Bachandeep Singh Bhathal[1], Dr. Gaurav Gupta[2]

Air Quality Prediction Using a Machine Learning Hybrid Model in Punjab

Journal of Electrical Systems

Abstract: - Clean air is an essential component for the health and survival of humans and wildlife. Atmospheric pollution has been closely linked to several significant diseases, including cancer, which highlights the critical need to address air quality issues. However, with the rapid pace of industrialization and population growth, human activities such as transportation, household operations, agricultural practices, and industrial processes have significantly contributed to air pollution. This has led to air pollution becoming a major environmental and health concern, particularly in urban areas of developing countries like India. To ensure the maintenance of ambient air quality, it is imperative to conduct regular monitoring and forecasting of air pollution levels. Machine learning has emerged as an innovative and effective technique for predicting the Air Quality Index (AQI) compared to conventional forecasting methods. In this study, we applied AQI prediction methods to Punjab, India. The research focused on analysing 11 air contaminants and 9 meteorological parameters over a comprehensive timeframe spanning from July 2019 to September 2023. To achieve accurate predictions, several advanced machine learning models were employed, including LightGBM, Random Forest, Catboost, Adaboost, and XGBoost. Among these models, the Catboost model demonstrated superior performance, achieving an exceptionally high R^2 correlation coefficient of 0.9997. It also recorded a mean absolute error (MAE) of 0.62, a mean square error (MSE) of 0.60, and a root mean square error (RMSE) of 0.78, making it the most effective predictor in this study. On the other hand, the Adaboost model exhibited the least predictive capability, with an R^2 correlation coefficient of 0.9471.

Keywords: correlation, capability, predictive, comprehensive

INTRODUCTION

Population density, industrial activities, agricultural practices, thermal power plants, energy production sectors, automotive industries, and transportation systems each contribute uniquely to air pollution. These factors play a significant role in altering the quality of air in different regions (Ravindra, 2019; Ravindra et al., 2020). Air pollution not only damages ecosystems but also poses severe health risks to humans. Its adverse effects range from premature deaths and skin irritations to more serious conditions such as lung infections, respiratory tract illnesses, pneumonia, lung cancer, and even heart failure (Manisalidis et al., 2020). The extent of air pollution in a particular area is influenced by several critical factors, including particulate matter, gaseous pollutants, and meteorological conditions. These factors necessitate the estimation of the Air Quality Index (AQI), which is widely monitored by both government and non-government organizations across the globe (Bao & Zhang, 2020; L. Wu et al., 2018). Among the key pollutants contributing to AQI, Particulate Matter 2.5 ($PM_{2.5}$) and Particulate Matter 10 (PM_{10}) stand out as significant contributors. Other major pollutants include, Carbon Monoxide (CO), SulphurDioxide (SO_2), Nitric Oxide (NO), Nitrogen Oxides (NO_x), Nitrogen Dioxide (NO_2), Benzene (C_6H_6), Ozone (O_3),Toluene ($C_6H_5CH_3$) and Ammonia (NH_3). These substances are pivotal in determining air quality and have far-reaching effects on the environment. Elevated AQI levels resulting from these pollutants have numerous detrimental consequences for the environment. These include global warming, the formation of acid rain, the development of smog and aerosols, reduced visibility, and significant contributions to climate change (Balakrishnan et al., 2019). These environmental impacts, combined with the

[1]bachan0235@gmail.com[1], gaurav.shakti@gmail.com[2]
Assistant Professor, University Institute of Computing, Chandigarh University, Mohali[1]
Assistant Professor, Department of Computer Science and Engineering, Punjabi University, Patiala[2]

health risks associated with air pollution, underline the urgency of effective monitoring and mitigation strategies. Addressing these challenges requires a comprehensive understanding of the sources and contributors to air pollution, as well as the development of advanced predictive tools for AQI estimation.

Greenhouse gases (GHGs) are a major contributor to global warming and have far-reaching effects beyond climate change. These gases significantly influence plant-soil interactions, leading to adverse consequences for agriculture, the environment, and the economy. Their impact is particularly critical, as they threaten the sustainability of ecosystems and the livelihoods of populations dependent on agricultural and environmental stability (Malhi et al., 2021). In 2022, the World Health Organization (WHO) released a comprehensive report on global air quality, which analysed data from 2010 to 2019. This report provided a detailed examination of various air pollutants, including those known for their significant contributions to environmental and health issues. Among these pollutants, $PM_{2.5}$ was highlighted as a growing concern. The study, which assessed data from 6,743 cities across 117 countries, revealed a disturbing global increase in $PM_{2.5}$ levels. This rise has been linked to severe health consequences, including 1.7 million annual deaths in India alone, underscoring the alarming health burden posed by air pollution. The report also drew attention to India's critical air quality challenges. Among the 20 cities worldwide with the highest air pollution levels, 18 were located in India. This finding reflects the gravity of India's air quality crisis and its potential for severe health impacts in the coming years. The disproportionate representation of Indian cities in this list illustrates the urgency of addressing air pollution to safeguard public health and the environment (Gurjar et al., 2016; Guttikunda et al., 2014). These findings emphasize the need for targeted interventions, stringent regulations, and innovative strategies to mitigate air pollution and its associated consequences.

A high Air Quality Index (AQI) value signifies a dangerously polluted environment that poses severe health risks, making it a critical concern for human safety and sustainability. As a result, monitoring and forecasting AQI have become indispensable tools in the global effort toward sustainable development (Rybarczyk & Zalakeviciute, 2021). Various methods have been developed for AQI prediction, utilizing statistical, deterministic, physical, machine learning, and deep learning approaches. However, the inherent rigidity of traditional statistical and decision-making models often renders them unsuitable for addressing the complex and dynamic nature of air pollution. Recent advancements in sensor technology have simplified the detection of air pollution levels, enabling the automatic calculation of AQI. With the availability of extensive datasets, forecasting AQI has become more accessible and efficient (Bekkar et al., 2021). Among the various approaches, machine learning has proven to be highly effective, offering precise and consistent AQI predictions under diverse environmental conditions. Machine learning models leverage the growing volume of historical data to improve the accuracy of forecasts, making them a promising alternative to traditional statistical models, particularly for time-series forecasting. Unlike statistical models, which struggle to address the highly nonlinear and complex dynamics of pollutant concentrations, machine learning models excel in handling such challenges. These models are nonparametric and nonlinear, relying solely on historical data to identify correlations between independent variables and pollutants. This capability allows for the development of more accurate prediction models, even when the underlying dynamics of pollution processes are not fully understood. By accommodating the intricacies of air quality data, machine learning has emerged as a reliable and innovative approach, reinforcing its role in advancing AQI forecasting and supporting sustainable environmental management.

Punjab, located in the northern region of India, is a prominent state known for its agricultural significance and strategic importance. Often referred to as the “Granary of India”, Punjab plays a crucial role in India's food grain production. The state benefits from certain meteorological variables, such as strong winds during specific seasons, which help disperse pollutants and reduce air pollution levels. However, Punjab's geographical and climatic conditions, including its predominantly flat terrain and frequent occurrences of temperature inversions during the winter months, contribute to significant air pollution concerns.Winter temperature inversions trap pollutants close to the ground, leading to elevated levels of particulate matter and other pollutants. This situation is further exacerbated by seasonal activities such as crop residue burning, commonly known as stubble burning, which adds large quantities of particulate matter and gaseous pollutants to the atmosphere. These factors make air pollution a critical issue in Punjab, particularly during the post-harvest and winter seasons, necessitating effective monitoring and mitigation strategies to safeguard public health and the environment. Punjab is regarded as one of India's most agriculturally and industrially significant states, contributing extensively to the country's economy. While it is renowned for its agricultural dominance, the state also boasts a diverse industrial

base that includes textiles, hosiery, sports goods, agricultural machinery, bicycles, pharmaceuticals, and food processing industries. Punjab's industrial hubs, such as Ludhiana, Jalandhar, and Amritsar, are well-known for their manufacturing capabilities. Additionally, the state houses several thermal power plants that contribute to its energy production.Natural resources, such as fertile soil and an extensive canal irrigation system, have played a pivotal role in establishing Punjab as an agricultural powerhouse. However, its industrial and agricultural activities have also contributed to environmental challenges. Historical AQI data and the rapid growth of industries and vehicular emissions have identified Punjab as a region facing increasing air pollution issues, particularly in urban and peri-urban areas. These challenges underscore the need for effective pollution control measures to balance economic growth with environmental sustainability. This study focuses on analyzing open-source air quality data collected from the Central Pollution Control Board (CPCB) for the period from January 2018 to December 2023. Using this data, we employed five advanced machine learning algorithms - LightGBM, Random Forest, CatBoost, AdaBoost, and XGBoost - to predict the AQI of the city. Each model was selected for its unique strengths, ensuring robust predictions tailored to the specific characteristics of the dataset and the problem at hand.

- **LightGBM** is particularly effective in handling categorical features, providing a fast training speed and strong predictive performance.
- **Random Forest** excels in managing both numerical and categorical data and offers valuable feature importance rankings, aiding interpretability.
- **CatBoost** is recognized for its superior generalization ability and strong performance, particularly in handling categorical variables.
- **AdaBoost** enhances overall prediction accuracy by iteratively adjusting instance weights to focus on difficult-to-predict cases.
- **XGBoost** employs advanced regularization techniques to control overfitting and offers flexibility in model customization.

The combined strengths of these machine learning models make them highly effective for AQI prediction in a dynamic and complex urban and rural environment’s like Punjab. By leveraging the predictive capabilities of these models, this study provides valuable insights into air quality trends and supports data-driven decision-making for pollution management in the state.

MATERIALS AND METHODS

Study Area

This study focuses on analysing the Air Quality Index (AQI) for Punjab, a northern state in India, renowned for its agricultural and industrial significance as well as its rapid economic development. Punjab has a well-established network of air quality monitoring stations operated by the Central Pollution Control Board (CPCB) of India, which provides essential data for evaluating air quality trends and patterns. Geographically, Punjab is located between 29°30′ and 32°32′ North latitude and 73°55′ and 76°50′ East longitude, covering an area of approximately 50,362 square kilometres. Punjab is one of India’s leading agricultural states and serves as a vital industrial hub with diverse industries, including textiles, food processing, agricultural machinery, bicycles, pharmaceuticals, and sports goods. Cities like Ludhiana, Jalandhar, and Amritsar are industrial centres contributing significantly to the state’s economy. Recognized as a key contributor to India's economic and agricultural output, Punjab plays a crucial role in the nation's development.However, rapid industrialization and extensive agricultural practices, including stubble burning, have led to significant air quality challenges in the state. These issues are particularly pronounced during the post-harvest and winter seasons due to temperature inversions and high levels of particulate matter. This underscores the importance of AQI monitoring and forecasting in Punjab to address these challenges and promote sustainable development in the region.

Air Quality and Meteorological Datasets

The data for this study was obtained from the Central Pollution Control Board - Central Control Room for Air (CPCBCRR). To calculate the Air Quality Index (AQI) as per the guidelines set by the CPCB, various air pollutants, including $PM_{2.5}$, PM_{10}, CO, SO_2, NO, NO_X, NO_2, C_6H_6, O_3, $C_6H_5CH_3$, and NH_3 were measured. The AQI calculation considers the maximum value of any one particulate matter and any three gaseous pollutants,

reflecting the combined impact of key pollutants. In addition to pollutant concentrations, AQI is significantly influenced by local meteorological conditions.

Meteorological parameters monitored during the study included Temperature, Relative Humidity (RH), Wind Speed (WS), Wind Direction (WD), Solar Radiation (SR), Air Pressure (BP), Ambient Temperature (AT), Rainfall (RF), and Total Rainfall (TOT-RF). These parameters play a crucial role in shaping air quality patterns by influencing pollutant dispersion and accumulation. The dataset comprised a total of 43,543 observations collected over the period from January 1, 2019, to December 31, 2023. Each observation included 21 components: 11 air pollutants, 9 meteorological factors, and 1 AQI value, which served as the target variable for prediction. The raw dataset formed the foundation for this analysis, offering detailed insights into the interplay between pollutants and meteorological variables over the study period. The AQI data and its associated components were systematically organized and are presented in Table 1: Skewness and Kurtosis Values of Selected Features before and after Transformation. This comprehensive dataset enabled a robust examination of air quality trends and provided a reliable basis for predictive modelling.

Table 1: Skewness and Kurtosis Values of Selected Features before and after Transformation

S.No.	Attributes	Before Transformation		After Transformation	
		Skewness	Kurtosis	Skewness	Kurtosis
	$PM_{2.5}$ $(\mu g/m^3)$	2.41	14.77	2.41	2.12
	PM_{10} $(\mu g/m^3)$	2.30	12.62	2.30	1.71
	$CO(\mu g/m^3)$	4.06	33.58	4.06	4.37
	SO_2 $(\mu g/m^3)$	9.04	143.11	9.04	1.22
	$NO(\mu g/m^3)$	10.83	198.99	2.66	3.77
	NO_x $(\mu g/m^3)$	4.53	40.71	2.81	4.03
	NO_2 $(\mu g/m^3)$	5.12	71.70	2.23	3.20
	C_6H_6 $(\mu g/m^3)$	52.55	6247.61	10.28	5.49
	O_3 $(\mu g/m^3)$	1.61	3.46	1.61	3.46
	$C_6H_5CH_3$ $(\mu g/m^3)$	3.07	55.11	3.07	2.08
	NH_3 $(\mu g/m^3)$	4.15	47.55	1.30	2.04

Machine Learning Methods to Predict AQI

To predict the Air Quality Index (AQI) for the proposed state, several advanced machine learning models were employed. These included Light Gradient Boosting Machine (LightGBM), Random Forest, CatBoost, AdaBoost, and XGBoost. Each of these models was selected for its unique ability to handle complex datasets and provide accurate predictions. This dataset integrated air pollutant concentrations, meteorological factors, and AQI values, ensuring a comprehensive framework for predictive modelling. By leveraging these machine learning algorithms, the study aimed to achieve reliable AQI forecasts, which are critical for air quality management and sustainable urban planning.

- **LightGBM**

LightGBM (Light Gradient Boosting Machine) is a highly reliable and efficient tool for implementing gradient boosting in decision trees (Yihuan Zhou et al., 2022). It employs tree-based learning strategies, which make it an excellent choice for gradient boosting tasks. Its decentralized and optimized architecture ensures faster training and higher output efficiency, making it particularly suitable for large and complex datasets. One of LightGBM's standout features is its histogram-based approach to variable bucketing. This method groups continuous variables into discrete bins, significantly improving training speed and accuracy while reducing memory usage. LightGBM can handle large-scale datasets efficiently, and its support for parallel and GPU-based learning further enhances its performance, especially when processing computationally intensive tasks. In supervised learning scenarios, LightGBM is particularly effective. It allows for the prediction of a target variable Y using only input features X. The technique involves a supervised training set X and a loss function L (y, f(x)), where

the aim is to minimize the predicted loss function to approximate $\hat{f}$(x). This process ensures that LightGBM delivers precise and reliable predictions, making it a valuable tool in machine learning applications, including AQI forecasting and other predictive modelling tasks.

$$\hat{f} = \arg min_f \, E_{y,X} \, L(y, f(x))$$

- **Random Forest**

Random Forest Regression is a widely used machine learning-based regression technique that builds upon the principles of bagging and random subspace methods (Ganesh et al., 2021). This approach employs an ensemble of decision trees to generate a more accurate and stable prediction by combining the outputs of multiple trees. The process begins with the creation of a bootstrap sample D_b from the training dataset (D), which consists of N total examples. To form the bootstrap sample, n random instances are selected from the dataset, allowing for replacement during sampling. This ensures that some instances may appear multiple times in the bootstrap sample while others may be excluded. Using the input vector x, K distinct regression trees are constructed, each trained on a different bootstrap sample. During the training phase, the Random Forest algorithm introduces randomness by selecting a subset of features at each node to determine the best split, which enhances model generalization and prevents overfitting. For regression tasks, the prediction is made by averaging the outputs of all K regression trees. Mathematically, this can be expressed as:

$$\hat{h} = \frac{1}{K}\sum_{k=1}^{K} h_k(x)$$

where$h_k(x)$ represents the prediction made by the k-th regression tree. By aggregating predictions from multiple trees, Random Forest Regression minimizes variance, improves robustness, and delivers reliable performance, especially when dealing with complex datasets. This makes it a valuable tool for predictive modelling, including applications like AQI forecasting.

- **CatBoost**

CatBoost is an advanced machine learning framework built upon the principles of gradient boosting and decision trees, designed to enhance the predictive power of traditional boosting models (Zhang et al., 2020). Boosting is based on the concept that a combination of many weak models can produce a highly accurate prediction model. These weak models, which individually perform slightly better than random chance, are combined iteratively to minimize errors and improve overall performance. In gradient boosting, errors are reduced by sequentially fitting decision trees, where each tree learns from the mistakes of its predecessor. This iterative process continues until the loss function, which measures prediction errors, can no longer be significantly minimized. However, CatBoost introduces a unique approach to constructing decision trees, setting it apart from traditional gradient boosting models. CatBoost employs "oblivious trees," a distinctive type of decision tree in which all nodes at the same level use the same predictor and apply identical conditions. This uniformity allows for efficient computation, as the index of a leaf node can be determined using simple bitwise operations. Additionally, CatBoost integrates a random permutation, denoted by σ, to order the dataset D during tree construction. The dataset elements are arranged as $D_k = \{x_1, x_2, \ldots, x_{k-1}\}$, where $x_1, x_2, \ldots, x_{k-1}$ represent elements ordered by the permutation, and k indicates the k-th element of D under the permutationσ. Another standout feature of CatBoost is its handling of categorical features during decision tree construction. Instead of strictly adhering to traditional methods, CatBoost uses an innovative approach to define the encoded value $\hat{x}_{ik}$ for the i-th categorical value during the fitting of the decision tree h_{t+1}. This efficient encoding method significantly enhances the model's ability to generalize and reduces overfitting.

By combining these advanced methodologies, CatBoost delivers superior performance in a wide range of predictive tasks, especially when handling datasets with categorical variables and complex dependencies. This makes it a powerful tool for tasks like AQI forecasting and other data-intensive applications. Here $x_j^i = x_k^i$ is the indicator function.

$$\hat{k}_k^i = \frac{\sum x_j \in D_k 1 x_j^i = x_k^i . y_j + ap}{\sum x_j \in D_k 1 x_j^i = x_k^i + a}$$

- **Adaptive Boosting (AdaBoost) Regressor**

AdaBoost, short for Adaptive Boosting, is one of the pioneering boosting algorithms developed to address complex prediction challenges effectively (Mishra et al., 2020). The algorithm operates by iteratively improving the performance of weak learners, transforming them into a strong ensemble model capable of making highly accurate predictions. The key principle of AdaBoost is the assignment of weights to each training sample(x_i, y_i), denoted as$w_1, w_2, \dots, w_N$. Initially, all observations in the dataset are treated equally, with the fundamental learner giving each sample the same level of importance. However, as the learning process progresses, weights are dynamically adjusted. Higher weights are assigned to samples that were misclassified or poorly predicted by the weak learner, compelling subsequent iterations to focus more on these difficult cases. In each iteration, the weak learner is used to make predictions based on the weighted dataset. This ensures that the algorithm progressively reduces the error rate, as the weak learner becomes increasingly accurate in handling challenging observations. The process continues for t iterations or until the predetermined limit of base learning algorithms T_i is reached.

At the end of the training process, the outputs of all weak learners are combined to create a robust ensemble model. This final model aggregates the predictions of individual weak learners, weighted by their accuracy, to form a stronger predictor with enhanced reliability and predictive capability. AdaBoost's iterative adjustment of weights and combination of weak learners makes it a versatile and powerful tool for both classification and regression tasks. Its ability to focus on challenging samples and progressively refine predictions has made it a widely adopted algorithm in various applications, including AQI forecasting and other machine learning problems.

- **Extreme Gradient Boosting (XGBoost)**

Boosting is a machine learning technique that combines multiple weak classifiers to create a single, more effective predictive model. XGBoost (eXtreme Gradient Boosting), a refined version of Gradient Boosting, was developed to enhance computational efficiency, scalability, and generalization performance, making it a powerful tool for various predictive tasks (Mahesh et al., 2022). XGBoost builds upon the principles of Gradient Boosting but introduces several optimizations that significantly improve its performance. These include better handling of sparse data, regularization techniques to prevent overfitting, and parallelized computations for faster training. Such advancements allow XGBoost to outperform the original Gradient Boosting algorithm in terms of speed, accuracy, and scalability. A critical aspect of using XGBoost is the proper organization of data. Since XGBoost only accepts numeric input, all categorical data must be converted into numerical equivalents. This transformation is typically achieved through techniques like one-hot encoding, which represents categorical variables as binary vectors. Following the encoding, feature engineering and data cleaning steps are performed to ensure the quality and relevance of the input data.

XGBoost predicts the target variable using an iterative approach, where each tree in the model corrects the errors of the previous one. The process is governed by a universal function that minimizes the loss function and optimizes model performance. The formula used in XGBoost can be expressed as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \ \ where \ f_k \in \mathcal{F}$$

Here,f_k represents an individual decision tree, and $\mathcal{F}$ is the space of all possible decision trees. The model learns by sequentially fitting trees to minimize the loss function, improving predictions at each step. XGBoost's ability to handle large datasets, diverse feature types, and complex relationships makes it a preferred choice for a wide range of machine learning tasks, including AQI forecasting and other time-series predictions. Its computational efficiency and strong generalization capabilities further solidify its role as a leading algorithm in modern predictive analytics.

DATA PRE-PROCESSING

A total of 43,543 instances were gathered from open-source data spanning the period from January 2019 to December 2023. During pre-processing, missing values in the dataset were identified and removed, reducing the total instances to 41,245 with 20 characteristics. After handling missing data, a type conversion was performed on the AQI variable, transforming it from an object data type to a float data type to facilitate numerical analysis.

To evaluate the performance of various machine learning models, the dataset was split into two subsets: a training set comprising 80% of the data and a testing set comprising the remaining 20%. This division ensured that the models were trained on a substantial portion of the data while reserving a separate set for unbiased evaluation.

The predictive capability of the models was assessed using widely accepted evaluation metrics:

- **Root Mean Square Error (RMSE):** Measures the square root of the average squared differences between predicted and actual values, reflecting the model's overall prediction accuracy.
- **Mean Square Error (MSE):** Captures the average squared differences between predicted and actual values, emphasizing larger errors.
- **Mean Absolute Error (MAE):** Evaluates the average of the absolute differences between predicted and actual values, providing a straightforward interpretation of prediction errors.
- **R^2 (Coefficient of Determination):** Indicates how well the model explains the variance in the target variable, with values closer to 1 representing better performance.

These metrics allowed for a comprehensive evaluation of model accuracy and performance in forecasting AQI, ensuring that the chosen models provided reliable and precise predictions based on the dataset.

EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is an essential step in uncovering hidden patterns within datasets before applying machine learning methods. EDA plays a critical role in establishing relationships between variables, particularly in identifying how various air contaminants contribute to high AQI levels (Langer & Meisen, 2021). This study utilized EDA to analyse the trends and statuses of air contaminants from 2019 to 2023, categorizing pollutants based on their significant impact on AQI. The relationships among pollutants and their influence on AQI were visualized using a heatmap, presented in Figure-1 and Figure-2. The heatmap was constructed using a correlation matrix, which quantifies the strength and direction of relationships between variables. Correlation scores range from +1 to -1, where positive values indicate a direct relationship, and negative values represent an inverse relationship (Li et al., 2016). In this analysis, pollutants with a correlation coefficient greater than 0.5 were considered to have a strong positive impact on AQI, while negative correlations were observed to have minimal influence on AQI prediction.

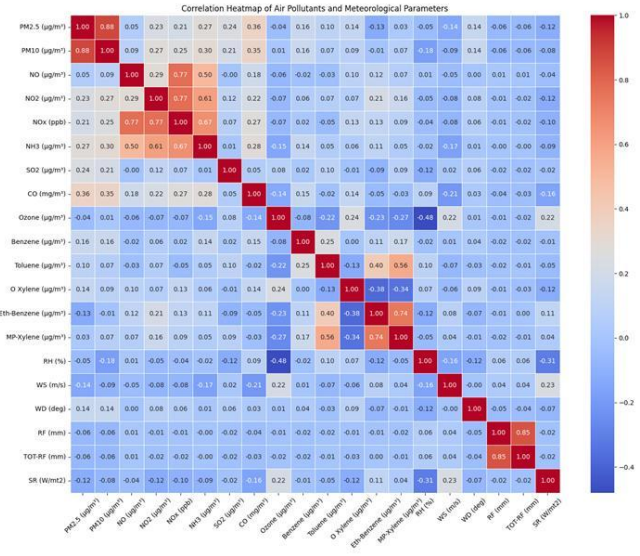

Figure-1: Correlation Heatmap of Air Pollutants and Meteorological Parameters

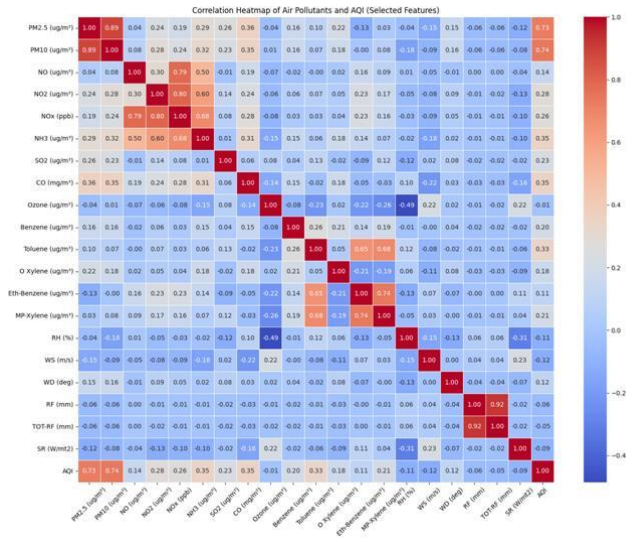

Figure-2: Correlation Heatmap of Air Pollutants and AQI

The heatmap revealed the following key correlations:

- $PM_{2.5}$ and AQI: Correlation coefficient of 0.73
- PM_{10} and AQI: Correlation coefficient of 0.74
- NO_2 and AQI: Correlation coefficient of 0.28
- CO and AQI: Correlation coefficient of 0.35
- NO_x and AQI: Correlation coefficient of 0.26

Particulate matter ($PM_{2.5}$ and PM_{10}) was identified as the most significant variable influencing AQI predictions, contributing 13% of the total variance, followed by NO_2, CO, and NO_x, which collectively accounted for 8%. Other pollutants showed weak correlations (less than 0.4) and contributed minimally to AQI predictions. Interestingly, meteorological factors were observed to have minimal influence on AQI predictions. These findings, as shown in Figure S3, align with the heatmap analysis, emphasizing the dominant role of particulate matter and select gaseous pollutants in AQI forecasting.

DATA TRANSFORMATION

Table 1 provides a summary of the skewness and kurtosis values for the major air pollutants before and after data processing. These statistical metrics are crucial for evaluating the distribution characteristics of datasets and ensuring they meet the assumptions required for various analyses.

- **Skewness** measures the asymmetry of a dataset relative to its mean. A dataset is considered symmetrical when the left and right sides of the distribution are equally distanced from the centre. In cases where skewness exists, it indicates a deviation from symmetry, with positive skewness suggesting a longer tail on the right and negative skewness indicating a longer tail on the left.
- **Kurtosis** assesses the "tailedness" of a dataset, comparing how heavy or light the tails are relative to a normal distribution. High kurtosis indicates heavy tails with extreme values, while low kurtosis reflects light tails and fewer outliers (Lord et al., 2021).

The datasets under analysis were not normally distributed initially, as indicated by skewness and kurtosis values outside the acceptable range for normality. To address this, the datasets were transformed into a normal distribution using statistical transformation techniques. Commonly applied methods included square root transformation, log transformation, and Box-Cox transformation, each selected based on the specific distribution characteristics of the data. After processing, the skewness values of the datasets were reduced, and the distributions became closer to symmetry. Similarly, kurtosis adjustments resulted in distributions more consistent with the expected normal shape. While the current datasets exhibited minor skewness, the applied transformations successfully improved their suitability for further analysis and modelling. These adjustments

were critical in preparing the data for machine learning and statistical analysis, ensuring that the datasets adhered to the assumptions required for accurate and reliable AQI predictions.

The dataset was adjusted to approximate a normal distribution by applying a square root transformation. According to the statistical values summarized in Table 1: Skewness and Kurtosis Values of Selected Features before and after Transformation, certain variables, including NO, NH_3, Xylene, and TOT-RF, exhibited significantly high skewness and kurtosis. Acceptable ranges for these metrics are between -3 and +3 for skewness and between -10 and +10 for kurtosis. However, these variables fell outside these ranges, necessitating transformation to bring the data closer to normality. The square root transformation was applied as an effective method to address these deviations. This approach reduced the skewness and kurtosis values of the affected variables, aligning them with acceptable levels and confirming a more normal distribution for the dataset. As noted by (Schneider & Wheeler-Kingshott, 2014), such transformations are essential for ensuring data suitability for statistical and machine learning models. Figure-3 shows the skewness and kurtosis ranges for various air pollutants before and after the square root transformation, highlighting significant improvements in the distribution of NO, NH3, and TOT-RF. Prior to transformation, the dataset exhibited characteristics of being either left-skewed or highly positively skewed. After applying the square root transformation, the data conformed to a normal distribution, enhancing its suitability for further analysis. These adjustments were critical for ensuring the robustness and reliability of subsequent AQI predictions and statistical analyses.

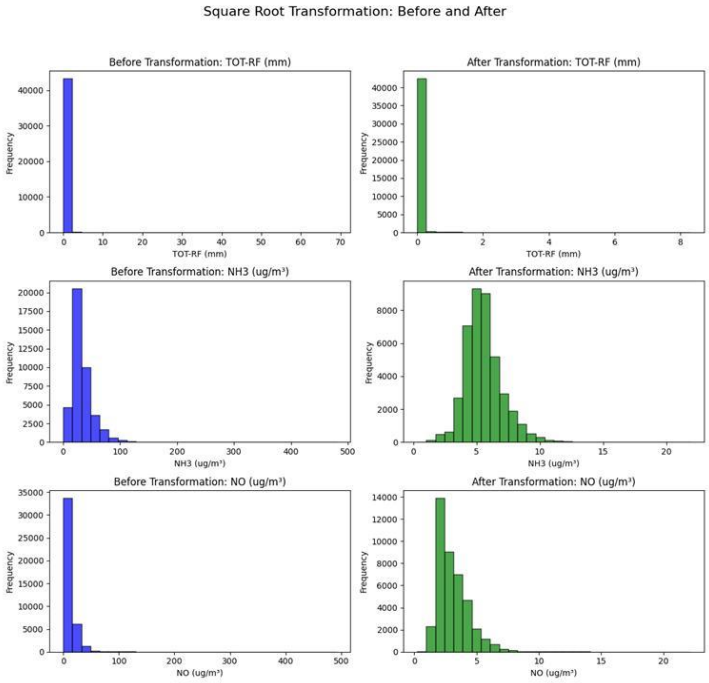

Figure-3: Square Root Transformation (Before and After)

RESULTS AND DISCUSSION

AQI Data Summary

Figure-5 illustrates the monthly and annual average concentrations of $PM_{2.5}$ and PM_{10}. The analysis reveals that PM_{10} concentrations consistently exceeded 100 µg/m³ during the months of January to March and October to December. However, during 2020, both PM_{10} and $PM_{2.5}$ concentrations dropped below 100 µg/m³ and 40 µg/m³, respectively, due to the nationwide COVID-19 lockdown, which significantly curtailed industrial and vehicular activities for five months (Singh & Chauhan, 2020). From April to September in other years, $PM_{2.5}$ and PM_{10} concentrations remained below 40 µg/m³ and 100 µg/m³, respectively. These observations clearly indicate seasonal variations in particulate matter levels, with elevated $PM_{2.5}$ and PM_{10} concentrations during the first six months of the year and reduced levels during the latter half.

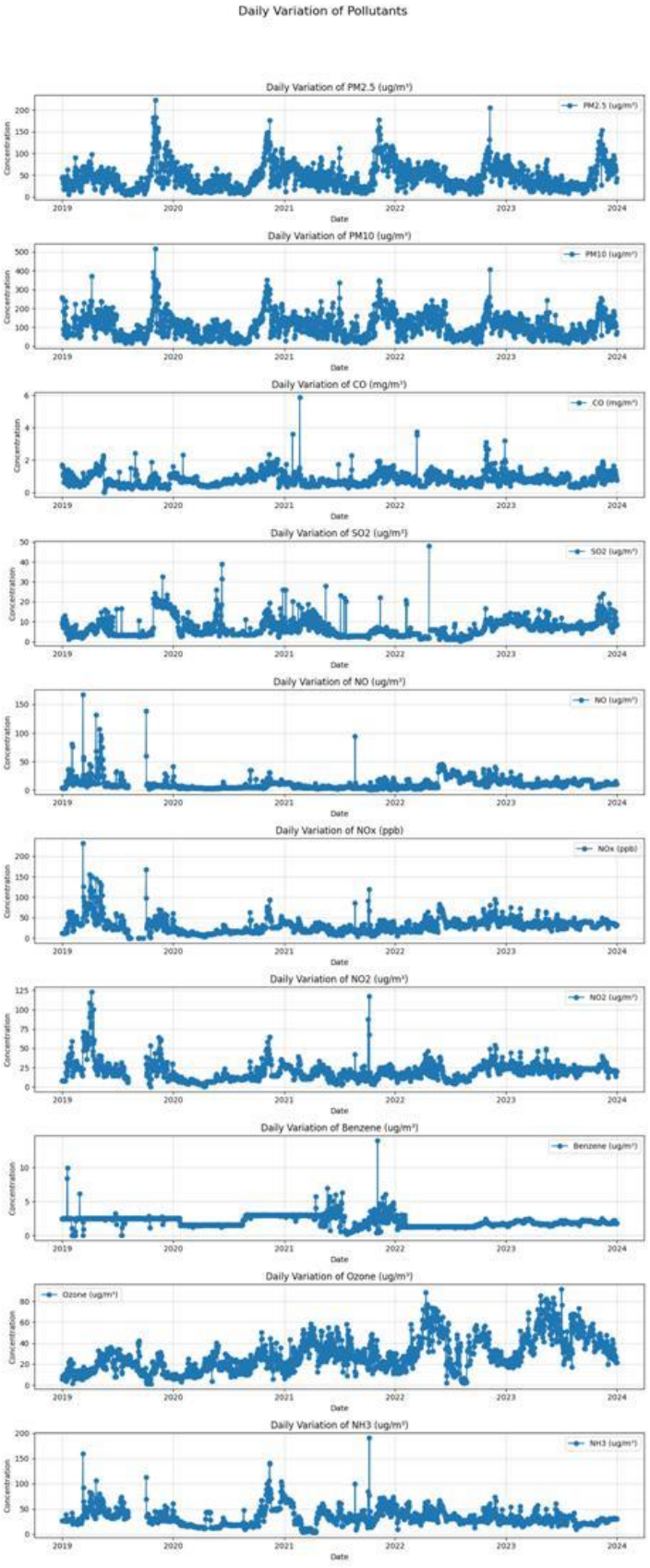

Figure-4: Daily Variation of Concentration of Different Pollutants

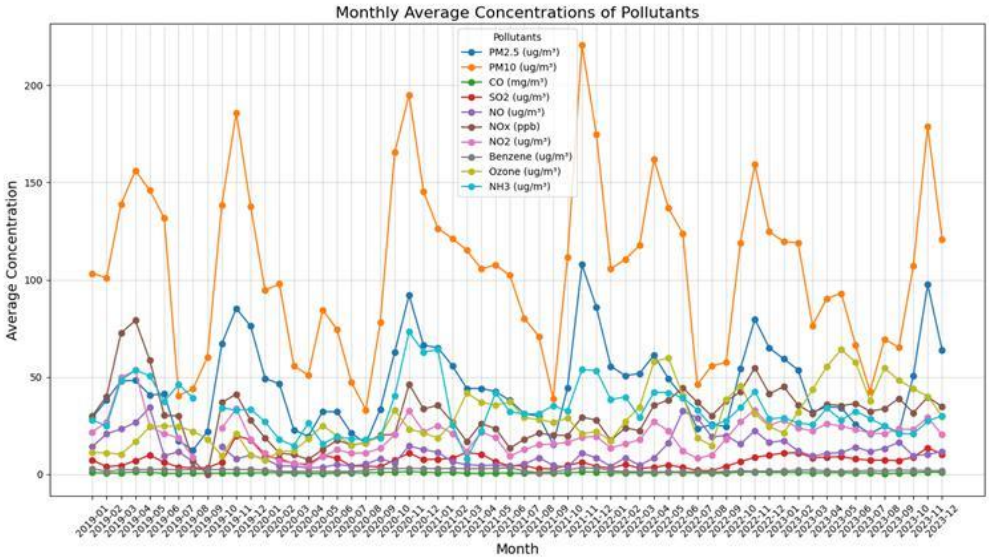

Figure-5: Monthly Average Concentrations of Pollutants

Figure-4 and Figure-5 further demonstrate that AQI values closely track PM_{10} concentrations, highlighting the significant role of particulate matter in determining AQI levels. For the five-year period from January 2019 to December 2023, the mean concentrations of $PM_{2.5}$ and PM_{10} were recorded at 43 µg/m³ and 106 µg/m³, respectively. These averages exceed the Central Pollution Control Board (CPCB) of India's National Ambient Air Quality Standards (NAAQS), which set the annual permissible limit for PM_{10} at 60 µg/m³. Maximum concentrations during the study period reached 611.9 µg/m³ for $PM_{2.5}$ and 999 µg/m³ for PM_{10}, highlighting severe pollution episodes. Over the five-year span, $PM_{2.5}$ and PM_{10} levels surpassed their respective permissible limits by 7.5% and 6%, reflecting consistent exceedance of regulatory thresholds. These findings emphasize the persistent challenge of managing particulate matter pollution in the region and its critical influence on air quality and public health.

The elevated levels of particulate matter in Punjab, India, can be attributed to a mix of industrial, agricultural, and anthropogenic activities. A significant contributor to air pollution in the region is the widespread practice of stubble burning, which releases large amounts of particulate matter into the atmosphere. This agricultural activity generates pollutants, including fine particulate matter, carbon monoxide, and various hydrocarbons, contributing to poor air quality across the state.Additionally, emissions from industrial processes, such as those in textile factories, food processing units, and other manufacturing sectors, contribute significantly to particulate matter levels. Vehicular emissions from Punjab's extensive road networks also play a crucial role in deteriorating air quality. Furthermore, biomass burning for residential heating and cooking in rural areas adds to the particulate matter burden, particularly during the winter months.Natural factors, such as dust from unpaved roads and agricultural fields, also contribute to the concentration of airborne particulates. These diverse sources collectively make particulate matter a major concern for Punjab, emphasizing the need for targeted measures to address industrial emissions, agricultural practices, and transportation-related pollution.Biomass combustion is a common practice in urban and semi-urban areas, further elevating particulate levels. Road traffic also plays a critical role in particulate matter pollution, releasing emissions from vehicles and suspended dust from roads. Furthermore, the state’s metal and petroleum industries contribute significantly to air quality degradation, emitting pollutants during manufacturing and processing activities (Serrano Cardona & Muñoz Mata, 2013). Together, these factors create a multifaceted challenge for managing particulate matter levels in Punjab. The diverse sources of pollution highlight the need for targeted mitigation strategies addressing both industrial emissions and natural contributions to improve air quality in the state.

IQAir's 2021 air quality analysis provided a comprehensive overview of pollution levels across 7,323 cities in 131 nations, union territories, and geographical regions. According to the report, the global annual average

concentration of $PM_{2.5}$ was 58.1 µg/m³. This far exceeds the World Health Organization (WHO) guideline of less than 5 µg/m³ and also surpasses the Indian Air Quality regulation limit of 40 µg/m³. India's air quality, in particular, remains a critical concern. The average $PM_{2.5}$ level in the country was reported to be 53.3 µg/m³, which is 11 times higher than the WHO's recommended threshold. Delhi, the nation's capital, recorded an annual $PM_{2.5}$ concentration of 85 µg/m³ - an alarming figure that is 17 times above the WHO's recommended limit. India continues to rank among the most polluted countries in Southeast Asia, as highlighted by the IQAir 2021 report. The analysis found that 12 of India's cities were listed among the top 15 most polluted cities worldwide. This underscores the severe and persistent air quality challenges faced by the country, necessitating urgent and targeted interventions to mitigate pollution levels and protect public health.

The fluctuation in air quality levels throughout the year can be attributed to atmospheric, meteorological, and temperature inversion effects (Khillare & Sarkar, 2012). During the summer months, warmer and denser air moves more rapidly, which facilitates the dispersion of pollutants. In contrast, during the winter season, temperature inversion occurs, trapping pollutants close to the ground and causing them to persist for longer periods. This phenomenon significantly contributes to increased concentrations of particulate matter in the atmosphere, making smog a prevalent issue, particularly in recent winters (Javed et al., 2021). Long winters, common in countries like India, exacerbate the problem by prolonging the duration of smog formation. These extended periods of smog not only degrade the surrounding air quality but also result in elevated AQI levels, posing significant risks to public health and the environment (Garg & Gupta, 2020). The AQI levels are also closely influenced by the amount of rainfall, as precipitation helps in settling particulate matter, thereby reducing air pollution. India’s seasonal variation is characterized by four distinct seasons:

- **Winter (December to February):** Marked by colder temperatures and increased smog due to temperature inversions.
- **Summer (March to June):** Characterized by warmer temperatures and better pollutant dispersion.
- **Monsoon (July to September):** Features high rainfall, which aids in reducing particulate matter concentrations.
- **Post-Monsoon (October to November):** A transitional period with reduced rainfall, during which air pollution begins to rise again (Bose and Roy Chowdhury, 2021).

These seasonal variations play a critical role in shaping the patterns of air pollution in India, highlighting the need for tailored interventions to mitigate the effects of atmospheric and seasonal factors on air quality.

Figure-6 compares AQI (Air Quality Index) with TOT-RF (Total Rainfall) to examine the relationship between rainfall and air quality. Although numerous factors influence AQI predictions, rainfall is a key environmental variable with a significant impact. Higher rainfall levels lead to a reduction in AQI, while lower rainfall results in higher AQI values. This inverse relationship highlights rainfall's role in mitigating air pollution by reducing particulate matter in the atmosphere. The analysis shows that Punjab receives substantially more rainfall from July to September compared to the other months of the year. Rainfall interacts with airborne pollutants, such as particulate matter, during precipitation, causing them to settle on the ground. This natural cleansing process reduces the concentration of pollutants in the air, leading to lower AQI values. Conversely, periods with little to no rainfall allow particulate matter and other pollutants to accumulate in the atmosphere, increasing AQI levels. The data analysis clearly demonstrates this relationship: higher rainfall corresponds to lower AQI levels, and vice versa. Additionally, rainfall contributes to surface water pollution when precipitation mixes with airborne pollutants and deposits them on the ground. Despite this, its role in improving air quality remains evident. The study's findings confirm a direct correlation between AQI, rainfall quantity, and climate, aligning with previous research by (Chandrappa & Chandra Kulshrestha, 2016). India experiences its heaviest rainfall during the monsoon season, spanning July to October. The winter months of November to February also see lower AQI levels, likely due to rainfall during the post-monsoon and winter periods. These seasonal patterns in precipitation contribute to the observed fluctuations in AQI, underlining the importance of climate and rainfall in air quality management.

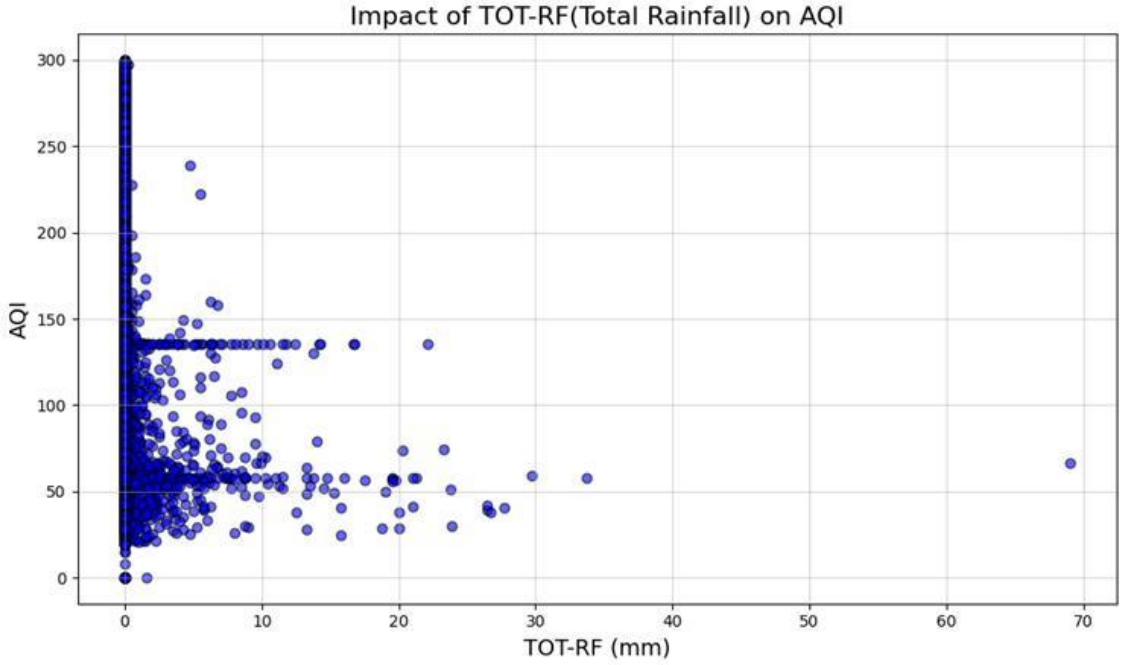

Figure-6: Impact of TOT-RF on AQI

GASEOUS POLLUTANTS

Error! Reference source not found. compares the concentrations of key pollutants, including CO, NO, NH_3, NO_2, NO_x, and SO_x, to analyse their impact on the Air Quality Index (AQI). Particulate matter and gaseous pollutants play a crucial role in AQI calculations. The analysis highlights that $PM_{2.5}$ follows a trend closely aligned with AQI throughout the year, indicating its significant contribution to AQI determination. Similarly, carbon monoxide (CO) also mirrors the AQI trend, underscoring its critical impact on air quality. The role of gaseous pollutants in shaping AQI categories is further detailed in **Error! Reference source not found.** and Table 2. The analysis shows that CO, NH_3, and SO_2 generally exhibit lower concentrations and fall within the "Good" AQI category (0–50) throughout most of the year. However, during winter months, nitrogen dioxide (NO_2) rises to levels that place it in the "Satisfactory" AQI category (51–100). This seasonal variation indicates that NO_2 significantly influences AQI during colder months, likely due to temperature inversions and reduced dispersion of pollutants. CO, in particular, follows the AQI trend consistently throughout the year, reinforcing its role as a major contributor to air quality levels. These findings confirm that while particulate matter, particularly $PM_{2.5}$, is the dominant factor in AQI determination, gaseous pollutants such as CO and NO_2 also play substantial roles, with their impact varying across seasons. This emphasizes the need for targeted interventions addressing both particulate matter and specific gaseous pollutants to improve air quality year-round.

Table 2: AQI Ranges fir different Pollutants

AQI Category	PM_{10} 24 hr	$PM_{2.5}$ 24 hr	NO_2 24 hr	O_3 24 hr	CO 8 hr	SO_2 24 hr	NH_3 24 hr	Pb 24 hr
Good (0-50)	0-50	0-30	0-40	0-50	0-1	0-40	0-200	0.05
Satisfactory (51-100)	51-100	31-60	41-80	51-100	1.1-2.0	41-80	201-400	0.6-1.0
Moderate (101-200)	101-250	61-90	81-180	101-168	2.1-10	81-380	401-800	1.1-2.0
Poor (201-300)	251-350	91-120	181-280	169-208	10.1-17	381-800	801-1200	2.1-3.0
Very Poor (301-400)	351-430	121-250	281-400	209-748	17.1-34	801-1600	1201-1800	3.1-3.5
Severe (401-500)	430+	250+	400+	748+	34+	1600+	1800+	3.5+

In recent years, pollutants such as benzene, ozone, xylene, and toluene have shown increasing significance, with ozone emerging as a notable contributor to air quality issues. Ozone concentrations tend to peak between November and February, coinciding with India's winter season. Unlike primary pollutants, ground-level ozone is

a secondary pollutant formed when sunlight interacts with volatile organic compounds (VOCs) and nitrogen oxides (NO_x). This photochemical reaction leads to ozone production, particularly in urban and industrial areas. Sources of VOCs include emissions from vehicles, the automotive sector, thermal power plants, bio refineries, chemical facilities, and various heavy industries. Punjab, with its industrial base, releases significant quantities of VOCs that interact with nitrous oxide to generate ozone (Manisalidis et al., 2020). During the winter, temperature inversions trap VOCs and other pollutants close to the ground, as the dense, cold air prevents their dispersion (Khillare & Sarkar, 2012). This phenomenon intensifies ozone production when sunlight reflects off the earth's surface and triggers a reaction between VOCs and NOx. In contrast, the behaviour of ground-level ozone during the summer is reversed. High temperatures in summer heat the air, allowing VOCs and NO_x to disperse, which reduces the formation of ground-level ozone. This seasonal variation highlights the complex interplay between meteorological conditions and pollutant behaviour. A similar pattern is observed with carbon monoxide (CO), further emphasizing the influence of seasonal factors on air pollutant concentrations.

Table 2 summarizes the annual average permissible limits for air pollutants contributing to AQI, as per the guidelines of the Central Pollution Control Board (CPCB) of India. The AQI is generally classified into six major categories: Good (0–50), Satisfactory (51–100), Moderate (101–200), Poor (201–300), Very Poor (301–400), and Severe (401–500). These categories reflect the cumulative impact of various pollutants, including ozone, $PM_{2.5}$, PM_{10}, VOCs, and CO, on air quality and public health. The findings underscore the need to address not only primary pollutants like particulate matter but also secondary pollutants like ozone, which are heavily influenced by industrial emissions and seasonal meteorological factors.

The temporal variations of pollutants from 2019 to 2023 are depicted in**Error! Reference source not found.**, illustrating how air pollution levels fluctuate over time. Month-to-month and year-to-year analyses reveal an overall upward trend in air pollution levels. Particulate Matter (PM), Carbon Monoxide (CO), and Ozone show seasonal patterns, with peak concentrations during winter due to temperature inversions and lower levels during the summer months (Khillare & Sarkar, 2012). The data highlights that in 2021, pollution levels for all measured pollutants increased significantly compared to prior years. This spike contrasts with 2020, where pollution levels were notably lower. This reduction in 2020 aligns with the impact of the COVID-19 lockdown, during which industrial and vehicular activities were drastically curtailed. In 2021, as heavy industries resumed continuous operations post-lockdown, air pollution levels surged, likely contributing to this noticeable increase. **Error! Reference source not found.** provides additional insights into the behaviour of specific pollutants such as benzene, xylene, and toluene. Unlike other pollutants, these substances do not exhibit a consistent temporal pattern. Their irregular concentrations suggest that their accumulation is closely linked to the state's industrial activities rather than natural or seasonal factors. This variability highlights the impact of industrial processes on the emission of volatile organic compounds (VOCs) and underscores the role of industrial operations in contributing to urban air pollution. These findings emphasize the importance of monitoring both seasonal and industrial influences on air quality. While temperature inversions and seasonal variations drive predictable fluctuations in pollutants like PM, CO, and ozone, the erratic behaviour of VOCs like benzene, xylene, and toluene points to the need for targeted strategies to address industrial emissions in Punjab.

In 2018, the Health Effects Institute (HEI) published a comprehensive report detailing the significant health impacts of air pollution on the population of India. According to the report, household burning, coal combustion, agricultural burning, anthropogenic emissions, transportation, diesel usage, and brick kilns were identified as the primary contributors to the release of major air pollutants. Estimates from 2015 indicate that air pollution was responsible for 10% of all deaths in India. The report attributed specific mortality figures to various sources of pollution: 0.169 million deaths resulted from coal combustion, 0.268 million from residual biomass burning, 0.1 million from dust, 0.06 million from agricultural burning, and 0.065 million from emissions related to transportation, brick kilns, and diesel usage. Alarmingly, projections suggest that by 2050, air pollution could cause up to 3.6 million deaths annually, an 84% increase compared to 2015 levels. Punjab, one of the most industrially and farming active states in North India, was highlighted in the report as having significant air pollution levels in 2019. The city recorded annual emissions of 47,800 tonnes of $PM_{2.5}$, 65,000 tonnes of PM_{10}, 3,250 tonnes of ozone (O_3), 182,100 tonnes of nitrogen oxides (NO_x), 188,550 tonnes of carbon monoxide (CO), 39,500 tonnes of volatile organic compounds (VOC), and 41,250 tonnes of sulphur dioxide (SO_2). Industrial emissions accounted for approximately 70% of these pollutants, originating from small, medium, and large-scale businesses in the region (Police et al., 2016). These findings underscore the critical role

of industrial activities in exacerbating air pollution and its associated health impacts, particularly in state like Punjab. The data highlights the urgent need for stringent pollution control measures targeting industrial emissions and other key sources. Without immediate intervention, the health burden from air pollution is projected to escalate significantly, threatening public health and environmental sustainability across India.

MACHINE LEARNING MODEL TO PREDICT AQI

To optimize performance, the hyper parameters of these models were fine-tuned to suit the dataset. The Grid Search technique was employed to identify the most effective hyper parameters, ensuring accurate predictions. Grid Search involves an exhaustive exploration of a predefined subset of the hyper parameter space for the chosen machine learning algorithms (Q. Wu & Lin, 2019). Table 3 summarizes these performance evaluation of the models on both the training dataset (representing 80% of the data) and the testing dataset (representing 20%). It provides insights into the accuracy and reliability of the models in forecasting AQI. Performance metrics such as Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), and R^2 were used to assess the models:

- **MAE** measures the average of the absolute differences between the actual and predicted values, providing a straightforward measure of error magnitude.
- **MSE** calculates the average of the squared differences between predicted and actual values. While MSE emphasizes larger errors, its values tend to be higher, necessitating the use of additional metrics.
- **RMSE** is the square root of MSE and is often preferred as it reduces the scale of error values, providing a more interpretable measure.

Table 3: Dataset Statistical Information for Air Pollutants

Index	PM2.5 (µg/m³)	PM10 (µg/m³)	NO (µg/m³)	NO2 (µg/m³)	NOx (ppb)	NH3 (µg/m³)	SO2 (µg/m³)	CO (mg/m³)	Ozone (µg/m³)	Benzene (µg/m³)	O Xylene (µg/m³)	Eth-Benzene (µg/m³)	MP-Xylene (µg/m³)	RH (%)	WS (m/s)	WD (deg)	RF (mm)	TOT-RF (mm)	SR (W/mt2)	AQI
Count	43543	43543	43543	43543	43543	43543	43543	43543	43543	43543	43543	43543	43543	43543	43543	43543	43543	43543	43543	43543
Mean	45.42	105.51	11.37	20.21	30.21	33.29	7.02	0.86	28.78	2.11	1.53	12.08	3.71	73.92	0.58	209.21	0.08	0.04	186.26	110.37
Std	33.09	72.48	15.92	15.87	25.02	20.93	7.34	0.57	22.87	1.38	0.81	21.97	3.98	25.24	0.31	76.83	0.27	0.82	159.88	53.69
Min	0.65	1.81	0.05	0.1	0	0.09	0.01	0	0.04	0	0.03	0.02	0.02	4.32	0.05	3.24	0	0	1.8	0
25%	22.1	54.25	4.35	10.75	15.48	21.06	3.44	0.5	11.89	1.5	0.87	1.7	1.42	54.59	0.27	141.02	0	0	100.25	66.19
50%	38.13	91.97	6.95	17.45	26.19	29.21	5.27	0.75	22.58	1.95	1.51	1.85	1.75	85.56	0.45	241.22	0	0	121.85	102.73
75%	59.82	138.185	13.71	24.79	37.75	39.08	8.45	0.96	38.75	2.5	2.25	5.29	7.35	95.45	0.7	267.02	0	0	182.89	135.65
Max	611.9	999.99	491.5	482.7	485.2	479.94	182.57	8.9	193.59	171.78	43.73	64.43	140.96	96.24	6.65	355.86	17.25	69	1305.95	300

It is essential to compare the performance of models across these metrics individually rather than directly comparing MSE, MAE, and RMSE. Furthermore, the values of MAE, MSE, and RMSE for the training and testing datasets should be comparable. A significant discrepancy between these datasets might indicate the presence of outliers or overfitting in the model. Thus, it is crucial to combine these error metrics with R^2 values to comprehensively evaluate model performance. Table 3 shows that Random Forest and CatBoost achieved the highest performance on the training dataset, with correlation coefficients of 0.9926 and 0.9997, respectively. The strong predictive accuracy of these models is attributed to their unique features:

Random Forest excels in determining feature importance by aggregating predictions from multiple decision trees. This ensemble approach reduces variance and enhances generalization.

CatBoost is specifically designed for high-speed performance on large datasets while delivering accurate predictions. It is particularly effective with categorical data and minimizes overfitting through advanced gradient boosting techniques.

Both models demonstrated excellent alignment between training and testing datasets, indicating their robustness and reliability in AQI prediction. The results confirm that Random Forest and CatBoost are highly effective for air quality forecasting, providing accurate and interpretable predictions for complex datasets.

IMPLICATIONS AND PERSPECTIVES

The study's findings highlighted that the Random Forest and CatBoost algorithms significantly outperformed other machine learning models, such as LightGBM, AdaBoost, and XGBoost, in accurately predicting AQI. While boosting algorithms like LightGBM and XGBoost are effective, they are susceptible to overfitting due to their reliance on tree-based methods. Additionally, parallelizing the training process in tree-based boosting

models can present challenges. In contrast, CatBoost incorporates advanced parameters specifically designed to minimize overfitting, enhancing its robustness in handling complex datasets. These results suggest that Random Forest and CatBoost could be effectively applied to other urban areas or regions with diverse air quality characteristics to evaluate their performance across various environmental scenarios. Their adaptability and predictive accuracy make them promising tools for studying air quality in different contexts.

Furthermore, the potential of these algorithms extends beyond research applications. Developing real-time AQI prediction systems based on Random Forest and CatBoost could provide policymakers with reliable and up-to-date information. Such systems would allow for proactive decision-making, enabling the timely implementation of measures to mitigate air pollution and protect public health. The scalability of these models also warrants further exploration. Investigating their application in regions with distinct pollution sources, meteorological conditions, and air quality challenges could reveal additional insights into their versatility and effectiveness. This study underscores the importance of leveraging advanced machine learning algorithms like Random Forest and CatBoost to build robust, scalable, and actionable AQI forecasting systems that contribute to better air quality management and planning.

CONCLUSION

The study analysed the forecast for Punjab Air Quality Index (AQI) for the period between 2019 and 2023. The results revealed a notable trend: AQI levels increased steadily in 2019 but experienced a significant decline in 2020 due to the nationwide lockdown implemented in response to the COVID-19 pandemic. This temporary reduction in AQI levels was attributed to reduced industrial, vehicular, and commercial activities. However, AQI levels began to rise again post-2020 as normal operations resumed. Among the factors influencing AQI, particulate matter - specifically $PM_{2.5}$ and PM_{10} - was identified as the most critical determinant of AQI values. In contrast, meteorological characteristics, such as temperature, humidity, and wind speed, had minimal impact on AQI predictions in this study. This underscores the dominant role of particulate pollutants in shaping air quality in Punjab. The machine learning models used in the study demonstrated high accuracy in predicting AQI levels. Random Forest and CatBoost emerged as the top-performing models, achieving maximum correlation coefficients of 0.9997 and 0.9926, respectively, for the training datasets. These results highlight the effectiveness of machine learning algorithms in forecasting AQI levels, showcasing their potential as reliable tools for air quality management. To extend the application of these models to other regions, further validation is required. Testing their performance under diverse air quality conditions - characterized by varying pollutant sources, climatic factors, and population densities - is crucial. Additionally, assessing the transferability of these models by applying them to different cities or countries will help evaluate their robustness and predictive capabilities across varied contexts. The study underscores the importance of leveraging machine learning algorithms like Random Forest and CatBoost for accurate AQI forecasting. These models not only provide valuable insights into air quality trends but also serve as critical tools for policymakers in implementing data-driven strategies to address air pollution.

REFERENCES

[1] Balakrishnan, K., Dey, S., Gupta, T., Dhaliwal, R. S., Brauer, M., Cohen, A. J., Stanaway, J. D., Beig, G., Joshi, T. K., Aggarwal, A. N., Sabde, Y., Sadhu, H., Frostad, J., Causey, K., Godwin, W., Shukla, D. K., Kumar, G. A., Varghese, C. M., Muraleedharan, P., … Dandona, L. (2019). The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: the Global Burden of Disease Study 2017. *The Lancet Planetary Health*, *3*(1), e26–e39. https://doi.org/10.1016/S2542-5196(18)30261-4

[2] Bao, R., & Zhang, A. (2020). Does lockdown reduce air pollution? Evidence from 44 cities in northern China. *Science of the Total Environment*, *731*. https://doi.org/10.1016/j.scitotenv.2020.139052

[3] Bekkar, A., Hssina, B., Douzi, S., & Douzi, K. (2021). Air-pollution prediction in smart city, deep learning approach. *Journal of Big Data*, *8*(1), 1–21. https://doi.org/10.1186/s40537-021-00548-1

[4] Chandrappa, R., & Chandra Kulshrestha, U. (2016). *Air Pollution and Disasters* (pp. 325–343). https://doi.org/10.1007/978-3-319-21596-9_8

[5] Garg, A., & Gupta, N. C. (2020). The Great Smog Month and Spatial and Monthly Variation in Air Quality in Ambient Air in Delhi, India. *Journal of Health and Pollution*, *10*(27).

https://doi.org/10.5696/2156-9614-10.27.200910

[6] Gurjar, B. R., Ravindra, K., & Nagpure, A. S. (2016). Air pollution trends over Indian megacities and their local-to-global implications. *Atmospheric Environment*, *142*, 475–495. https://doi.org/10.1016/j.atmosenv.2016.06.030

[7] Guttikunda, S. K., Goel, R., & Pant, P. (2014). Nature of air pollution, emission sources, and management in the Indian cities. *Atmospheric Environment*, *95*, 501–510. https://doi.org/10.1016/j.atmosenv.2014.07.006

[8] Javed, A., Aamir, F., Gohar, U., Mukhtar, H., Zia-UI-Haq, M., Alotaibi, M., Bin-Jumah, M., Marc (Vlaic), R., & Pop, O. (2021). The Potential Impact of Smog Spell on Humans' Health Amid COVID-19 Rages. *International Journal of Environmental Research and Public Health*, *18*(21), 11408. https://doi.org/10.3390/ijerph182111408

[9] Khillare, P. S., & Sarkar, S. (2012). Airborne inhalable metals in residential areas of Delhi, India: distribution, source apportionment and health risks. *Atmospheric Pollution Research*, *3*(1), 46–54. https://doi.org/10.5094/APR.2012.004

[10] Langer, T., & Meisen, T. (2021). System Design to Utilize Domain Expertise for Visual Exploratory Data Analysis. *Information*, *12*(4), 140. https://doi.org/10.3390/info12040140

[11] Li, H., Fan, H., & Mao, F. (2016). A Visualization Approach to Air Pollution Data Exploration—A Case Study of Air Quality Index (PM2.5) in Beijing, China. *Atmosphere*, *7*(3), 35. https://doi.org/10.3390/atmos7030035

[12] Lord, D., Qin, X., & Geedipally, S. R. (2021). Exploratory analyses of safety data. In *Highway Safety Analytics and Modeling* (pp. 135–177). Elsevier. https://doi.org/10.1016/B978-0-12-816818-9.00015-9

[13] Mahesh, T. R., Vinoth Kumar, V., Muthukumaran, V., Shashikala, H. K., Swapna, B., & Guluwadi, S. (2022). Performance Analysis of XGBoost Ensemble Methods for Survivability with the Classification of Breast Cancer. *Journal of Sensors*, *2022*, 1–8. https://doi.org/10.1155/2022/4649510

[14] Malhi, G. S., Kaur, M., & Kaushik, P. (2021). Impact of Climate Change on Agriculture and Its Mitigation Strategies: A Review. *Sustainability*, *13*(3), 1318. https://doi.org/10.3390/su13031318

[15] Manisalidis, I., Stavropoulou, E., Stavropoulos, A., & Bezirtzoglou, E. (2020). Environmental and Health Impacts of Air Pollution: A Review. *Frontiers in Public Health*, *8*. https://doi.org/10.3389/fpubh.2020.00014

[16] Mishra, S., Mishra, D., & Santra, G. H. (2020). Adaptive boosting of weak regressors for forecasting of crop production considering climatic variability: An empirical assessment. *Journal of King Saud University - Computer and Information Sciences*, *32*(8), 949–964. https://doi.org/10.1016/j.jksuci.2017.12.004

[17] Police, S., Sahu, S. K., & Pandit, G. G. (2016). Chemical characterization of atmospheric particulate matter and their source apportionment at an emerging industrial coastal city, Visakhapatnam, India. *Atmospheric Pollution Research*, *7*(4), 725–733. https://doi.org/10.1016/j.apr.2016.03.007

[18] Ravindra, K. (2019). Emission of black carbon from rural households kitchens and assessment of lifetime excess cancer risk in villages of North India. *Environment International*, *122*, 201–212. https://doi.org/10.1016/j.envint.2018.11.008

[19] Ravindra, K., Singh, T., Pandey, V., & Mor, S. (2020). Air pollution trend in Chandigarh city situated in Indo-Gangetic Plains: Understanding seasonality and impact of mitigation strategies. *Science of the Total Environment*, *729*. https://doi.org/10.1016/j.scitotenv.2020.138717

[20] Rybarczyk, Y., & Zalakeviciute, R. (2021). Assessing the COVID-19 Impact on Air Quality: A Machine Learning Approach. *Geophysical Research Letters*, *48*(4). https://doi.org/10.1029/2020GL091202

[21] Schneider, T., & Wheeler-Kingshott, C. A. M. (2014). Q-Space Imaging: A Model-Free Approach. *Quantitative MRI of the Spinal Cord*, 146–155. https://doi.org/10.1016/B978-0-12-396973-6.00010-1

[22] Serrano Cardona, L., & Muñoz Mata, E. (2013). Paraninfo Digital. *Early Human Development*, *83*(1), 1–11. https://doi.org/10.1016/j.earlhumdev.2006.05.022

[23] Singh, R. P., & Chauhan, A. (2020). Impact of lockdown on air quality in India during COVID-19 pandemic. *Air Quality, Atmosphere and Health*, *13*(8), 921–928. https://doi.org/10.1007/s11869-020-00863-1

[24] Wu, L., Li, N., & Yang, Y. (2018). Prediction of air quality indicators for the Beijing-Tianjin-Hebei region. *Journal of Cleaner Production*, *196*, 682–687. https://doi.org/10.1016/j.jclepro.2018.06.068

[25] Wu, Q., & Lin, H. (2019). Daily urban air quality index forecasting based on variational mode decomposition, sample entropy and LSTM neural network. *Sustainable Cities and Society*, *50*(March), 101657. https://doi.org/10.1016/j.scs.2019.101657

[26] Zhang, Y., Zhao, Z., & Zheng, J. (2020). CatBoost: A new approach for estimating daily reference crop evapotranspiration in arid and semi-arid regions of Northern China. *Journal of Hydrology*, *588*. https://doi.org/10.1016/j.jhydrol.2020.125087