loss functions. Gradient Boosting Machine exhibits high efficiency for both applications of regression and classification [25]. XGBoost, LightGBM, and CatBoost libraries offer well-optimised implementations of GBM. Hyperparameters such as the learning rate, tree depth, and number of boosting rounds must be aligned to achieve the optimum results. No single gradient boosting algorithm is universally superior; instead, the choice depends on the application requirements. CatBoost is preferable when predictive accuracy is critical [26], LightGBM is ideal for large datasets requiring fast training, and XGBoost provides a well-balanced option with strong performance and scalability. The XGBRegressor [27] library is specifically used as a regression with gradient boosting implementation, created to minimise a specified loss function. It builds an ensemble of decision trees one after the other, with each new tree concentrating on reducing the residual errors of its predecessors. To compare traditional boosting techniques with the ensemble model, XGBRegressor's ability to integrate both L1 (alpha) and L2 (lambda) regularization is a noteworthy advantage, as it reduces overfitting and enhances generalization. It can be fine-tuning the complexity of a model for training speed and predicted accuracy by adjusting its hyperparameters, including learning_rate, n_estimators, max_depth, and subsample. The models were trained on earlier time periods and tested on later periods to evaluate performance under temporal distribution shifts. Hyperparameters learning rate, number of estimators, and maximum tree depth (GBM and XGBoost) were tuned based on validation performance.

$$£^{(t)} = \sum_{i=1}^{n} l(\tilde{y}i, \tilde{y}_{i)}^{t} + \sum_{k=1}^{t} \Omega f_k \quad (4)$$

$$\Omega(f) = rT + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \quad (5)$$

$$\mathcal{L}^t \approx \sum_{i=1}^{n} [g_i f_i(x_{i)} + \frac{1}{2} h_i f_i(x_i^2)] + \Omega(f_t) \quad (6)$$

$$\omega_j = \frac{\Sigma_{i \in Ijgi}}{\sum_{i \in Ij} h_i + \lambda} \quad (7)$$

$$y_i = \sum_{k=1}^{T} f_k(x_i) \quad (8)$$

Eq. (4) to Eq. (8) collectively define the regularized boosting framework adopted in this study. These formulations govern loss minimization, tree complexity control, and incremental error correction, forming the basis of the XGBM implementation used for multi-pollutant prediction [28]. $l(\tilde{y}i, \tilde{y}_{i)}^{t}$ = Loss function (e.g., Mean Squared Error and Mean Absolute Error calculation for the difference between actual and predicted pollutant concentration) and Ωf_k as a Regularisation term (Overfitting in transfer learning). $\Omega(f)$ for preventing overfitting, a regularisation term is added to the main loss function. The equation illustrates the estimated loss function at boosting step t, employing a second-order Taylor expansion. ω_j Optimal weight assigned to leaf node j. y_i Final predicted pollutant concentration for the data point i. T: Total number of boosting iterations. $f_k(x_i)$ Output of the kth

weak learner of a model. Each weak learner comprehensively resolves the residual errors from previous trees, leading to improved pollutant prediction accuracy. By aggregating losses across multiple pollutants, XGBM ensures balanced forecasting and prevents domination by a single pollutant. Additionally, an iterative learning of XGBM makes it more suitable for robust air quality monitoring. The XGB Regressor optimises a regularized objective that includes both a loss component and a penalty for complexity in order to construct additive decision trees. By utilising gradient information (first-order) along with Hessian information (second-order), it effectively modifies tree structures while managing overfitting through L1/L2 regularization.

IV. Results and Discussion

A. Advanced Machine Learning Models GBM and XGBM

Ensemble learning techniques, particularly XGBRegressor, offer a strong and logical solution to train weak learner features in terms of air quality monitoring. The developed improvements on collected data indicated that the boosting framework effectively captured complex non-linear relationships between pollutants and meteorological variables, leading to improved prediction accuracy for multiple pollutants such as $PM_{2.5}$, PM_{10}, CO, SO_2, and NO_2. The combination training of L1 and L2 regularization within the objective function-controlled model complexity, which reduced the overfitting issue and gave stable performance across major metro cities and three different phases, includes pre-COVID, COVID-19, and post-COVID. An aggregated loss across major pollutants allowed the model to maintain balanced forecasting without bias toward a single pollutant, to support more effective evaluation of air quality. The model's performance was evaluated using MSE, MAE, R^2, and RMSE, as defined in Eq. (9) to Eq. (12):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (9)$$

$$\text{MAE} = \frac{1}{n}\Sigma|y_i - \hat{y}_i| \quad (10)$$

$$\text{Coefficients of determination } (R^2 \text{ score}) = 1 - \frac{\Sigma(y_i - \bar{y})^2}{\Sigma[y_j - \bar{y}]^2} \quad (11)$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{\eta}(y_i - \hat{y}_i^2)} \quad (12)$$

The graph presented in Fig. 1 showed R^2 score comparisons of GBM and XGBM models across five major metropolitan cities, Bengaluru, Chennai, Delhi, Kolkata, and Ahmedabad, within time periods pre-COVID, COVID-19, and post-COVID. Bengaluru consistently performed best across all phases, while Kolkata and Ahmedabad achieved lower prediction accuracy in the pre-COVID and during COVID phases, with some post-COVID improvement. During COVID, it showed sharp variations across cities, particularly XGBM showed exceptionally high R^2 score in Chennai and Delhi, even exceeding unity in Chennai, which indicated possible overfitting due to limited pandemic data. XGBM generally outperformed GBM, particularly in the post-COVID period, with notable gains in cities like Delhi and Chennai. Overall, model performance varies across cities and timeframes due to different pollution dynamics, with XGBM proving more robust, especially after COVID.

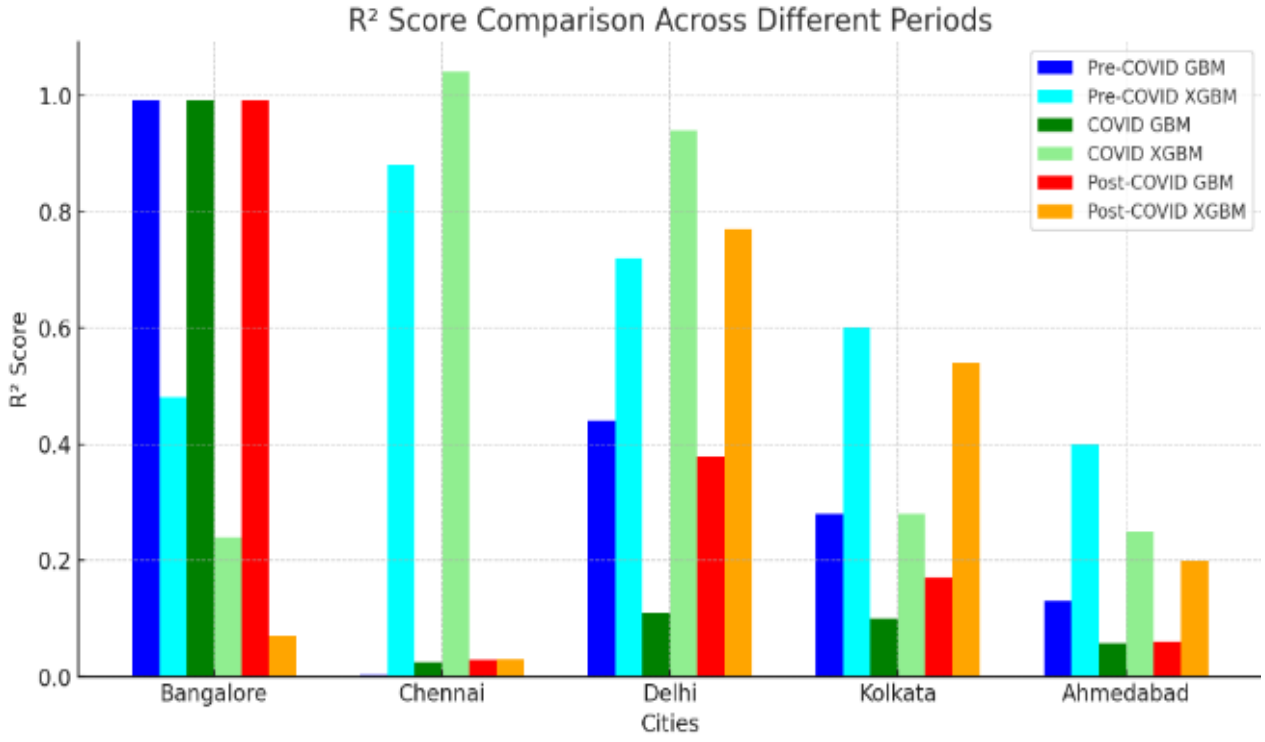

Fig. 1. Comparison of R^2 scores for GBM and XGBM models from metropolitan cities in the time of Pre-COVID, COVID-19, and Post-COVID.

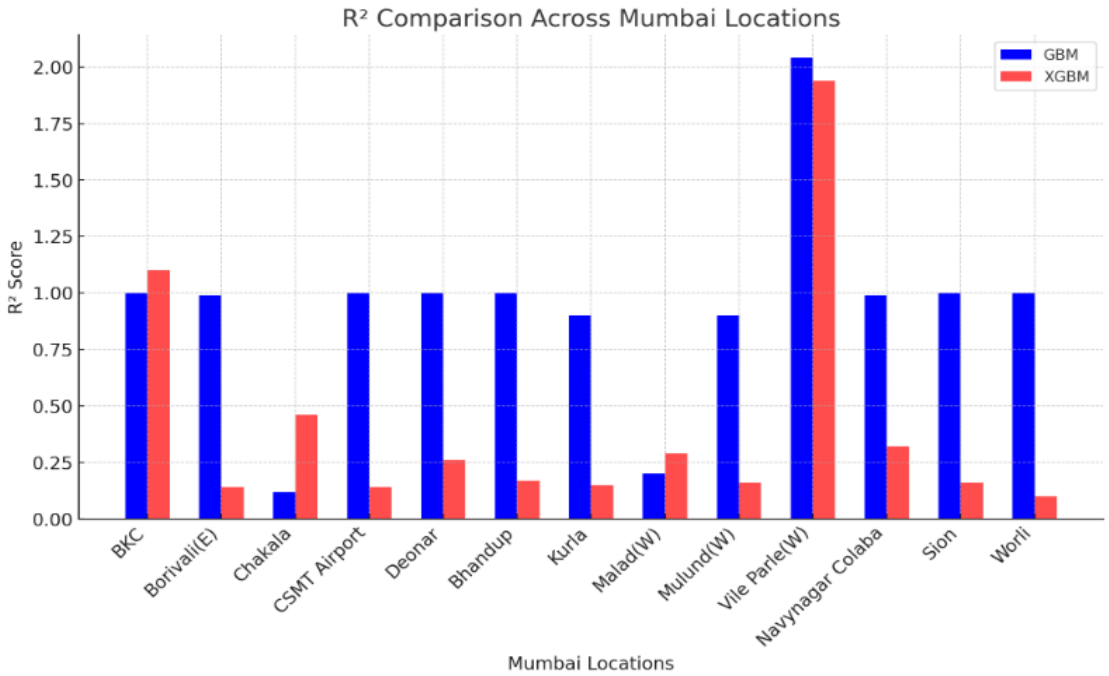

Fig. 2. R^2 and MSE values of GBM and XGBM models across various locations of Mumbai city.

The graph shown in Fig. 2 of different regions of Mumbai city compares the R^2 analysis for GBM and XGBM. The GBM model consistently demonstrated superior R^2 scores when compared to XGBM, signifying better or similar performance. Locations such as Chakala, Malad(W), and Kurla showed relatively lower R^2 scores between 0.25 and 0.5 for both models, particularly for XGBM, which may reflect higher variability in emissions, traffic influence, or data noise in these areas. Notably, regions such as BKC and Borivali(E) highlighted better R^2 scores, nearly around 1 for both models, while Vile Parle(W) achieved particularly out-of-the-ordinary results, recording R^2 scores greater than 2 for both, which showed overfitting challenges in all temporal phases. Conversely, locations like Chakala, Bhandup, and Worli revealed considerably lower scores for XGBM, indicating a weak model fit. It concludes that GBM surpasses or equals XGBM in predictive accuracy across the majority of locations.

GBM provided more consistent and stable performance across Mumbai locations, while XGBoost's performance was more location-dependent, excelling in some areas but underperforming in others. Specifically, the locations Vile Parle West and Bandra Kurla Complex (BKC) indicate an overfitting and calculation issue. This indicated that spatial heterogeneity played an important role in model performance and that model selection may need to be location-specific for urban air quality forecasting.

B. Comparison between the Existing Performance of SVM and the Model Performance

The two line graphs, in Fig. 3, illustrate the performance trends of Gradient Boosting Machine (GBM) and Support Vector Machine (SVM) models from 2015 to 2019, along with their cumulative totals. In the first graph [Fig. 3(a)], which presented R^2 values, the SVM model (represented by the red dashed line) consistently surpassed the GBM model (indicated by the blue solid line) throughout all years, with R^2 values ranging from approximately 0.76 to 0.81, signifying a stronger model fit.

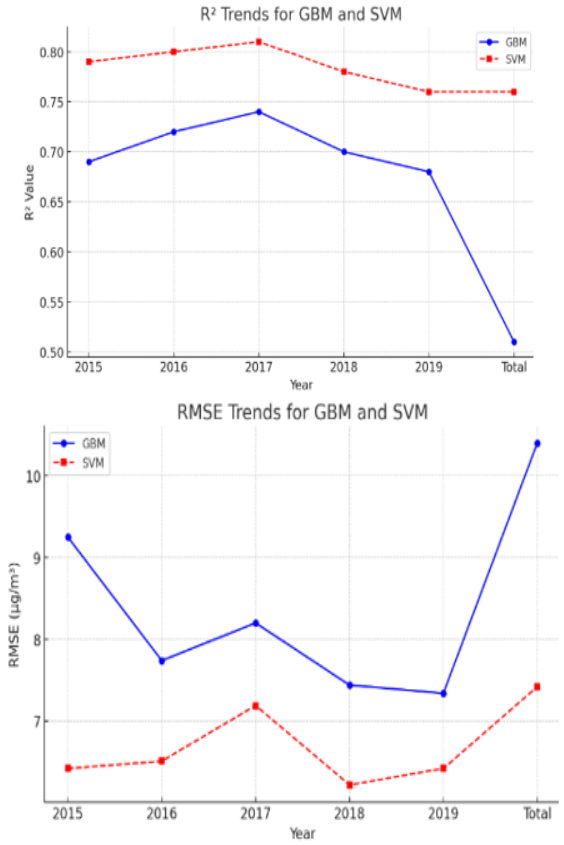

Fig. 3. (a) R^2 trend for GBM and SVM, (b) RMSE trends of GBM and SVM in 2015-2019.

The highlighted R^2 score of the GBM model presented lower R^2 values, peaking at around 0.74 in 2017 and showing a significant decline of roughly 0.51 throughout 2019, implying that it may lack consistency overall. Fig. 3(b) shows RMSE (Root Mean Squared Error) results with GBM and SVM. The outcomes from SVM continue to perform better, consistently achieving lower RMSE values, which reflect lower prediction errors throughout all years from 2015 to 2019. The GBM, on the other hand, demonstrated higher RMSE, with the error fluctuating and escalating sharply in the total metric, exceeding 10 μg m^{-3}.

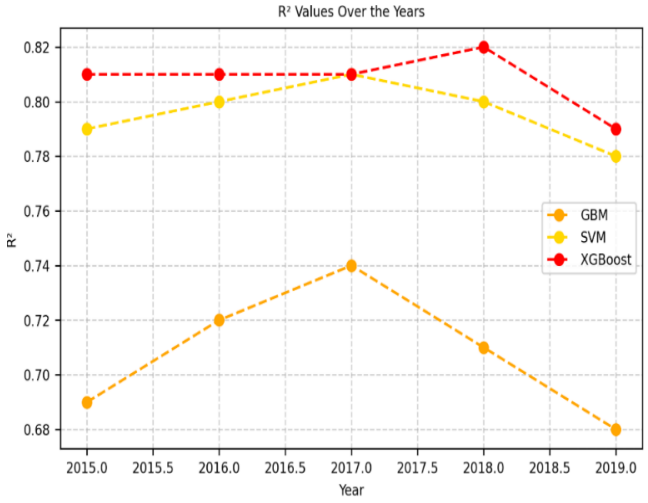

Fig. 4. Year-wise comparison of R^2 score and RMSE performance from 2015 to 2019 of GBM, SVM, and XGBM.

Fig. 4 highlights model performance trends comparison between existing results and outcomes from Indian metro cities over the years by a line plot of R^2 (coefficient of determination) values from 2015 to 2019 for three machine-learning models, GBM, SVM, and XGBoost, used for prediction tasks. XGBoost overall consistently achieved the highest R^2 values across all years, indicating the strongest explanatory power and best predictive performance among the three models. Its performance remained relatively stable, peaking around 0.82 in 2018, followed by a slight decline in 2019. SVM showed moderate and fairly stable performance, with R^2 values generally ranging between 0.78 and 0.81. It showed improvement from 2015 to 2017, and highlighted a slight decrease afterwards. In contrast, GBM listed the lowest R^2 values throughout that period, starting nearly at 0.69 in 2015, improving to about 0.74 in 2017, and then declining again by 2019.

The graph highlighted two key insights: first, XGBoost outperformed both GBM and SVM in terms of prediction accuracy and model consistency, and second, all models experienced a slight performance drop in 2019. Overall, the plot supported the conclusion that XGBoost provided more reliable and robust predictions over time compared to the other approaches. XGBM typically surpassed GBM in all three phases, especially during the pre-COVID and COVID-19 phases. However, the performance of a model in the post-COVID period declined noticeably across various cities. This may improve with possible alterations in data patterns or difficulties in model generalization following the pandemic [29].

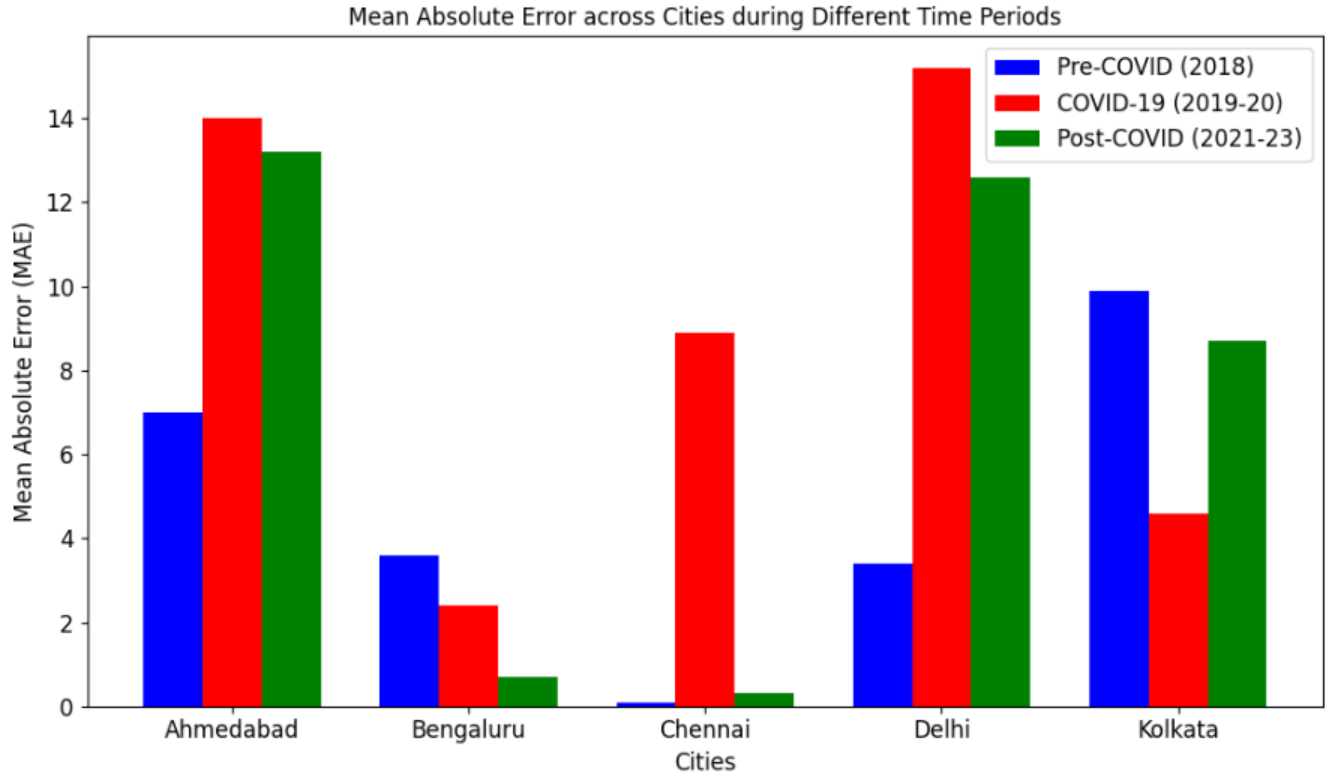

Fig. 5. Mean absolute error of air quality monitoring across five indian metropolitan cities (Ahmedabad, Bengaluru, Chennai, Delhi, and Kolkata) during three different periods: Pre-COVID (2018), During COVID-19 (2019-20), and Post-COVID (2021-23).

Fig. 5 highlights that in Ahmedabad, the Pre-COVID MAE was approximately 6.9, which surged to about 14 during the COVID-19 period and slightly declined to 13.2 in the Post-COVID phase [30]. This indicated a significant increase in prediction errors during the pandemic, with only a small improvement afterwards. In contrast, Bengaluru indicated a steady improvement with MAE values dropping from approximately 3.5 Pre-COVID to 2.4 during COVID-19, and further down to 0.8 in Post-COVID. This showed a superior result in all three phases in Bengaluru. Chennai showed variations, with a very low MAE of about 0.1 in Pre-COVID, a sharp increase to approximately 9 during COVID-19, and a return to around 0.2 in the Post-COVID phase. This pattern highlights temporary disruption during the pandemic, possibly due to sudden changes or anomalies in the data. Delhi, however, saw a steep increase from 3.5 in Pre-COVID to 15 during COVID-19, with only a slight reduction to 12.5 in Post-COVID. The persistently high MAE due to very high AQI during the winter season implies ongoing challenges in modelling or underlying unpredictability. MAE performance in Kolkata dropped notably during COVID-19 from around 9.9, pre-COVID, to 4.7, and increased again to 8.7 in post-COVID, indicating that while the pandemic might have temporarily simplified patterns [31], complexity resumed in the following years. Ahmedabad and Delhi show high values in Post COVID, while Bengaluru and Chennai show recovery.

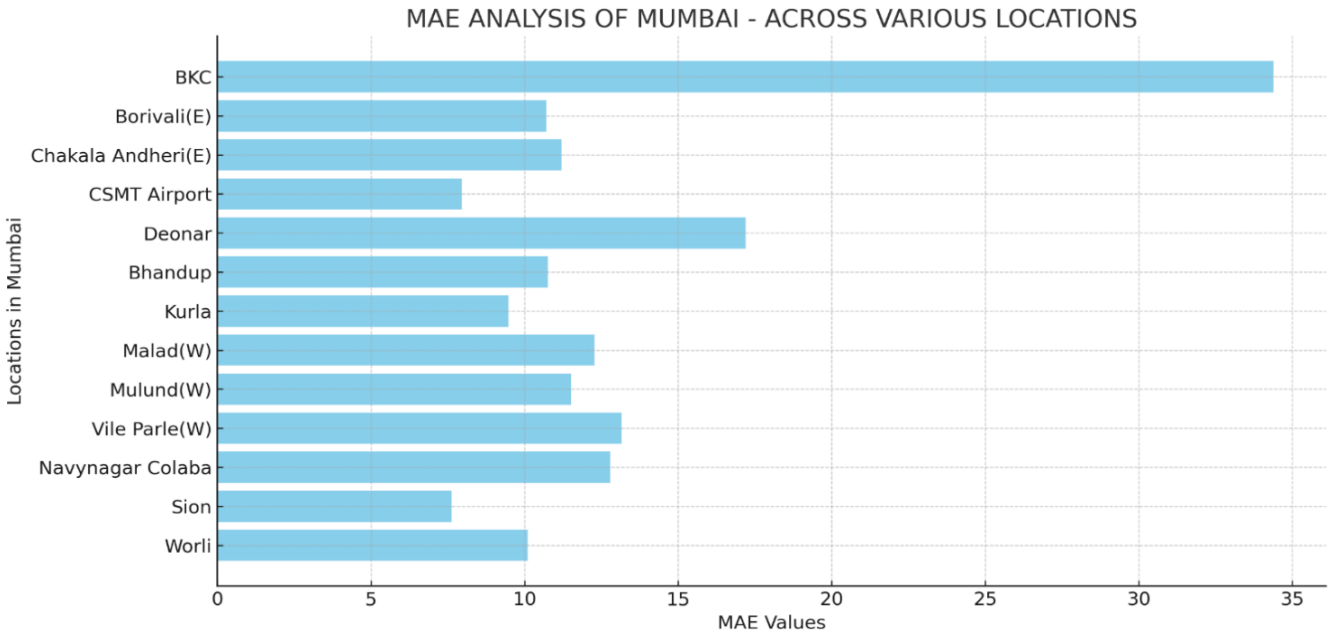

Fig. 6. Mean absolute error of air quality monitoring across thirteen locations of Mumbai city during three different periods: Pre-COVID (2018), During COVID-19 (2019-20), and Post-COVID (2021-23).

TABLE II. COMPARISON OF EXISTING AIR QUALITY PREDICTION STUDIES WITH THE PROPOSED WORK

Aspect	Existing studies	Present study
Geographic Scope	Mostly single-city or limited regional studies	Multiple Indian metropolitan cities (Ahmedabad, Bengaluru, Chennai, Delhi, Kolkata, Mumbai)
Temporal Coverage	Short-term or limited historical data	Long-term data (2015–2023)
Modelling focus	Prediction accuracy-centric	Accuracy, robustness, and spatial–temporal variability
Algorithms	ML, ensemble, and hybrid DL models	SVM, GBM, and XGBoost with comparative evaluation
Consideration of socio-economic disruptions	Rarely considered	Explicit analysis of pre-COVID, COVID, and post-COVID phases
Spatial heterogeneity analysis	Limited	City-wise and location-specific performance analysis
Scientific interpretation	Minimal	Interpretation of pollution dynamics and model behaviour
Practical relevance	Primarily forecasting	Decision support for air quality management and policy

An analysis of comprehensive Mean Absolute Error (MAE) across all locations of Mumbai is shown in the graph in Fig. 6, which highlights the predicted accuracy of a particular model on the dataset. Various regions, such as Bandra-Kurla Complex (BKC), highlight the highest MAE value, which shows a major difference between actual and predicted values. This indicated that there may be inconsistencies in the data or that environmental or contextual elements are complicated. Deonar and Chakala Andheri (East) highlighted equal prediction accuracy challenges. In contrast,

locations such as Kurla, CSMT Airport, and Sion showed relatively lower MAE values, which reflect better model accuracy or more stable data conditions. Overall, the graph highlighted the spatial variability of prediction performance across Mumbai, underlining the importance of location-specific analysis and potential model refinement to improve accuracy in high-error zones.

Exploring the factors of air pollutants concentration and predicting Major air pollutants, such as particulate matter ($PM_{2.5}$ and PM_{10}), nitrogen dioxide (NO_2), sulphur dioxide (SO_2), carbon monoxide (CO), carbon dioxide (CO_2), and ozone (O_3), and AQI index more accurately is a common concern for the scientific community and policymakers. To make a more accurate prediction of air quality, this study addressed the problem of air quality monitoring from different Indian metropolitan cities, including Ahmedabad, Bengaluru, Chennai, Delhi, Kolkata, and Mumbai (from various highly polluted locations). The study analysed a comparison between traditional baseline and ensemble machine learning techniques such as SVM, GBM, and XGBM [32]-[35]. After the feature extraction, an XGBoost model is constructed, and the extracted features are used as input to monitor the air quality. The obtained prediction accuracy is much higher than that of the support vector machine and gradient boosting machine [36]-[38].

Table II shows the comparison of existing air quality prediction studies with proposed machine learning methods. Support vector machine, Gradient boosting machine, and Extreme gradient boosting machines explicitly analyse city-wise and location-specific performance on different evaluation metrics within three different temporal phases of pre-, during-, and post-COVID.

V. Implications for Policy and Decision Making

The results of this study highlight that machine-learning models for air quality monitoring can offer value beyond numerical prediction accuracy when their performance is interpreted in relation to local pollution characteristics. Cities such as Bengaluru and Chennai showed stable and consistent model behaviour across different time periods, suggesting that data-driven forecasting approaches can be effectively used to support routine air quality advisories and short-term planning in these regions. In contrast, large metropolitan areas such as Delhi, Ahmedabad, and several locations within Mumbai exhibited higher prediction variability, reflecting complex and rapidly changing emission patterns that are more difficult to capture using meteorological data alone. These can be attributed to seasonal factors such as winter smog, industrial emissions, and vehicular density. These findings suggest that policymakers should prioritise region-specific interventions, such as stricter emission control during peak pollution seasons and enhanced monitoring in high-variability zones.

These outcomes recommend that a single, uniform forecasting strategy may not be appropriate for all urban environments. For cities with highly variable pollution dynamics, machine-learning predictions should be complemented with additional information such as seasonal emission patterns, local activity data, and expert-driven thresholds. Periods of sudden socio-economic change, such as the COVID-19 lockdown, further demonstrate that predictive models need regular reassessment to remain reliable under changing conditions.

At a practical level, the observed location-specific differences in model performance support the need for zone-level air quality management rather than city-wide generalizations. Areas with consistently higher prediction errors may require improved monitoring coverage or targeted mitigation efforts. The models rely primarily on pollutant and meteorological data, without incorporating external factors such as traffic density or industrial activity, which may influence prediction accuracy. Overall, this study reinforces the importance of using machine-learning tools as part of a broader decision-support framework, where model outputs inform but do not solely determine air quality management and policy actions, including those guided by agencies such as the Central Pollution Control Board (CPCB). Potential data inconsistencies and missing values in real-world datasets may affect model generalization.

VI. Conclusion

This research work successfully presented a comprehensive analysis of traditional and advanced ensemble machine learning techniques to monitor air quality across major Indian metropolitan cities using long-term, high-resolution air quality data spanning pre-COVID, COVID-19, and post-COVID phases. The work provides valuable insights into how pollution patterns and model performance evolve under varying urban and socio-economic conditions. The comparative evaluation demonstrated that ensemble-based approaches, particularly Extreme Gradient Boosting Machine, consistently outperform traditional models such as Support Vector Machine by capturing complex, non-linear relationships between meteorological variables and different pollution dynamics. The results further disclose pronounced spatial heterogeneity in predictive performance, with cities such as Bengaluru, Kolkata, and Chennai exhibiting stable and improved accuracy, while substantially populated and industrialized metropolitan cities like Delhi, Ahmedabad, and several Mumbai zones, such as Vile Parle(W) and Bandra Kurla Complex, remain challenging due to highly dynamic emission patterns. The COVID-19 period uncovered significant disruptions in pollution behaviour, underscoring the limitations of logical models trained solely on meteorological data. These findings emphasise the need for adaptive, location-specific modelling strategies in urban air quality surveillance. Overall, the study emphasizes the need for location-specific and adaptive air quality modelling strategies. Machine-learning models can serve as effective decision-support tools when their limitations are clearly understood and when they are used alongside domain knowledge and regulatory frameworks.

To ensure robustness, each model was evaluated over multiple runs, and average performance metrics were reported. Although confidence intervals were not explicitly calculated, performance consistency across multiple cities and temporal phases provides a reliable assessment of model stability. Future work will incorporate statistical significance testing and confidence interval estimation. Future scope can expand

the model to include more cities and larger labelled datasets to further strengthen prediction performance. Further move to deep learning frameworks to train larger datasets by improving the integrity of the domain-adaptation mechanism so that it can perform more effectively across regions with different pollution dynamics. With these advancements, the overall system can evolve into a more reliable and scalable air-quality monitoring solution that supports policy development and public-health awareness planning.

AUTHORS' CONTRIBUTION STATEMENT

The corresponding author (Khushbu Chauhan) was primarily responsible for collecting datasets and pre-processing them and also, developing the proposed algorithms and conducting a detailed performance assessment to verify their alignment with the study's research goals. Khushbu Chauhan also prepared the initial version of the manuscript, carried out the visual interpretation of the results, and assisted in shaping the overall organization of the study. The second author (Kruti Sutaria) provided ongoing supervision, guided the formulation of the methodological framework, and undertook the formal analytical components of the study. Kruti Sutaria further reviewed, revised, and enhanced the manuscript to ensure technical soundness and completeness.

ACKNOWLEDGMENT

The authors would like to sincerely appreciate the Engineering and Technology department of Parul University (Gujarat) for providing them with the opportunity to work on this latest issue. We would also like to express our gratitude to the Central Pollution Control Board (CPCB), State Pollution Control Boards (SPCBs), and Multiple Corporations for supplying accurate, real-time data on Ahmedabad, Bengaluru, Chennai, Delhi, Kolkata, and Mumbai cities for the research and analysis.

DECLARATION STATEMENTS

The author declares that there are no conflicts of interest related to this research. Additionally, the authors did not receive any funding. This manuscript is the author's original work and has not been previously published or submitted for review to any other journal or conference.

REFERENCES

[1] Clark, Colin. "Urban population densities." Journal of the Royal Statistical Society. Series A (General) vol. 114, no. 4, pp. 490-496, 1951. doi.org/10.2307/2981088.

[2] Fent, Thomas. "Department of Economic and Social Affairs, Population Division, United Nations Expert Group Meeting on Social and Economic Implications of Changing Population", Age Structures: New York, vol. 386, pp.451-452, 2008. doi: 10.1007/s10680-008-9165-7.

[3] EPA, Air Quality Index. "A guide to air quality and your health." USA: EPA, 2009.

[4] Five major air pollutants: https://www.airnow.gov/aqi/aqi-basics/.

[5] Air sensor guidebook: EPA United states environmental protection agency EPA 600/R-14/159 JUN 2014.

[6] Pan, P., Malarvizhi, A. S., & Yang, C. "Data Augmentation Strategies for Improved PM2. 5 Forecasting Using Transformer Architectures", Atmosphere, vol. 16, no.2, pp.127, 2025. https://doi.org/10.3390/atmos16020127.

[7] Alrasheedi, N. D. N., Masseran, N., Tajuddin, R. R. M. "A Systematic Literature Review on the Estimation of High Air Pollution Periods Using Machine Learning Approaches", Contemporary Mathematics, pp. 3184-3208, 2025. https://doi.org/10.37256/cm.6320255760.

[8] Muhammad, A. S., Che Rose, F. Z., Marsani, M. F. "Trend analysis and performance of machine learning models for agroclimatology parameters in Bosso, Nigeria", Theoretical and Applied Climatology, vol. 156, no. 4, pp. 1-16, 2025. https://doi.org/10.1007/s00704-025-05438-7.

[9] Kumari, S., Choudhury, A., Karki, P., Simon, M., Chowdhry, J., Nandra, A., Garg, M. C. "Next-Generation Air Quality Management: Unveiling Advanced Techniques for Monitoring and Controlling Pollution", Aerosol Science and Engineering, pp. 1-22, 2025. https://doi.org/10.1007/s41810-024-00281-1.

[10] Satish, M., Biswas, S. K., Purkayastha, B. "Predictive Purity: Advancements in Air Pollution Forecasting through Machine Learning", Optical Memory and Neural Networks, vol. 34, no. 2, pp. 256-271, 2025. https://doi.org/10.3103/S1060992X25700031.

[11] Waghmare, S., & Ghadvir, G. "Assessing urban air quality of Pune city using AI-based predictive model: a data-driven approach for forecasting air quality index", Asian Journal of Civil Engineering, pp.1-11, 2025. https://doi.org/10.1007/s42107-025-01334-7.

[12] Ansari, A., & Quaff, A. R. "Advanced machine learning techniques for precise hourly air quality index (AQI) prediction in Azamgarh, India", International Journal of Environmental Research, vol. 19, no. 1, pp. 15, 2025. https://doi.org/10.1007/s41742-024-00684-5.

[13] Dawar, I., Singal, M., Singh, V., Lamba, S., Jain, S. "Predicting air quality index using machine learning: a case study of the Himalayan city of Dehradun", Natural Hazards, vol. 121, no. 5, pp. 5821-5847, 2025. https://doi.org/10.1007/s11069-024-07027-9.

[14] Desai, T., Kapadia, S., Halani, M., Zinzuwadia, P., Shah, K., Shah, M., Prajapati, M. "Comparative analysis of machine learning algorithms for air quality index prediction", Machine Learning for Computational Science and Engineering, vol. 1, no. 1, pp. 14, 2025. https://doi.org/10.1007/s44379-025-00014-2.

[15] Sharma, D., Thapar, S., Sachdeva, K. "Enhancing particulate matter prediction in Delhi: insights from statistical and machine learning models", Environmental Monitoring and Assessment, vol. 197, no. 7, pp. 723, 2025. https://doi.org/10.1007/s10661-025-14121-3.

[16] Ahmad, N., & Kumar, V. "Spatio-Temporal Forecasting using a Hybrid BiGRU-1DCNN Model for PM 2.5 Concentrations in Delhi, India (2018-2023)", Across Multiple Monitoring Stations. Water, Air, & Soil Pollution, vol. 236, no. 7, pp. 459, 2025. https://doi.org/10.1007/s11270-025-08103-x.

[17] Saini, J., Dutta, M., & Marques, G. "Machine learning for indoor air quality assessment: A systematic review and analysis", Environmental Modeling & Assessment, vol. 30, no. 2, pp. 417-434, 2025. https://doi.org/10.1007/s10666-024-10001-1.

[18] Naveed, K., Umer, T., Asghar, A. B., Aslam, M., Ejsmont, K., Metwally, A. S. M., Thanh, K. N. "Machine Learning Assisted Predictive Urban Digital Twin for Intelligent Monitoring of Air Quality Index for Smart City Environment", Environmental Modelling & Software, pp. 106559, 2025. https://doi.org/10.1016/j.envsoft.2025.106559.

[19] Alimohammadi, H., & Chen, S. N. "Performance evaluation of outlier detection techniques in production timeseries: A systematic review and meta-analysis", Expert Systems with Applications, vol. 191, pp. 116371, 2022. https://doi.org/10.1016/j.eswa.2021.116371Get rights and content.

[20] Dun, M., Xu, Z., Chen, Y., & Wu, L. "Short-Term Air Quality Prediction Based on Fractional Grey Linear Regression and Support Vector Machine", Mathematical Problems in Engineering, vol.2020,no.1,pp.8914501,2020. https://doi.org/10.1155/2020/8914501.

[21] Mogollón-Sotelo, C., Casallas, A., Vidal, S., Celis, N., Ferro, C., & Belalcazar, L. "A support vector machine model to forecast ground-level PM2. 5 in a highly populated city with a complex terrain", Air Quality, Atmosphere & Health, vol.14, no.3, pp.399-409, 2021. https://doi.org/10.1007/s11869-020-00945-0.

[22] Moazami, S., Noori, R., Amiri, B. J., Yeganeh, B., Partani, S., & Safavi, S. "Reliable prediction of carbon monoxide using developed support vector machine", Atmospheric Pollution Research, vol.7, no.3, pp.412-

418, 2016. https://doi.org/10.1016/j.apr.2015.10.022.

[23] Akhmadkhon, A. "THE ROLE OF GRADIENT BOOSTING MACHINES IN MODERN ECONOMIC ANALYSIS", Universum: технические науки, vol.6, no.1 (130), pp. 11-14, 2025.

[24] Bentéjac, C., Csörgő, A., Martínez-Muñoz, G. "A comparative analysis of gradient boosting algorithms", Artificial Intelligence Review, vol. 54, no.3, pp.1937-1967, 2021. https://doi.org/10.1007/s10462-020-09896-5.

[25] Singh, U., Rizwan, M., Alaraj, M., Alsaidan, I. "A machine learning-based gradient boosting regression approach for wind power production forecasting: A step towards smart grid environments", Energies, vol.14, no.16, pp.5196, 2021. https://doi.org/10.3390/en14165196.

[26] Lei, T. M., Ng, S. C., & Siu, S. W. "Application of ANN, XGBoost, and other ML methods to forecast air quality in Macau", Sustainability, vol.15, no.6, pp.5341, 2023. https://doi.org/10.3390/su15065341.

[27] Matandirotya, N. R., & Burger, R. "An assessment of NO2 atmospheric air pollution over three cities in South Africa during 2020 COVID-19 pandemic", Air Quality, Atmosphere & Health, vol.16, no.2, pp.263-276, 2023. https://doi.org/10.1007/s11869-022-01271-3.

[28] Kumar, S., Lodhi, A., Chauhan, A., & Kumar, A. "Comparative Analysis of Forecasting Models for Air Quality Index Prediction", In 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS) IEEE, pp. 190-195, 2022. DOI: 10.1109/ICICCS53718.2022.9788302.

[29] Zarrar, H., & Dyo, V. "Drive-by air pollution sensing systems: Challenges and future directions", IEEE Sensors Journal, vol.23, no.19, pp.23692-23703, 2023. DOI: 10.1109/JSEN.2023.3305779.

[30] Liu, Z., Chen, K., Ning, Z., Wang, L., Zheng, Z. "Research on air quality prediction based on correlation analysis and XGBoost", In 2023 International Conference on Electronics and Devices, Computational Science (ICEDCS) IEEE, pp. 761-766, 2023.

[31] Fu, B., Liu, T., Dong, H., Zhu, M., & Yang, Y. "XGBoost-Based Prediction Model for Mine Signal Classification", In 2024 7th International Conference on Mechatronics and Computer Technology Engineering (MCTE*)* IEEE, pp. 1683-1686, 2024.

[32] Liu, B., Tan, X., Jin, Y., Yu, W., & Li, C. "Application of RR-XGBoost combined model in data calibration of micro air quality detector", *Scientific Reports*, vol. 11, no.1, pp.15662, 2021.

[33] Cui, Z., Huang, C., Huang, Z., & Chen, A. "Air quality prediction and early warning based on Prophet-XGBoost combined model", In Fourth International Conference on Applied Mathematics, Modelling, and Intelligent Computing (CAMMIC 2024), SPIE. vol. 13219, pp. 586-597, 2024.

[34] Ma, X., Fang, C., Ji, J. "Prediction of outdoor air temperature and humidity using Xgboost", In IOP conference series: earth and environmental science, IOP Publishing. vol. 427, no. 1, pp. 012013, 2020.

[35] Naizabayeva, L., Sembina, G., Aliman, A., Satymbekov, M., Barlykbay, N., Seilova, N. "Air Pollution Forecasting in Almaty using Ensemble Machine Learning Models", Journal of Applied Data Sciences, vol.*6*, no.4, pp.2461-2476, 2025.

[36] Winarto, R., Purnomo, M. H., Anggraeni, W. "Enhancing Predictive Emissions Monitoring Performance: Data Preprocessing for XGBoost-Based Model Algorithm", In 2025 17th International Conference on Knowledge and Smart Technology (KST) IEEE, pp. 278-283, 2025.

[37] Wang, Z., Wu, X., & Wu, Y. "A spatiotemporal XGBoost model for PM2. 5 concentration prediction and its application in Shanghai", Heliyon, vol.*9*, no.12, 2023.

[38] Liumei Zhang, Yangna Ji, Tianshi Liu, Jiao Li "PM2.5 Prediction Based on XGBoost", 7th International Conference on Information Science and Control Engineering (ICISCE) IEEE, 2020.