Research Article

International Journal of Advanced Technology and Engineering Exploration, Vol 12(123)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2023.10102602

A seasonal autoregressive technique for forecasting $PM_{2.5}$ trend in Delhi

Varsha P. Desai[1*], Priyanka P. Shinde[2], Kavita S. Oza[3] and Rajanish K. Kamat[4]
D Y Patil Agriculture & Technical University, Talsande, Kolhapur, Maharashtra, India[1]
Government College of Engineering, Karad, Maharashtra, India[2]
Department of Computer Science, Shivaji University, Kolhapur, Maharashtra, India[3]
Dr. Homi Bhabha State University, Mumbai, Maharashtra, India[4]

Received: 17-December-2023; Revised: 28-January-2025; Accepted: 09-February-2025

Abstract
Delhi's air pollution is a severe and persistent environmental problem that frequently reaches dangerous levels. Alarmingly high particulate matter ($PM_{2.5}$) levels are commonly observed in Delhi, particularly during winter. The World Health Organization (WHO) has set acceptable guidelines for $PM_{2.5}$ concentrations at 10 µg/m³ for an annual average and 25 µg/m³ for a 24-hour average. However, $PM_{2.5}$ concentrations in Delhi's air often significantly exceed these limits. $PM_{2.5}$ pollution has a profound impact on both the environment and public health worldwide. It is a major contributor to respiratory and cardiovascular diseases, lung cancer, breathing problems, and strokes. Additionally, it harms the environment by causing acid rain, damaging vegetation, and reducing visibility. Accurate forecasting of $PM_{2.5}$ concentrations is therefore essential for implementing health precautions, protecting public health, and providing decision-making guidelines to minimize exposure to its harmful effects. The seasonal autoregressive integrated moving average (SARIMA) model is a robust, adaptable, and precise method for predicting time-series data with seasonal and non-seasonal components. It is particularly useful for real-time forecasting, especially in scenarios where seasonality plays a significant role. SARIMA effectively utilizes past data to predict future $PM_{2.5}$ trends. The proposed model provides valuable insights for forecasting $PM_{2.5}$ concentrations in Delhi. Air quality index (AQI) data from 2015 to 2022 was used to train the model, which was developed using the SARIMA technique to predict trends for the next five years with 95% accuracy. This predictive capability enables informed decision-making to promote a sustainable environment and safeguard public well-being. By accurately forecasting $PM_{2.5}$ concentrations, the model supports proactive policy decisions and helps protect residents from the adverse effects of air pollution.

Keywords
$PM_{2.5}$ pollution, Air quality index (AQI), Air pollution in Delhi, SARIMA model, Time-series forecasting, Environmental health.

1.Introduction
Delhi, India's capital, is one of the most polluted cities in the world. The major reasons for heavy pollution in Delhi are vehicular emissions, rapidly growing construction activities, tremendous industrialization, and stubble burning [1]. The air quality index (AQI) is a significant parameter for measuring air pollution concentration in a specific geographical area. Various concentrations of air pollutants are considered when computing the AQI. It primarily includes particulate matter (PM) percentage, sulfur dioxide, nitrogen dioxide, carbon monoxide, and ground-level ozone. AQI is the standardized method of communicating air quality information to the people.

Today, air quality is regularly monitored in urban areas and published online, allowing people to access real-time information about pollution levels in their surroundings. This enables individuals to take precautionary measures, such as wearing protective masks or limiting outdoor activities during periods of high pollution. Additionally, it raises public awareness about air pollution and encourages proactive efforts to mitigate its impact. By fostering transparency and promoting environmental sustainability, accessible air quality data plays a crucial role in protecting public health and driving collective action against pollution.

As per the survey, AQI ranges from approximately 0-500, where a more excellent AQI indicates more air pollution that may cause significant health issues.

*Author for correspondence

According to the Pollution Control Board PM is the major contributor in air pollution. It causes a severe impact on the human respiratory system [2]. $PM_{2.5}$ negatively impacts the environment in addition to being hazardous to human health. It damages ecosystems, contributes to air pollution, and obscures vision. Predicting $PM_{2.5}$ levels is essential for preserving environmental health, promoting sustainable development, and lessening air pollution's financial and climatic effects. Therefore, accurate prediction of the PM index in air pollution is crucial to enhance air quality, protect the environment and reduce the economic cost of air pollution [3].

Several factors, such as the need for informed policymaking, the influence on the environment, public health concerns, and economic consequences, motivate research on air pollution. Heart attacks, lung cancer, bronchitis, asthma, and other respiratory and cardiovascular conditions are all mostly brought on by air pollution. Understanding the extent of these impacts and developing strategies to mitigate them is critical. Rapid urbanization and population growth in cities like Delhi have exacerbated air pollution [4]. To comprehend these processes and create pollution-reduction strategies for urban planning research is essential. The scientific foundation for air quality norms and regulations is provided by research, which aids governments in putting various plans and policies into action to safeguard the environment and public health.

Seasonal autoregressive integrated moving average (SARIMA) is specifically designed to handle time series data with seasonal patterns, where the data shows a repeating pattern at regular intervals. It successfully captures and models the seasonality in the data by combining moving averages, differencing, and seasonal autoregression. SARIMA fits the model to the data's unique seasonal and non-seasonal behaviour, typically yielding more accurate forecasts than conventional models [5].

The research objective is to develop a time series model for early prediction of $PM_{2.5}$ concentration in Delhi based on historical data. Predicting $PM_{2.5}$ levels helps individuals and authorities take preventive measures to reduce exposure, especially for vulnerable populations such as children, the elderly, and those with pre-existing health conditions. Predictive models can help urban planners and policymakers anticipate areas at risk of high pollution levels, guiding land use, transportation, and infrastructure development decision to minimize exposure and improve air quality.

The study contributes to the existing literature by providing insights into air pollution in Delhi and its impact on the environment. Seasonal decomposition, the Interquartile method (IQR), residuals analysis, and seasonality trends were used to develop a time series model. It provides potential solutions for air pollution prediction and assists the government in generating strategies to suggest preventive measures in advance for a sustainable environment.

Various aspects of air pollution were highlighted in this paper, and an autoregressive technique was proposed for forecasting $PM_{2.5}$ index levels. This technique is a key step towards implementing strategies for emission reduction and adopting necessary air pollution control measures. This work aims to address existing issues by developing a fine-tuned model to accurately predict PM2.5 in Delhi.

The article is divided into several sections. Section 2 presents the literature review, providing a comprehensive overview. Section 3 discusses the materials and methods. Section 4 elaborates on the results, while Section 5 analyzes their impact. Finally, Section 6 presents the conclusion.

2.Literature review

As per the central pollution control board (CPCB) report in winter, the average $PM_{2.5}$ concentration in Delhi ranged from 115.7 - 235.6 micrograms per cubic meter (µg/m³) in the last few years. Now, the government of Delhi is taking essential steps to address the pollution problem, including motivation for public transportation, promotion for the use of electric vehicles, prohibition on the use of diesel-based equipment, etc. [6]. A graded response action plan (GRAP) has been implemented to control the intensity of air pollution levels, including halting erection activities, banning garbage burning and firecrackers, closing schools and colleges during periods of high pollution, and improving the air quality in the city [7].

By identifying areas with high pollution levels, policymakers can target resources and make efforts to take them to areas where they are most needed. It can result in more efficient resource use and reduced overall pollution levels. $PM_{2.5}$ and PM_{10} are the most prominent pollutants in Delhi [8]. It is an airborne suspension of solid particles and liquid droplets. It can cause significant respiratory issues, lung cancer, and heart disease [9]. PM contains tiny particles about 2.5 micrometers or less in diameter that may enter the bloodstream and penetrate deep into the lungs, causing' enduring health complications [10]. This air pollution harms people's health and the environment, like vegetation, soil, water bodies, and acid rain, which leads to biodiversity reduction in agricultural productivity [11]. It is critical to address Delhi's air pollution problem on a priority basis to safeguard public health and citizen welfare and promote sustainable development.

Air pollution data shows that over six years, around twenty-three heavily populated cities in India have been analyzed. It is observed that the people living in Delhi are 12% more susceptible to the health issues mentioned above issues than people from other cities. Feature selection is done through a correlation-based technique; further, it filters pollutants that affect AQI. Logarithmic transformations were used to identify skewed features, and exploratory data analysis (EDA) was used to uncover the dataset's hidden patterns [12]. A comparative study was performed on machine learning-based algorithms for AQI prediction with and without integration of synthetic minority oversampling technique (SMOTE) resampling techniques. The XGBoost model predicts the highest accuracy, whereas the support vector machine (SVM) model predicts the lowest accuracy. Practically, all machine learning models demonstrated enhancements in the model testing phase, and the Gaussian naive Bayes (GNB) model produced the best target prediction results [13].

The study conducted by Kothandarman, focused on predicting a significant pollutant parameter, using meteorological data. The author employed various models, such as k-nearest neighbor (KNN), linear regression (LR) and random forest (RF), to forecast ambient parameters in air pollution. Each machine learning model's performance was calculated using statistical metrics such as correlation coefficient, mean square error (MSE), and root mean square error (RMSE). The proposed models accurately predicted air quality parameters, but there is scope for forecasting most pollutant parameters [14].

In this study, researchers discussed air quality forecasting with machine learning algorithms. The author emphasized the significance of air quality because it has a significant impact on the environment as well as human health. The article provided an overview of various machine learning methods, including decision tree (DT), regression models, and artificial neural networks (ANN), and how these algorithms can evaluate data from various sources, including weather patterns, pollution levels, and other environmental factors to categories air pollution parameters [15]. The study highlighted the implementation of machine learning practices for air quality extrapolation. They proposed an efficient optimization method using multiple task learning (MTL) with a regularization technique. This research reveals that consecutive hour-related regularization helps improve performance over the standard regression model. MTL framework can train multiple models simultaneously to learn more generalizable features [16]. By minimizing the number of model parameters and employing structured regularization techniques, the performance of the trained model is improved [17].

The accuracy of air pollution prediction is significantly improved by incorporating spatiotemporal and meteorological features. The researchers found that hybrid artificial intelligent models, which combine multiple algorithms and methodologies, are more effective in accuracy and stability, although they require more computational resources [18]. These models, including fuzzy logic, SVM, ANN, convolutional neural networks (CNN), and long short-term memory networks (LSTM), have been successfully used for air pollution prediction. The study also acknowledged the challenges of accurate and efficient models, mainly when dealing with complex and dynamic systems [19].

The dynamic and unpredictable nature of pollutants in space and time makes air quality prediction difficult. Regular air quality checking, exploration, and analysis are essential in developing countries like India due to the severe consequences of air pollution on human and animal health, plants, natural resources, monuments, climate, and the environment [20].

The world health organization (WHO) states that the increasing concentration of $PM_{2.5}$ is hazardous, as it can enter deep into the lungs and bloodstream, leading to several health problems such as breathing problems, lung and cardiovascular diseases. It is essential for the Delhi government, along with the support of policymakers and public health advocates, to implement strict control measures to reduce air pollution. Continuous efforts are essential for reducing $PM_{2.5}$ concentration in Delhi to protect public health and prevent future hazards [21, 22].

An in-depth examination of intelligent modelling approaches for time-series air quality forecasting is carried out, with ensemble learning and metaheuristic optimization techniques employed for analysis. This study addresses non-linearity, interactions, tracking, and identifying pollution hotspots. Various challenges include slow convergence, being stuck in a local optimal state, models designed for specific locations, chemical composition, and meteorological variables [23, 24].

A comprehensive evaluation of measurement studies, air quality sensors, and recommendations was done to manage indoor air quality (IAQ) effectively. The study observes the effects of indoor air pollution on human health and the financial costs associated with inadequate IAQ. The performance of these sensors varies greatly [25], and because local laws, infrastructure, and climates change from one place to another, some guidelines may not be appropriate or practicable everywhere [26].

Development of an internet of things (IoT) based system to track air quality and integrate with a mobile application that promotes awareness and health protection by assisting users in tracking and forecasting pollution levels along their travel routes. A three-phase monitoring system measures dangerous gases and shows the air quality in parts per million (PPM) and utilization using MQ2 and MQ7 sensors [27]. Determine dangerous gas concentrations with accuracy and show the air quality in PPM. Elaborates on current air quality updates and how the system supports environmental health initiatives. It can raise public knowledge of air quality [28] and assist authorities in making well-informed decisions regarding controlling air pollution. To ensure accuracy, they need to be calibrated regularly. Data security hazards could exist, just like with any IoT system. Measures should be taken to ensure the security of the data being transmitted and stored [29].

It is crucial to measure different air contaminants in various areas to continue maintaining high air quality. Forecasting future levels of air pollution is feasible using machine learning algorithms and historical data. In order to estimate air quality, researchers look various machine learning models, including RF, DT, and LR. For the AQI, the RF algorithm yields the most accurate prediction. Real-time monitoring of many air contaminants by the technology enables quick reaction to variations in air quality. The system is flexible and able to adjust varied conditions because it may be used to detect different types of air pollution in different locations [30]. However, certain types of pollutants may not be detected by the sensors that can gather pollution data. This underscores the importance of regular sensor calibration to ensure accuracy and reliability [31].

Examine different approaches based on machine learning and big data for air quality predictions. The study examines published research findings on assessing air quality through deep learning, DT, artificial intelligence (AI), and other techniques. For experiments, visual data from cameras are used to monitor and evaluate changes in PM_{10} levels over time. Backpropagation ANN is a genetic algorithm with ANN (GANN) used to predict SO_2 and NO_2 levels. The accuracy of the RF predictor for AQI is 81%. PM_{10} concentration is measured with a DustTrak meter [32]. This algorithm was calibrated using the regression method within significant data architecture. They monitor air quality while using the current surveillance system to observe the surroundings. Due to device malfunctions and connectivity issues, poor sensor data quality might impact the accuracy of the air quality assessment. There is a lack of integrated real-time, considerably extensive data-based air quality evaluation and monitoring environments for intelligent cities [33].

It discusses analyzing air quality using big data techniques, distinguished by multi-source heterogeneity, dynamic mutability, and spatial-temporal correlation [34]. Researchers classify approaches for air quality monitoring, forecasting, and traceability. This analyzes different models and algorithms employed in these procedures. In order to notify the public and initiate preventative pollution management actions, early warning systems are essential. During modelling with spatial-temporal series data, various challenges are posed, such as large data volumes, nonlinear correlation, high dimensions, and dynamic change [35].

An improved machine learning model for forecasting India's main cities' AQI is presented in this study. To improve prediction accuracy, it combines the DT algorithm with grey wolf optimization (GWO). Metrics like R-Square, RMSE, MSE, MAE, and accuracy are used to assess the model's performance. The model's accuracy is 88.98% in predicting Delhi's AQI is impressive [36], which could significantly improve air quality management. Occasionally, the convergence speed of the GWO method may be slow, resulting in a longer time to discover the best solution

[37]. Deep learning models, particularly Recurrent Neural Networks (RNN), LSTM, and gated recurrent units (GRU), were used to anticipate air quality based on pollution and meteorological time-series data from the AirNet dataset. To identify the optimal architecture and model parameters for precise air quality prediction, a thorough examination of various RNN models and their variants is conducted in this study. The analysis showed that the GRU model slightly outperforms the RNN and LSTM architectures for PM_{10} concentration prediction [38]. The system employs historical time series data to derive fuzzy rules and construct a neuro-fuzzy network for forecasting air pollutant concentrations and environmental factors. The system uses hybrid learning algorithms to optimize the neuro-fuzzy network's parameters [39].

ANN, in particular, is a machine learning technique that has gained popularity and effectiveness over more conventional techniques when modelling PM pollutants. The daily average of PM concentrations has been predicted using ANN, considering current air quality requirements and laws linked to meteorological factors. Predicting urban and PM_{10} concentrations is studied using five ANN models, a deterministic modelling system, and a linear statistical model [40]. The outcomes demonstrate that ANNs have higher prediction accuracy than other techniques. The daily average PM_{10} concentration for the day ahead is predicted using an ANN model.

Understanding trends, seasonality, and cycles in time-dependent datasets requires using of time series models, which examine and forecast patterns in data gathered over time These models aid in recognizing past patterns, facilitating precise prediction and preparation for upcoming occurrences or values. They are extensively utilized in economics, finance, environmental monitoring, and weather forecasting. Time series models enable academics and businesses to make data-driven decisions and adjust to shifting trends by identifying seasonal and cyclical patterns. *Table 1* aim to make time-based data's underlying structure and dependencies visible for perceptive analysis and forecasting.

Table 1 Comparison between time series models

Objective	Other Time series models	SARIMA
Handling seasonality	Strongly seasonal trends in data can cause problems for ARIMA models, which do not explicitly account for seasonality [41].	Time series data with seasonal components are mainly modeled and forecasted using SARIMA [42].
Flexibility	Non-seasonal data may not get the same results from Holt-Winters exponential smoothing, which is mainly intended for seasonal data [43].	SARIMA models can handle both seasonal and non-seasonal data [44].
Parameter customization	Basic exponential smoothing does not allow this degree of parameter modification, and it might not be flexible enough to represent intricate seasonal cycles [45].	Providing greater control over the modeling process [46].
Trends and Cycles	Moving average or simple exponential smoothing might capture trends but are only effective in capturing cyclic patterns if tailored explicitly for such tasks [47].	Well-suited for stationary data capture both long-term trends and seasonal cycles [48].
Statistical rigor	Machine learning techniques like neural networks and SVM might not offer the same statistical validation and interpretability as SARIMA [31].	SARIMA models have a strong statistical basis; they may be rigorously tested and validated [45].
Forecast accuracy	Naive techniques or simpler models that fail to represent the complexity of the data may result in less accurate forecasts [46].	For seasonal data, SARIMA frequently provides more trustworthy and accurate estimations. As a result, the model performs better overall, and forecast errors are decreased [49].
Computational efficiency	Even though deep learning models may capture complex patterns, they frequently need a large amount of processing power and could only be practical for real-time forecasting with substantial infrastructure [49].	SARIMA models are more complex than standard ARIMA models, but they are still comparatively computationally efficient [50].

$PM_{2.5}$ is the most dominant contributor to air pollution in Delhi. These pollutants cause risks such as respiratory issues and cardiovascular diseases. The review explores time series, machine learning models and big data techniques for better air quality forecasting [51]. However, researchers need to conduct more studies on precise air quality parameter prediction for specific cities in India. SARIMA model is computationally and statistically more efficient; therefore, implementation of SARIMA for more accurate $PM_{2.5}$ prediction has major scope for the researcher to contribute to a sustainable environment. Overall, $PM_{2.5}$ prediction is an essential metric for pollution control and plays a critical role in protecting public health and the environment. Considering this need, an autoregressive technique was applied to forecast the $PM_{2.5}$ index in Delhi.

3.Material and methods

Data preparation and preprocessing

The AQI dataset was collected from Kaggle and the AQI website covering 2015 to 2022. The concentration of PM2.5 in Delhi is severe, so it is necessary to predict future values that help the Delhi government take earlier precaution measures. Therefore, the $PM_{2.5}$ feature is selected from the dataset for analysis. The significant purpose of the study is to detect the sources of air pollution and develop models to forecast future air quality concentrations of $PM_{2.5}$ in Delhi. The dataset contained 29,532 records with 16 columns, and the authors selected 2,922 records from Delhi for the study. During data preprocessing, Python's isna() method was used to identify null values and return a Boolean mask with true values for null values and false values for non-null values. There were no missing values for $PM_{2.5}$ in the dataset. The dataframe.resample() method is used for frequency conversion and resampling of time series data. The IQR identifies outliers in the dataset. Sixty-two outliers were detected and removed from the dataset. This cleaned dataset is further used for model development. For model training, 70% of the records were used, and the remaining 30% were used for testing the model. *Table 2* depicts that according to observation, the average concentration of $PM_{2.5}$ from the year 2015 to 2022 is 178.24, the maximum concentration is 339.56, and the minimum concentration is 10.24.

Table 2 Statistical summary of dataset

$PM_{2.5}$	**Concentration** (µg/m^3)
Mean	178.25
Median	108.33
Max	339.56
Min	10.24

SARIMA model

SARIMA is a more advanced version of the autoregressive integrated moving average (ARIMA); adding seasonal components, enhances ARIMA's capabilities and improves its suitability for time series data exhibiting distinct seasonal trends, which is used for time series forecasting. The benefits of SARIMA over other time series analysis techniques are depicted in *Table 1*.

SARIMA technique is designed to handle time series data that reveals seasonal patterns. The model uses three primary parameters, namely p, d, and q, representing the time series' autoregressive, differencing, and moving average components. Additionally, there is another parameter called m, which means the number of observations per season. This parameter determines the lag between seasonal observations and is set to 12 for monthly data during model development. The model parameters can be estimated using maximum probability estimation or other optimization methods, and the model can be used to generate forecasts for future periods. *Figure 1* shows the stepwise approach for SARIMA Model.

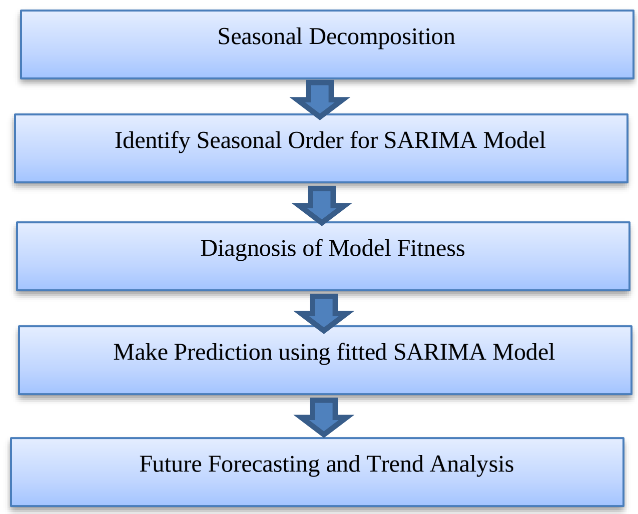

Figure 1 Stepwise approach for SARIMA model

In the time series analysis, the seasonal decomposition technique separates time series data into different components like trend, seasonality, and residual (noise). This process supports understanding patterns in the data, making forecasting and analyzing trends easier. The seasonal component captures the repeating patterns or cycles in the data at regular intervals, such as monthly, quarterly, or yearly. The residual plot is used to identify the noise. RSME is the difference between observed and predicted values used to determine model fitness and further train data for future forecasting and trend analysis (Equation 1).

It is expressed as SARIMA $((p, d, q) \times (P, D, Q)_m)$ (1)

Where p: non-seasonal autoregressive order

d: non-seasonal differencing order
q: non-seasonal moving average order
P: seasonal autoregressive order
D: seasonal differencing order
Q: seasonal moving average order
m: seasonal period for annual seasonality

EDA

To effectively control air quality, it is imperative to comprehend the trends and patterns in $PM_{2.5}$ concentration. This EDA compiles $PM_{2.5}$ concentration data using SARIMA modeling to project future $PM_{2.5}$ levels and evaluate potential increases over the coming years. *Figure 2* depicts the model decomposition and seasonability trend for $PM_{2.5}$ concentration from 2015 to 2022. It is revealed that the average level of $PM_{2.5}$ has been increasing since 2021. Visualizing the time series helps us to understand the underlying patterns, trends, and seasonality. Human activities and meteorological conditions such as temperature, wind speed, direction, humidity, and rainfall, significantly impact on concentration of $PM_{2.5}$ level which exhibits strong variation in the seasonal pattern. In Delhi, high $PM_{2.5}$ pollution was observed in the winter season from October to February. Both human and environmental factors are converging for severe air pollution. SARIMA has the capability to forecast future trends through handling these seasonal variations effectively.

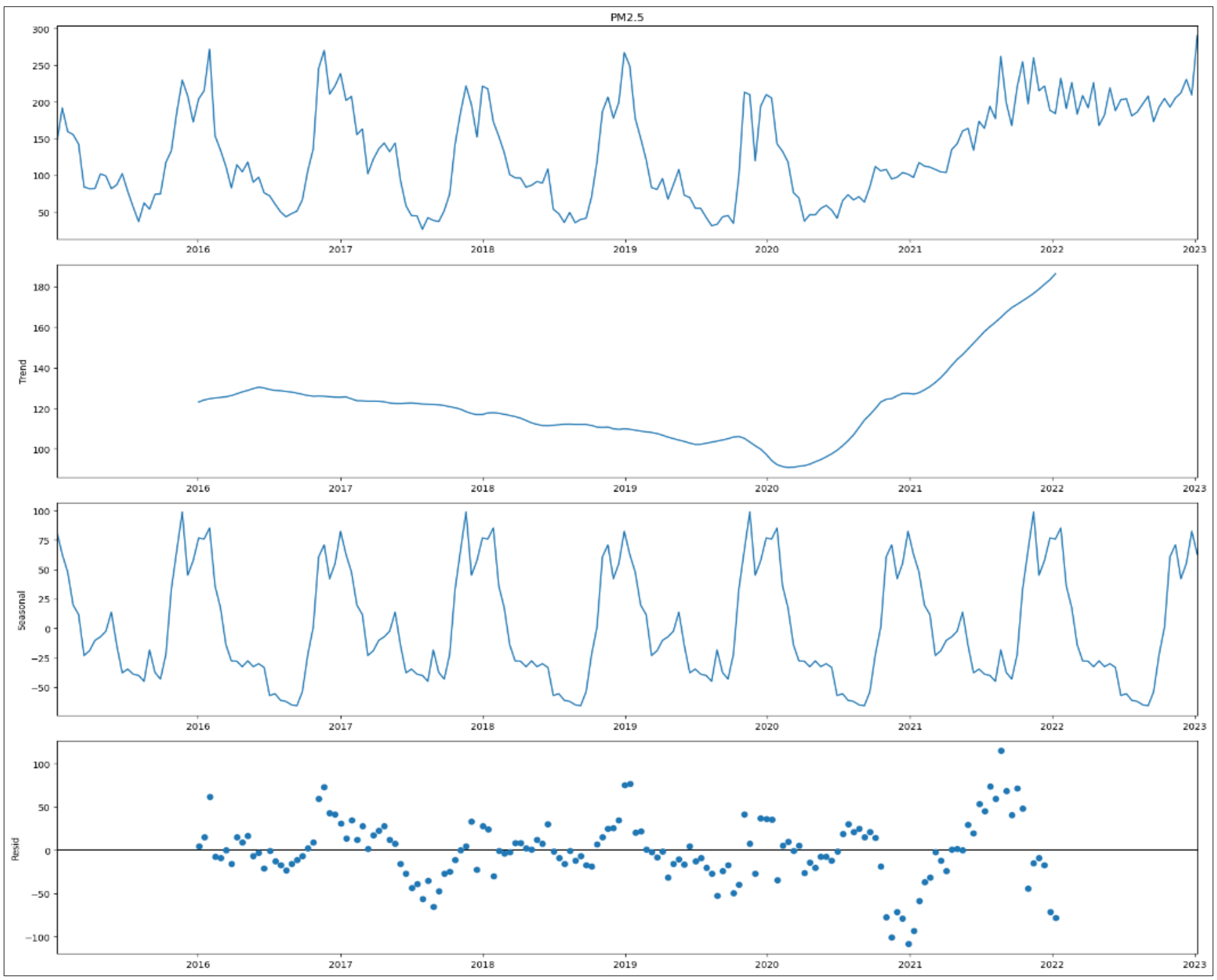

Figure 2 Model decomposition and PM2.5 seasonable trend analysis

Algorithm

Input: Historical AQI data from 2015 to 2022 with columns Date and AQI.
Output: 5-year AQI forecast with model accuracy.

1.Load libraries for data handling, visualization, and time series analysis dplyr, forecast, ggplot2
2. Import the AQI data.

3. Convert data in time-order format
4. For Each AIQ Value
    Handle Missing Values
    Remove Outliers and anomalies
5. Plot AQI over time to observe trends, seasonal patterns, and any outliers.
6. Analyze Stationary data
7. For Each Value
    Find lagging of data series
8. Train Model
    Apply function to determine optimal SARIMA parameters
    $((p, d, q)\times(P, D, Q)_m)$
    Fit SARIMA model on AQI data with identified parameters.
9. Use the SARIMA model to forecast AQI values for 5 years
10. Plot the forecast to display future AQI trends
11.Calculate residual r_t at time t is $r_t=O_t-F_t$,
    Where O_t is Observed value at time t, F_t is Forecasted value at time t.
11. Calculate root mean square error (RMSE) to assess accuracy:
    If validation data exists, compute RMSE on the validation set.
    If only training data is available, use in-sample predictions to calculate RMSE.
12. Output the 5-year forecast plot and display forecasted AQI values.

4.Results

Residuals are considered to measure the differences between observed values and predicted values from a model, and they are often analyzed to evaluate the capability of the model's fit. The histogram displays *Figure 3*, the frequency of residuals in different bins or intervals, while the estimated density plot provides a smooth curve that approximates the underlying probability density function. The red kernel density estimate (KDE) line closely follows the green N (0,1) line, indicating that the residual values are normally distributed. The residual diagnostics facilitate the evaluation of the model's fit to the data. The residuals reveal that how observed and predicted values differ, and should resemble white noise, as the proposed model has successfully identified all of the systematic patterns in the data.

Figure 4 presents a Q-Q plot where the residuals are plotted against the expected $PM_{2.5}$ values from the dataset, following a standard normal distribution N(0,1). The blue dots in the plot align with the expected values' linear trend, indicating that the residuals are normally distributed. Any deviations from this trend may suggest non-normality or other distributional issues.

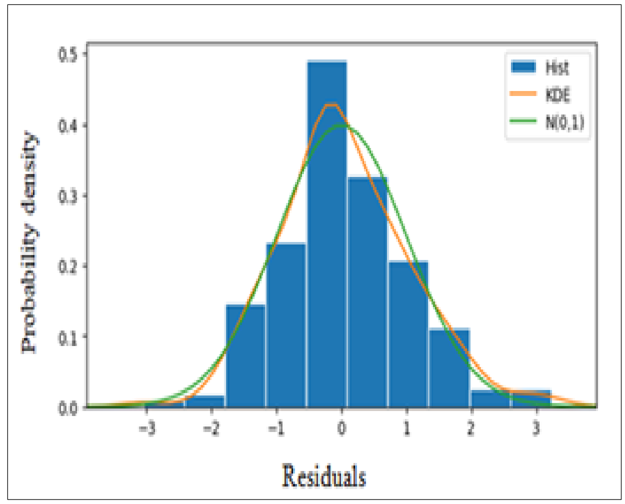

Figure 3 Histogram for estimated density

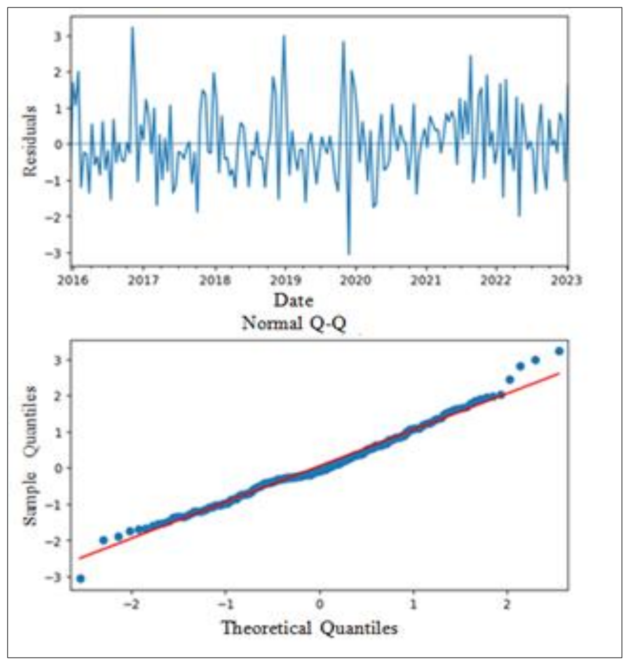

Figure 4 Standardized residual and normal Q-Q

The proposed model gives 95% confidence intervals with no autocorrelation, indicating that the residuals are approximately white noise. The Ljung-Box test yielded p-values greater than 0.05 for all tested lags, confirming that the residuals exhibit no significant autocorrelation. These results suggest that the SARIMA model has adequately captured the data’s temporal dependencies and seasonal structures.

Seasonal variations affect $PM_{2.5}$ concentration. The SARIMA model effectively captures this variation through in-depth residual diagnostics, performance metrics evaluation, cross-validation, and the inclusion of relevant exogenous variables, which is essential for generating reliable and actionable forecasts.

5.Discussion

A proposed model has been developed using SARIMA to forecast the average $PM_{2.5}$ concentration in Delhi, utilizing eight years of data from 2017 to 2022. *Figure 5* shows the $PM_{2.5}$ trends for actual values and sample predictions. A sensitivity analysis should be conducted by adjusting the SARIMA parameters and evaluating the model's output. By examining the results, determine which parameter combinations produce the best model performance. This analysis makes understanding the SARIMA model's sensitivity to variations in each parameter easier.

Figure 6 shows the model fit and $PM_{2.5}$ forecast in Delhi. The performance of the trained model was calculated using the RMSE metric. A value of 48.83 indicates a good fit between the model and the actual $PM_{2.5}$ concentration data. For this proposed model, 1.96×RSME gives 95% accuracy. The 95% confidence interval provides a range where the future value is expected to lie with 95% probability. *Figure 7* depicts the proposed model's outcome with past and future trends of $PM_{2.5}$ concentrations in Delhi. The trained autoregressive model has been forecast for average $PM_{2.5}$ concentrations for the next five years, which ranged between 175 to 250 μg/m.

This adaptive SARIMA model updates its parameters periodically based on current dynamic data. The concentration of $PM_{2.5}$ will increase in the upcoming years. The results of this proposed model will further help the Delhi government take precautionary measures and implement pollution control practices to reduce $PM_{2.5}$ concentration in advance.

Limitations

Complexity can arise in SARIMA models, particularly when dealing with large datasets or tuning several parameters (p, d, q, P, D, Q, m). This complexity can make the model more difficult to interpret and increase calculation times, especially when working with large time series or high-dimensional data.

A complete list of abbreviations is listed in *Appendix I.*

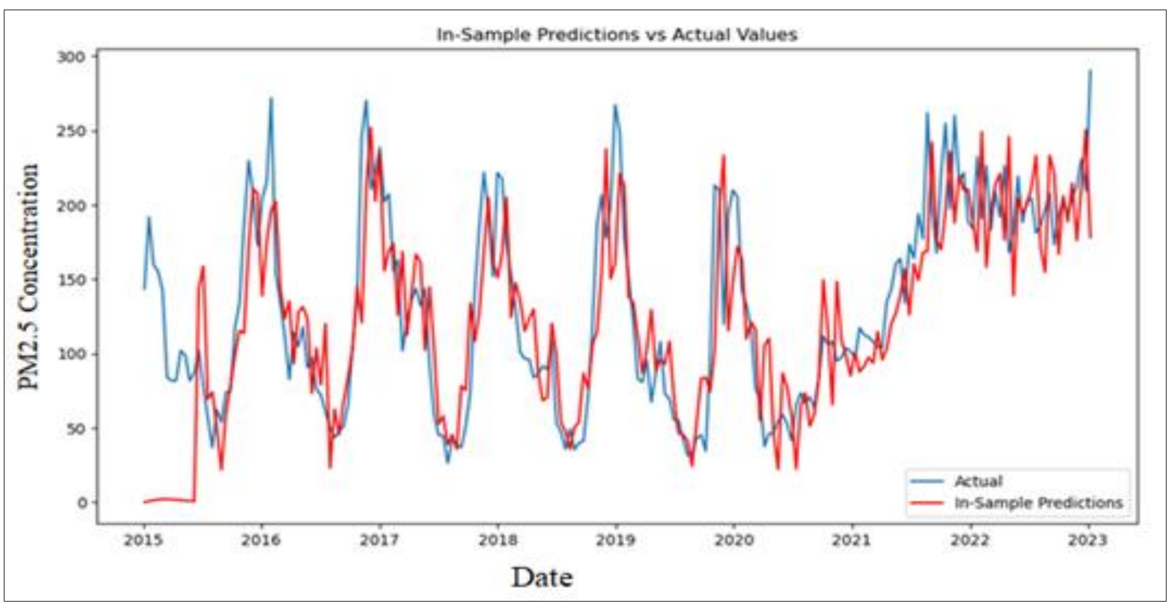

Figure 5 PM2.5 actual and predicted values

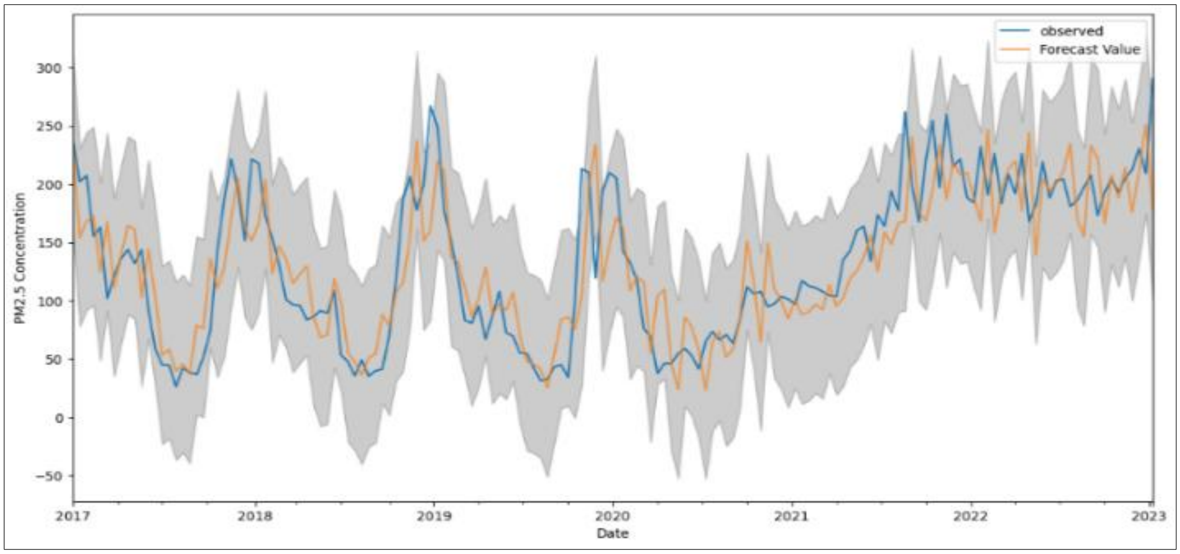

Figure 6 SARIMA for $PM_{2.5}$ forecast at Delhi

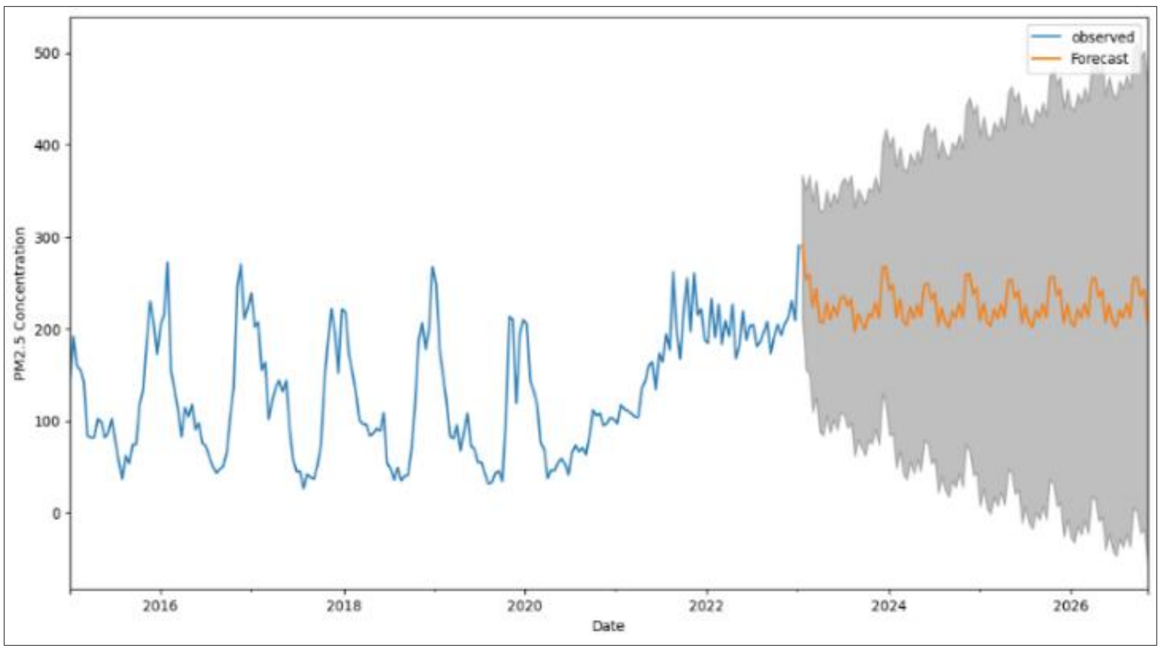

Figure 7 Predicted $PM_{2.5}$ concentration for next 5 Years (2023-2027)

6.Conclusion and future work

Accurately predicting $PM_{2.5}$ concentrations is essential for protecting public health and well-being, supporting sustainable development, and fostering economic growth. The trained SARIMA model has been used to forecast the average $PM_{2.5}$ concentration in Delhi for the next five years. The performance of the trained model was estimated using the RMSE metric. The computed results showed a minor error with an RMSE of 48.83, indicating that future $PM_{2.5}$ forecasting using trend analysis achieves 95% accuracy. This performance surpasses that of earlier autoregressive models. Continuous air quality monitoring and enhancement of the proposed model are essential to ensure its effectiveness and accuracy over time. Accurate prediction of $PM_{2.5}$ pollutants in Delhi helps the government and policymakers to take necessary steps to protect public health and ensure a sustainable environment. Therefore, implementing the proposed $PM_{2.5}$ forecasting model can contribute to future trend analysis and assist policymakers in improving air quality in Delhi through necessary control measures.

The proposed model offers valuable insights for future trend analysis. Further ensemble learning techniques could enhance the accuracy of the SARIMA model. An extended version of SARIMA, incorporating exogenous variables, will further improve predictive accuracy. Accurate $PM_{2.5}$ predictions will assist policymakers and public health officials in implementing essential measures to protect the public from the harmful effects of air pollution.

Acknowledgment
None.

Conflicts of interest
The authors have no conflicts of interest to declare.

Data availability
The data considered in this study is publicly accessible and sourced from Air Quality Data in India (2015 - 2022) (https://www.kaggle.com/datasets/fedesoriano/air-quality-data-in-india).

Author's contribution statement
Varsha P. Desai: Conceptualized, data collection, performed experiments, model development, analysis and interpretation of data, and wrote the manuscript. **Priyanka P. Shinde:** Contributed to formal analysis, review and editing. **Kavita S. Oza:** Contributed to data interpretation and reviewed the manuscript. **Rajanish K. Kamat:** Supervised the study, provided critical review and feedback on the manuscript.

References

[1] https://www.indiatoday.in/india/story/air-pollution-health-risk-india-annual-economic-cost-over-usd-150-billion-1928843-2022-03-24. Accessed 10 January 2023.
[2] Lu F, Xu D, Cheng Y, Dong S, Guo C, Jiang X, et al. Systematic review and meta-analysis of the adverse health effects of ambient PM2.5 and PM10 pollution in the Chinese population. Environmental Research. 2015; 136:196-204.
[3] https://www.msn.com/en-in/news/other/summer-action-plan-for-pollution-to-be-submitted-to-delhi-cm-arvind-kejriwal-on-april-24/ar-AA1a7DWL. Accessed 15 February 2023.
[4] www.cpcb.nic.in. Accessed 20 May 2022.
[5] Yin C, Liu K, Zhang Q, Hu K, Yang Z, Yang L, et al. SARIMA-based medium-and long-term load forecasting. Strategic Planning for Energy and the Environment. 2023; 42(2):283-306.
[6] Pant P, Shukla A, Kohl SD, Chow JC, Watson JG, Harrison RM. Characterization of ambient PM2.5 at a pollution hotspot in New Delhi, India and inference of sources. Atmospheric Environment. 2015; 109:178-89.
[7] https://www.outlookindia.com/,www.outlookindia.com/website/story/india-news-delhi-air-quality-will-continue-to-remain-bad-in-coming-winters-5-key-reasons/400980. Accessed 14 March 2021.
[8] Hill W, Lim EL, Weeden CE, Lee C, Augustine M, Chen K, et al. Lung adenocarcinoma promotion by air pollutants. Nature. 2023; 616(7955):159-67.
[9] Pope IIICA, Dockery DW. Health effects of fine particulate air pollution: lines that connect. Journal of the Air & Waste Management Association. 2006; 56(6):709-42.
[10] Singh T, Kaur A, Katyal SK, Walia SK, Dhand G, Sheoran K, et al. Exploring the relationship between air quality index and lung cancer mortality in India: predictive modeling and impact assessment. Scientific Reports. 2023; 13(1):1-12.
[11] https://www.indiaspend.com/pollution/india-has-9-of-worlds-10-most-polluted-cities-but-few-air-quality-monitors-792521. Accessed 14 March 2021.
[12] Selokar A, Ramachandran B, Elangovan KN, Varma BD. PM 2.5 particulate matter and its effects in Delhi/NCR. Materials Today: Proceedings. 2020; 33:4566-72.
[13] Kumar K, Pande BP. Air pollution prediction with machine learning: a case study of Indian cities. International Journal of Environmental Science and Technology. 2023; 20(5):5333-48.
[14] Kothandaraman D, Praveena N, Varadarajkumar K, Madhav RB, Dhabliya D, Satla S, et al. Intelligent forecasting of air quality and pollution prediction using machine learning. Adsorption Science & Technology. 2022; 2022:1-15.
[15] Sharma R, Shilimkar G, Pisal S. Air quality prediction by machine learning. International Journal of Scientific Research in Science and Technology. 2021; 8:486-92.
[16] Desai VP, Kamat RK, Oza KS. Rainfall modeling and prediction using neural networks: a case study of Maharashtra. Disaster Advances. 2022; 15(3):39-43.
[17] Zhu D, Cai C, Yang T, Zhou X. A machine learning approach for air quality prediction: model regularization and optimization. Big Data and Cognitive Computing. 2018; 2(1):1-15.
[18] Sharma E, Deo RC, Prasad R, Parisi AV. A hybrid air quality early-warning framework: an hourly forecasting model with online sequential extreme learning machines and empirical mode decomposition algorithms. Science of the Total Environment. 2020; 709:135934.
[19] https://weather.com/en-IN/india/pollution/news/2022-07-07-pollution-in-delhi-patna-5-times-higher-than-cpcb-safe-limits. Accessed 25 December 2022.
[20] Tiwari S, Kumar A, Mantri S, Dey S. Modelling ambient PM2.5 exposure at an ultra-high resolution and associated health burden in megacity Delhi: exposure reduction target for 2030. Environmental Research Letters. 2023; 18(4):1-12.

[21] Chae S, Shin J, Kwon S, Lee S, Kang S, Lee D. PM10 and PM2.5 real-time prediction models using an interpolated convolutional neural network. Scientific Reports. 2021; 11(1):1-9.
[22] Subramaniam S, Raju N, Ganesan A, Rajavel N, Chenniappan M, Prakash C, et al. Artificial intelligence technologies for forecasting air pollution and human health: a narrative review. Sustainability. 2022; 14(16):1-36.
[23] Raubitzek S, Corpaci L, Hofer R, Mallinger K. Scaling exponents of time series data: a machine learning approach. Entropy. 2023; 25(12):1-45.
[24] Ansari M, Alam M. An intelligent IoT-cloud-based air pollution forecasting model using univariate time-series analysis. Arabian Journal for Science and Engineering. 2024; 49(3):3135-62.
[25] Zhang H, Srinivasan R. A systematic review of air quality sensors, guidelines, and measurement studies for indoor air quality management. Sustainability. 2020; 12(21):1-38.
[26] Abhilash MS, Thakur A, Gupta D, Sreevidya B. Time series analysis of air pollution in Bengaluru using ARIMA model. In Ambient communications and computer systems 2017 (pp. 413-26). Springer Singapore.
[27] Shah HN, Khan Z, Merchant AA, Moghal M, Shaikh A, Rane P. IOT based air pollution monitoring system. International Journal of Scientific & Engineering Research. 2018; 9(2):62-6.
[28] Arif M, Katafygiotou M, Mazroei A, Kaushik A, Elsarrag E. Impact of indoor environmental quality on occupant well-being and comfort: a review of the literature. International Journal of Sustainable Built Environment. 2016; 5(1):1-11.
[29] Sokhi RS, Moussiopoulos N, Baklanov A, Bartzis J, Coll I, Finardi S, et al. Advances in air quality research–current and emerging challenges. Atmospheric Chemistry and Physics Discussions. 2021; 2021:1-33.
[30] Mane MV, Kumar D, Agarwal K. Detection and prediction of air pollution using machine learning and deep learning techniques. In international conference on computing, communication, and intelligent systems 2022 (pp. 145-50). IEEE.
[31] Liu X, Lu D, Zhang A, Liu Q, Jiang G. Data-driven machine learning in environmental pollution: gains and problems. Environmental Science & Technology. 2022; 56(4):2124-33.
[32] Mahendra HN, Mallikarjunaswamy S, Kumar DM, Kumari S, Kashyap S, Fulwani S, et al. Assessment and prediction of air quality level using ARIMA model: a case study of surat city, Gujarat State, India. Nature Environment & Pollution Technology. 2023; 22(1):199-210.
[33] Molaei SN, Salajegheh A, Khosravi H, Nasiri A, Abadi AR. Prediction of hourly PM10 concentration through a hybrid deep learning-based method. Earth Science Informatics. 2024; 17(1):37-49.
[34] Sarkar P, Saha M. Machine learning-based detection of sudden air pollutant level changes: impacts on public health. International Journal of Information Technology. 2024:1-9.
[35] Mehmood K, Bao Y, Cheng W, Khan MA, Siddique N, Abrar MM, et al. Predicting the quality of air with machine learning approaches: current research priorities and future perspectives. Journal of Cleaner Production. 2022; 379:134656.
[36] Ghazali S, Ismail LH. Air quality prediction using artificial neural network. In proceedings of the international conference on civil environmental engineering sustainability, Johor Bahru, Malaysia 2012 (pp. 1-15).
[37] Maleki H, Sorooshian A, Goudarzi G, Baboli Z, Tahmasebi BY, Rahmati M. Air pollution prediction by using an artificial neural network model. Clean Technologies and Environmental Policy. 2019; 21:1341-52.
[38] Natarajan SK, Shanmurthy P, Arockiam D, Balusamy B, Selvarajan S. Optimized machine learning model for air quality index prediction in major cities in India. Scientific Reports. 2024; 14(1):1-18.
[39] Athira V, Geetha P, Vinayakumar R, Soman KP. Deepairnet: applying recurrent networks for air quality prediction. Procedia Computer Science. 2018; 132:1394-403.
[40] Lin YC, Lee SJ, Ouyang CS, Wu CH. Air quality prediction by neuro-fuzzy modeling approach. Applied Soft Computing. 2020; 86:105898.
[41] Bhatt K, Sarangi RK, Kambli M. Comparative analysis of ocean color forecasting models: a case study on Indian southern Peninsula using Sarima and deep learning approaches. International Research Journal of Modernization in Engineering Technology and Science. 2023; 5(11):3167-75.
[42] Parviz L. Comparative evaluation of hybrid sarima and machine learning techniques based on time varying and decomposition of precipitation time series. Journal of Agricultural Science and Technology. 2020; 22(2):563-78.
[43] Huang W, Li T, Liu J, Xie P, Du S, Teng F. An overview of air quality analysis by big data techniques: monitoring, forecasting, and traceability. Information Fusion. 2021; 75:28-40.
[44] Guo Q, He Z, Li S, Li X, Meng J, Hou Z, et al. Air pollution forecasting using artificial and wavelet neural networks with meteorological conditions. Aerosol and Air Quality Research. 2020; 20(6):1429-39.
[45] Yu R, Yang Y, Yang L, Han G, Move OA. RAQ–a random forest approach for predicting air quality in urban sensing systems. Sensors. 2016; 16(1):1-18.
[46] Ukachukwu M, Uzoamaka N, Elochukwu N. Application of machine learning to air pollution studies: a systematic review. Journal of Energy Research and Reviews. 2023; 15(2):1-11.
[47] Terbuch A, O'leary P, Khalili-motlagh-kasmaei N, Auer P, Zöhrer A, Winter V. Detecting anomalous multivariate time-series via hybrid machine learning. IEEE Transactions on Instrumentation and Measurement. 2023; 72:1-11.

[48] Suri RS, Jain AK, Kapoor NR, Kumar A, Arora HC, Kumar K, et al. Air quality prediction-a study using neural network based approach. Journal of Soft Computing in Civil Engineering. 2023; 7(1):93-113.
[49] Mohamad AF, Jasin AM, Asmat A, Canda R, Ismail J, Soom AB. Sales analytics dashboard with ARIMA and SARIMA time series model. In 13th symposium on computer applications & industrial electronics (ISCAIE) 2023 (pp. 106-12). IEEE.
[50] Dubey AK, Kumar A, García-díaz V, Sharma AK, Kanhaiya K. Study and analysis of SARIMA and LSTM in forecasting time series data. Sustainable Energy Technologies and Assessments. 2021; 47:101474.
[51] Al-alola SS, Alkadi II, Alogayell HM, Mohamed SA, Ismail IY. Air quality estimation using remote sensing and GIS-spatial technologies along Al-Shamal train pathway, Al-Qurayyat city in Saudi Arabia. Environmental and Sustainability Indicators. 2022; 15:1-10.

Dr. Varsha P. Desai holds a Gold Medal in MCA and a Master's degree in Computer Applications. She earned her doctoral degree in Computer Science from Shivaji University. Currently, she serves as an Assistant Professor at D Y Patil Agriculture & Technical University, Talsande, Kolhapur, with 16 years of teaching experience. She has published over 45 research papers in national and international journals indexed by WOS, Scopus, and UGC. Dr. Desai is a member of the FOSSEE Club at IIT Bombay and has received recognition for her academic excellence and research contributions. Her research interests include Data Analytics, Artificial Intelligence, Data Mining, Cybersecurity, E-Learning, and Enterprise Resource Planning.
Email: varshadesai9@gmail.com

Dr. Priyanka P. Shinde is an Assistant Professor at Government College of Engineering, Karad. She earned her Ph.D. from Shivaji University, specializing in data analytics for the healthcare industry. With over 40 research papers published in esteemed national and international journals, including IEEE, Springer, and Scopus-indexed publications, she has made significant contributions to her field. She has also represented her institute multiple times at the India International Science Festival Young Scientist Conference. Her research interests include machine learning, data science, big data, and soft computing.
Email: ppshinde.gcek@gmail.com

Dr. Kavita S. Oza has been working in machine learning applications for over a decade. In the past five years, she has guided five research scholars in the fields of machine learning, artificial intelligence, and data science. Her contributions include the publication of 13 books in Computer Science and over 50 research articles in various national and international journals. She is also a life member of the Computer Society of India.
Email: kso_csd@unishivaji.ac.in

Prof. (Dr.) Rajanish K. Kamat is currently the Vice Chancellor of Dr. Homi Bhabha State University, Mumbai. With a passion for teaching and research, he has been a Professor in the Department of Computer Science and a distinguished university faculty member for over 25 years. His academic contributions include over 200 publications in reputed journals and conferences such as IEEE, Elsevier, and Springer, along with 16 reference books published by international publishers, including Springer (UK) and River Publishers (Netherlands). His books are available in more than 400 libraries worldwide. Dr. Kamat has successfully guided 20 Ph.D. scholars and secured research grants totaling ₹24 crores from funding agencies such as UGC, DST, and MHRD. He is a recipient of the Young Scientist Award under the Fast Track Scheme of the Department of Science and Technology. Among his recent initiatives are the establishment of the Faculty Development Centre (FDC) under the MHRD PMMMNMTT in Cyber Security and Data Sciences, the creation of a Centre of Excellence in VLSI System Design supported by RUSA, and coordination of two international projects—EQUAMBI and the Internationalized Master Degree Education in Nano Electronics—under the Erasmus+ programme. He is also a Fellow of the Institution of Electronics and Telecommunications Engineers.
Email: vc.drhbsu@gmail.com

Appendix I

S. No.	Abbreviations	Description
1	AQI	Air Quality Index
2	AI	Artificial Intelligence
3	CPCB	Central Pollution Control Board
4	CNN	Convolutional Neural Networks
5	DT	Decision Tree
6	EDA	Exploratory Data Analysis
7	GNB	Gaussian Naive Bayes
8	GANN	Genetic Algorithm with ANN
9	GRAP	Graded Response Action Plan
10	GRU	Gated Recurrent Unit
11	GWO	Grey Wolf Optimization
12	IAQ	Indoor Air Quality
13	IQR	Interquartile Method
14	IoT	Internet of Things
15	KDE	Kernel Density Estimate
16	KNN	K-Nearest Neighbor

17	LR	Linear Regression
18	LSTM	Long Short-Term Memory Networks
19	MSE	Mean Square Error
20	MTL	Multiple Task Learning
21	$PM_{2.5}$	Particulate Matter 2.5
22	PPM	Parts Per Million
23	RF	Random Forest
24	RNN	Recurrent Neural Network
25	RMSE	Root Mean Square Error
26	SMOTE	Synthetic Minority Oversampling Technique
27	SARIMA	Seasonal Autoregressive Integrated Moving Average
28	SVM	Support Vector Machine
29	WHO	World Health Organization