Open Access Research Journal of
Science and Technology
Journals home page: https://oarjst.com/
ISSN: 2782-9960 (Online)

OARJ OPEN ACCESS RESEARCH JOURNALS

(RESEARCH ARTICLE)

Integrating statistical and machine learning models for advancing air quality forecasting: A comparative time series analysis

Pranjali Subhash Sonone * and Abhay Kamlakar Khamborkar

Department of Statistics, Institute of Science, R. T. Road, Civil Lines, Nagpur, Maharashtra, India-440001.

Open Access Research Journal of Science and Technology, 2025, 14(02), 018-028

Publication history: Received on 03 June 2025; revised on 08 July 2025; accepted on 10 July 2025

Article DOI: https://doi.org/10.53022/oarjst.2025.14.2.0090

Abstract

Air pollution is a global issue, as it is known to cause various health problems, including skin irritation, respiratory issues, and lung diseases. The environment becomes contaminated due to poor air quality, and inhaling it harms our health. Therefore, it is essential to forecast air quality and take action to protect both humans and the environment in the modern world. In this study, air quality data from the Nagpur region are collected and studied. Several statistical techniques estimate an air quality index based on air pollution data. These statistical models are compared to determine the model's accuracy and precision. Traditional models (ARIMA, SARIMA, and Exponential smoothing) and machine learning techniques (SVM, LSTM, and Prophet) are employed to forecast AQI over time. R-squared, Root Mean Squared Error and Mean Absolute Error were then used to compare the prediction performance of each model. The results demonstrate that while conventional models may still capture linear trends and seasonal patterns, machine learning models outperform them in handling non-linear interactions and long-term dependencies. With the highest R-squared value and the lowest RMSE of any model, LSTM outperformed every model in predictive performance. It was observed that precision can be further increased by considering meteorological factors such as solar radiation, wind direction, wind speed, and relative humidity.

Keywords: Predictive Modelling; Time Series; Machine Learning; Random Forest; Air Quality Index

1. Introduction

Air quality forecasting is essential for public health, as it helps identify potential issues and take appropriate action. The rising level of air pollution has led to several diseases and fatalities. The number of cases of cardiovascular and respiratory conditions is increasing daily. According to a WHO estimate from 2021, millions of premature deaths and significant health problems like cardiovascular and respiratory disorders are caused by poor air quality. Therefore, public health must be protected, as air pollution is the second leading risk factor for non-communicable diseases. In urban areas with heavy traffic levels and large-scale industries, it is also noted that the air quality index has reached its limits. Additionally, Metropolitan Cities like Nagpur, a rapidly growing city, are grappling with the issue of increasing air pollution. Several statistical methods of time series analysis are employed to make the forecast, including machine learning methods such as SVM, Prophet, and LSTM, as well as conventional methods like ARIMA, SARIMA, and exponential smoothing. After modelling, metrics such as Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) were used to compare the performance of each model. The results will help statisticians develop an accurate model for predicting air pollution and enable environmental authorities to implement more effective pollution control measures.

Research has shown that deep learning models, such as LSTM, can capture complex temporal patterns in datasets, making them suitable for long-term forecasts. In contrast, time series models such as SARIMA are more ideal for short-term predictions[1]. It was also found that hybrid models using artificial neural networks (ANN) and fuzzy inference

* Corresponding author: Pranjali Sonone.

Copyright © 2025 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution Liense 4.0.

systems (FIS) demonstrated improved accuracy and stability[2]. Various studies have been conducted on the interactions between air pollutants and meteorological factors, showing that incorporating meteorological factors enhances predictive performance[3], [4], [5]. The Prophet Forecasting Model (PFM) was also used to predict pollutants such as PM2.5, NO2, and O3, improving forecasting accuracy. Several studies have demonstrated the effectiveness of hybrid models that combine traditional time series methods, such as ARIMA and SARIMA, with machine learning approaches, including Long Short-Term Memory (LSTM), Random Forest, and Support Vector Machines (SVM)[1], [6], [7]. Additionally, various studies suggest that integrating multiple forecasting approaches yields better prediction accuracy and robustness[8], [9] Research also found that deep learning models require large datasets and substantial computational power, whereas statistical methods, such as ARIMA, are easier to implement but may struggle to handle complex dependencies effectively[10], [11].

2. Materials and methods

2.1. Data Collection and Site Description

The Air Quality Index (AQI) data for the Nagpur region was collected from the Central Pollution Control Board's (CPCB) SAMEER software. A civil lines monitoring station was chosen because of urban pollution patterns caused by traffic, commercial activities, and industrial emissions. This study used the AQI record from January 1, 2017, to December 31, 2024.

2.2. Model Development

Various time series models were employed to predict the air quality index. These models were selected based on their ability to handle data features, long-term dependencies, seasonality, and nonlinearity. The variables used in this study include date, AQI, and meteorological variables such as temperature, wind speed, and relative humidity.

2.2.1. ARIMA Implementation

The time series forecasting technique known as ARIMA works well with stationary datasets that exhibit a regular pattern. It is made up of three parts: Moving Average (MA), Integrated (I), and Autoregressive (AR), which are indicators of a popular statistical technique for time series forecasting. The Autoregressive Integrated Moving Average (ARIMA) model works exceptionally well with stationary data that exhibits regular patterns. It consists of three components:

- Autoregressive (AR): Modelling of the relationship between observations and their past values
- Integrated (I): Differentiation to make the series stationary

Moving Average (MA) captures the relationship between an observation and past forecast errors. ARIMA is a good baseline model for air quality forecasting because it can be used to identify seasonal variations and linear trends.

Mathematical model

$$y_t = c + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots + \emptyset_p y_{t-p} + \theta_1 e_{t-1} + \emptyset_2 e_{t-2} + \cdots + \emptyset_q e_{t-q} + e_t$$

Where,

y_t: AQI value at time t,
c: Constant term,
$\emptyset_1, \ldots, \emptyset_p$: Coefficients for past AQI values (autoregressive terms),
$\emptyset_1, \ldots, \emptyset_q$: Coefficients for past forecast errors (moving average terms),
e_t: White noise (error),
p, d, q: Orders of autoregression, differencing, and moving average, respectively.

2.2.2. SARIMA

SARIMA extends ARIMA by adding seasonal components to the forecasting time series. As the AQI data contains seasonal patterns, the SARIMA model captures four major seasonal components: daily cycles (24-hour periods), Weekly patterns (7-day cycles), Monthly variations, and Annual seasonal changes.

Mathematical model

SARIMA extends ARIMA to handle seasonality, such as weekly or yearly patterns in AQI data.

Model equation

$$y_t = (ARIMA\ terms) + \Phi_{1y_{t-s}} + \ldots + \Phi_{1y_{t-P_s}} + \Theta_{1e_{t-s}} + \ldots + \Theta_{Qe_{t-Q_s}}$$

Where,

Φ, Θ: Seasonal autoregressive (SAR) and moving average (SMA) coefficients,
s: Seasonal period (e.g., $s = 7$ for weekly AQI patterns),
P, D, Q: Seasonal AR, differencing, and MA orders

2.2.3. Exponential Smoothing

Exponential smoothing is an effective method for time series forecasting because it assigns decreasing weights to previous observations, thereby giving more weight to recent observations. This method also considers that the most recent data point should be given greater weight because it is more informative than older values. When there are prominent trends or seasonality in the dataset, Simple ES gives historical observations exponentially decaying weights. The Holt-Winters ES forecasting method is another extension of ES that considers trends and seasonality in the data.

Mathematical model

Simple Exponential Smoothing: $\hat{y}_{t+1} = \alpha y_t + (1-\alpha)\hat{y}_t$, where α is the smoothing parameter, $(0 < \alpha < 1)$.

Holt's Linear Trend Method: $l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1})$, $b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$

$$\hat{y}_{t+h} = l_t + hb_t$$

Where l_t is the level, b_t is the trend, and β is the trend smoothing parameter.

Holt-Winters Seasonal Model

$$y_t = l_{t-1} + b_{t-1} + s_{t-s}$$

2.2.4. Support Vector Machines (SVM)

SVM is a practical supervised learning method that primarily handles regression and classification tasks. It can work with high-dimensional data and capture intricate patterns. This study uses SVM to model the relationships between pollutants and meteorological variables in forecasting air quality. SVM can also handle time series data by converting lag observations into input feature.

Mathematical model

Simple Exponential Smoothing: $\hat{y}_{t+1} = \alpha y_t + (1-\alpha)\hat{y}_t$, where α is the smoothing parameter, $(0 < \alpha < 1)$.

Holt's Linear Trend Method: Method: Method:$l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1})$, $b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$

$$\hat{y}_{t+h} = l_t + hb_t$$

Where l_t is the level, b_t is the trend, and β is the trend smoothing parameter.

Holt-Winters Seasonal Model

$$y_t = l_{t-1} + b_{t-1} + s_{t-s}$$

2.2.5. Long Short-Term Memory (LSTM)

LSTM, an advanced neural network (RNN) model, models sequential data with long-term dependencies. They overcome the gradient issues by using gating mechanisms.

- Forget Gate: Removes irrelevant past information.
- Input Gate: Updates the memory with new data.
- Cell State Update: Maintains long-term dependencies.
- Output Gate: Identifies the final output.

LSTM models are highly effective for forecasting air pollution data and capturing complex relationships between pollutants, weather conditions, and seasonal trends. Still, they require substantial training data, computational power, and careful hyperparameter tuning for optimal performance.

Mathematical model

$$\text{Forget gate: } f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_{f.}\right)$$

Where, f_t Forget gate activation at time t (values between 0 and 1),W_f:Weight matrix for the forget gate,b_f: Bias term for the forget gate,h_{t-1}: Hidden state from the previous time step, x_t: Input at the current time step,σ: Sigmoid activation function, ensuring the output is between 0 and 1.

$$\text{Input Gate: } i_t: \sigma(W_i[h_{t-1}, x_t] + b_i) \;\&\; \tilde{C}_t = tanh(W_C[h_{t-1}, x_t] + b_C)$$

i_t: Input gate activation at time t,
$\tilde{C}_t$:Candidate values for the cell state,
W_i, W_C: Weight matrices for the input gate and candidate cell state,
b_i, b_C: Bias terms for the input gate and candidate cell state.
Update Cell State $C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$

Where,

C_t: Updated cell state at time t,
C_{t-1}: Previous cell state,
$\odot$ Element-wise multiplication,
Output Gate: o_t: $\sigma(W_o[h_{t-1}, x_t] + b_o)$,

Where,

$h_t = o_t \odot tanh(C_t)$,
o_t: Output gate activation at time t,
h_t: Hidden state

2.2.6. Prophet

The Prophet model is a forecasting tool developed by Facebook to handle time series data with seasonal patterns. This model facilitates the handling of missing data and outliers. Prophet divides time series into three main components: Trend, which models long-term patterns using a piecewise linear or logistic growth function. Seasonality to capture recurring patterns (daily, weekly, yearly) using the Fourier series, Holidays/Events for incorporating special events (festivals, holidays)

$$\text{Mathematical model: } y(t) = g(t) + s(t) + h(t) + \in_t$$

Where,

$g(t)$: Trend function,
$s(t)$: Seasonal effects,
$h(t)$: Holiday effects,
$\in_t$: Error term

3. Results and Discussion

The correlation matrix represents the relationships between air pollutants, meteorological factors, and AQI values. Here, the colours range from deep red (strong positive correlation) to deep blue (strong negative correlation), indicating correlation levels. It can be observed that PM2.5 and PM10 exhibit a high positive correlation (r = 0.88), indicating a strong correlation with air pollution levels. There is a strong association of NO, NO2, and NOx with particulate matter. Meteorological factors exhibit a negative correlation with most pollutants. The AQI is highly correlated with PM2.5 (r = 0.73) and PM10 (r = 0.78).

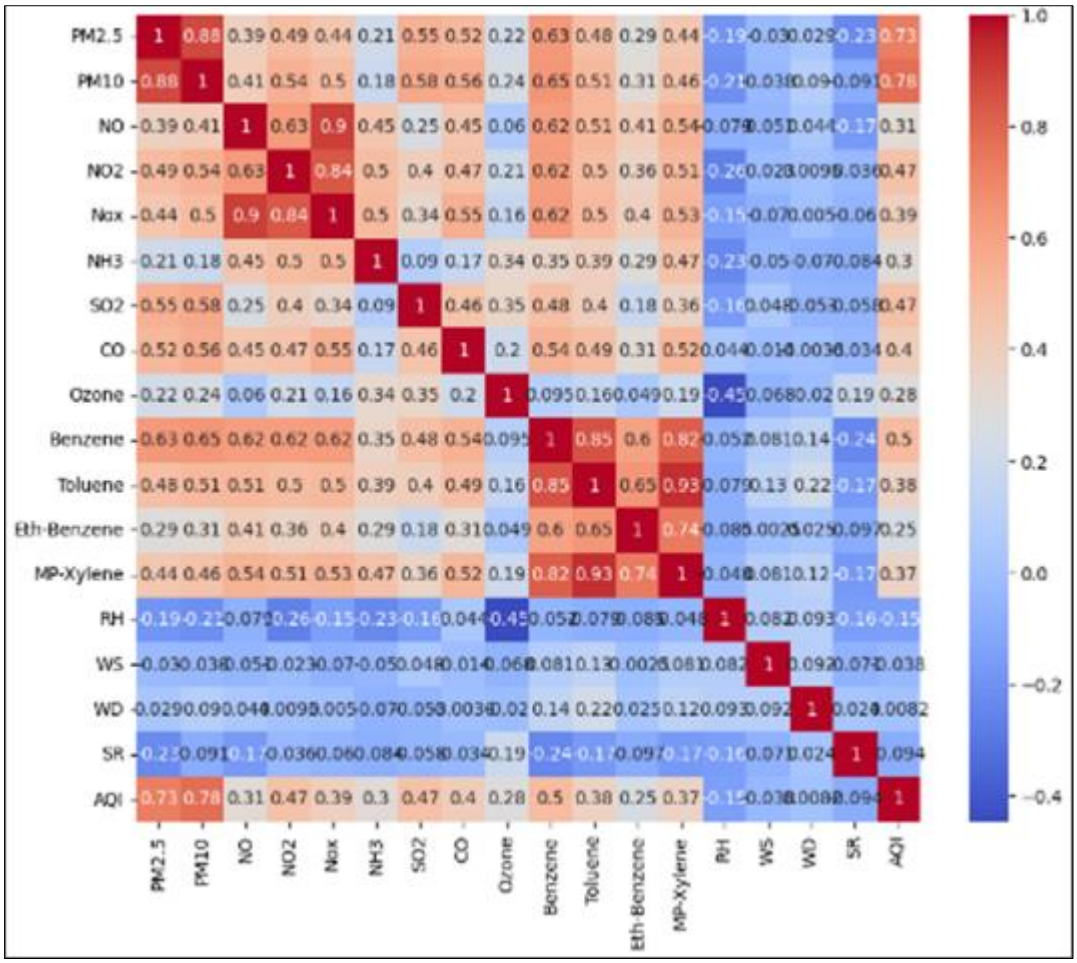


Figure 1 Correlation Matrix

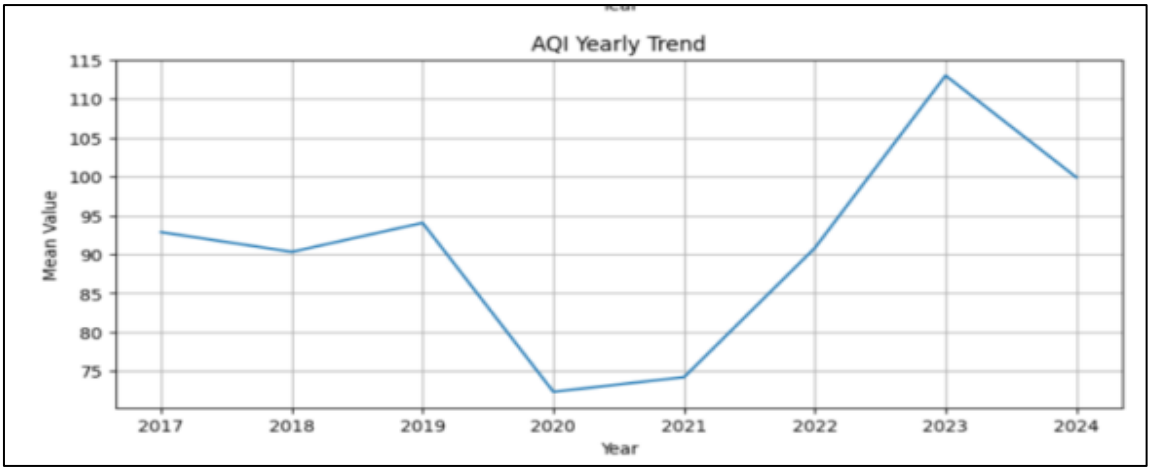


Figure 2 AQI Yearly Trend

The AQI trend from 2017 to 2024 shows that the AQI level decreased in 2020, probably due to reduced activities during the COVID-19 lockdown. After that, a sharp increasing trend was observed from 2021 onward, peaking in 2023. This trend indicates worsening air quality and the urgent need for targeted pollution control measures.

3.1. SARIMA

The AQI data, along with meteorological variables, were used to build a forecasting model, SARIMAX (1,1,1) X (1,1,1,12). Key performance metrics RMSE and MAE have values of 99.83 and 91.07, respectively, which show moderate predictive accuracy. Meteorological factors like Solar radiation (SR), wind speed (WS), and wind direction (WD) show no significant correlation with AQI except relative humidity (RH). The Autoregressive (AR) and moving average (MA) term highlights strong dependency with time. The seasonal moving average term also proves highly significant seasonal AQI fluctuations. The model diagnostics show that the residuals are non-normal since they fail the Jarque-Bera test for normality, but show no considerable autocorrelation (Ljung-Box test) or heteroskedasticity. The "Actual vs. Fitted AQI Values" plot shows a general agreement between forecasted and actual AQI trends, though residual analysis sometimes indicates some overestimation and underestimation. The residual histogram establishes a relatively symmetric distribution around zero, indicating unbiased forecasting. Following the initial lag, autocorrelation is negligible in ACF and PACF plots. However, because the predicted values are smoother than the actual data, the SARIMAX model performs poorly in representing AQI variability, suggesting that more fine-tuning or additional predictors are needed to improve accuracy.

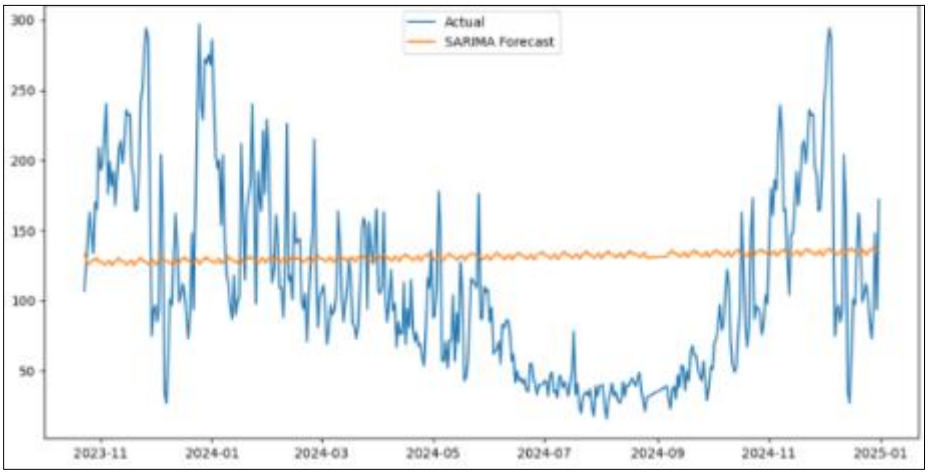


Figure 3 Sarima Forecast vs. Actual

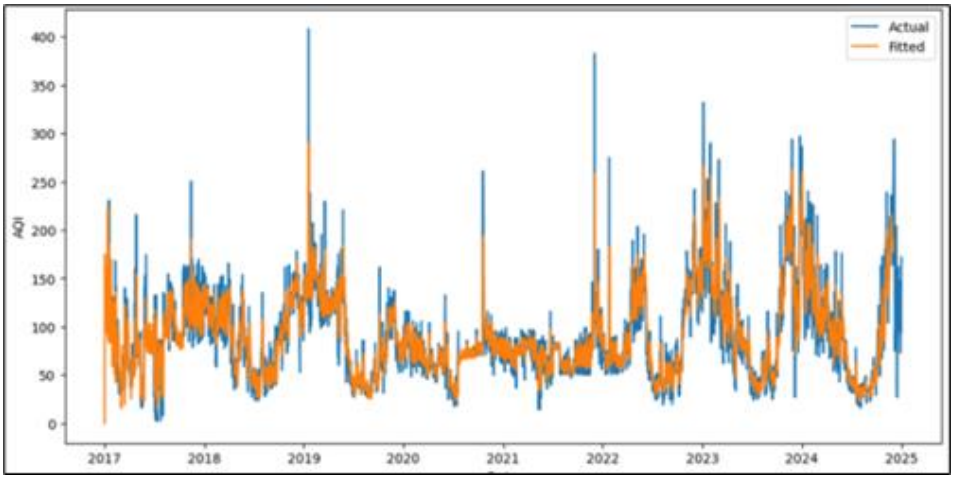


Figure 4 Actual vs. Predicted

3.2. Exponential Smoothing

The exponential smoothing models, such as SES, Holt's, and Holt-Winters', exhibit greater SSE (2.36×10^6, 2.16×10^6, and 2.17×10^6, respectively) and MSE (813.99, 744.82, and 748.00) than the SARIMAX model. Although their AIC and BIC values are lower (19,463 to 19,920), the fitted values from exponential smoothing cannot replicate the sharp changes in real AQI data. The predicted trend remains relatively smooth and upward. At the same time, SARIMAX includes seasonal trends, exogenous factors such as humidity, and short-run dependence, making it a stronger model for AQI prediction.

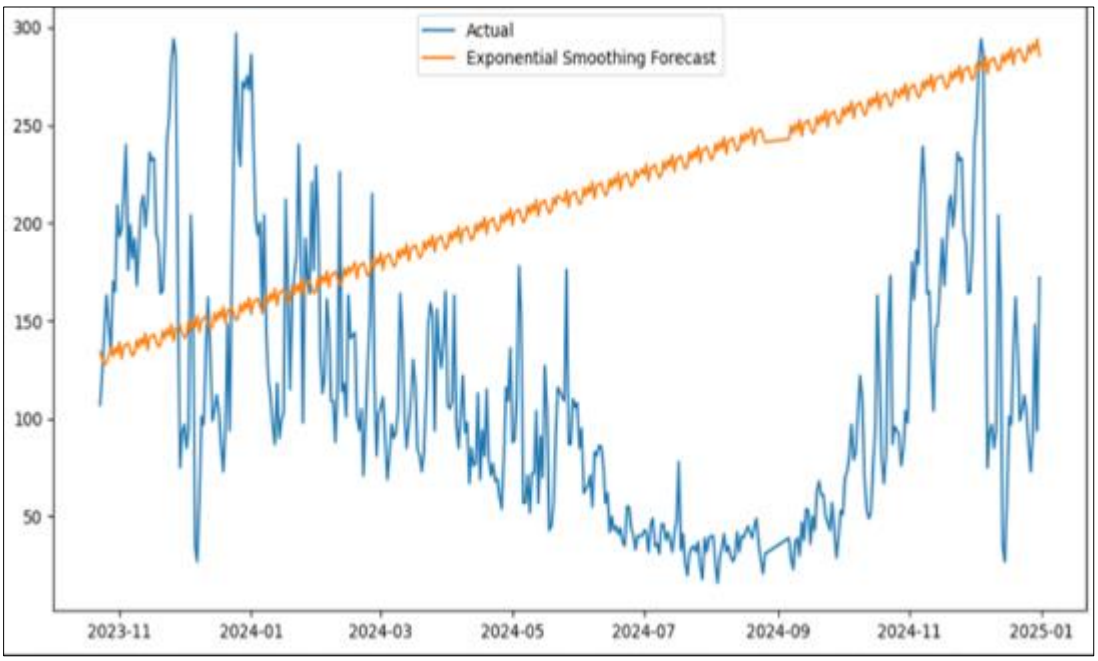


Figure 5 Exponential Smoothing Forecast vs. Actual

Exponential smoothing residuals have higher deviations and remove periodic variations, unlike SARIMAX residuals, which are more temporally dependent and provide a better fit. The line graph also confirms that SARIMAX fits the data better than exponential smoothing, which poorly explains the variation in AQI. SARIMAX(1,1,1)x(1,1,1,12) is the best choice for AQI forecasting because it has lower AIC, BIC, and MSE, makes more accurate predictions, and can accurately account for exogenous and seasonal factors.

3.3. Prophet

Although the Prophet model accurately captures seasonality and long-term trends in AQI forecasting, its evaluation metrics suggest that it could improve. The model shows notable differences between expected and actual values, indicating sensitivity to outliers, with an MSE of 1319.89 and an RMSE of 36.33. It is less resilient to short-term fluctuations because, on average, predictions deviate by this margin, as indicated by the MAE of 27.04. Furthermore, 41.07% of the AQI variance is explained, as indicated by the R-squared value of 0.41, meaning that a sizable portion remains unaccounted for. The model's overall performance could be improved using meteorological factors like RH, WS, WD, and SR as regressors to capture AQI variations better and increase predictive accuracy.

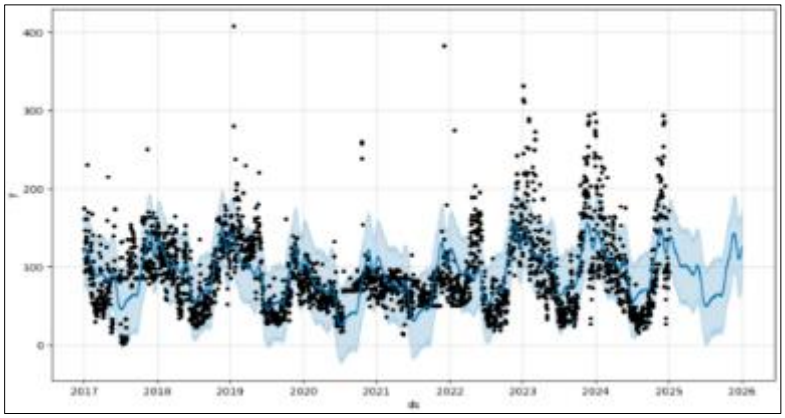


Figure 6 AQI Forecast (without meteorological variables)

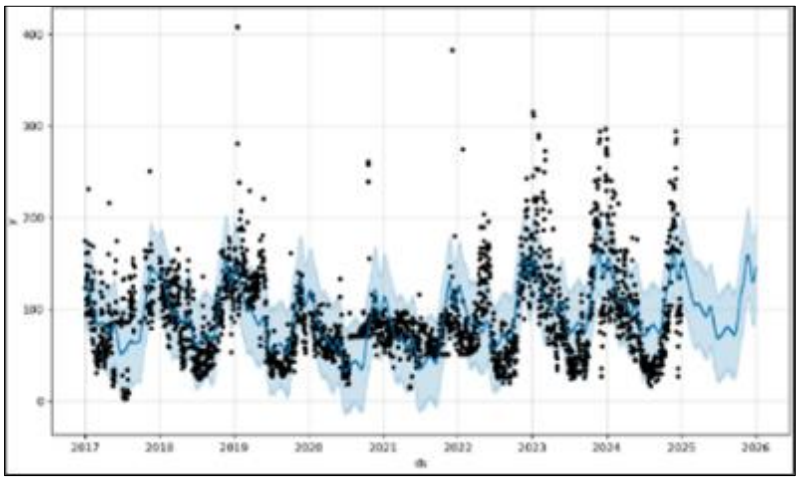


Figure 7 AQI Forecast (with meteorological variables)

The prophet model exhibits marginal improvement when additional regressors (RH, WS, WD, and SR) are included. The RMSE of 35.97 decreases average prediction error while the MAE falls to 26.8, indicating more accurate predictions. Moreover, the R-squared value improved to 0.436, indicating that the model can now explain 43.61% of the variance in AQI, an increase from the original 41.07%. The minor improvements suggest that incorporating meteorological variables captures a greater portion of the AQI variation and enhances the model's predictive utility. Additional enhancements could improve predictive accuracy, incorporating other pertinent features and more sophisticated modelling methods.

3.4. SVM

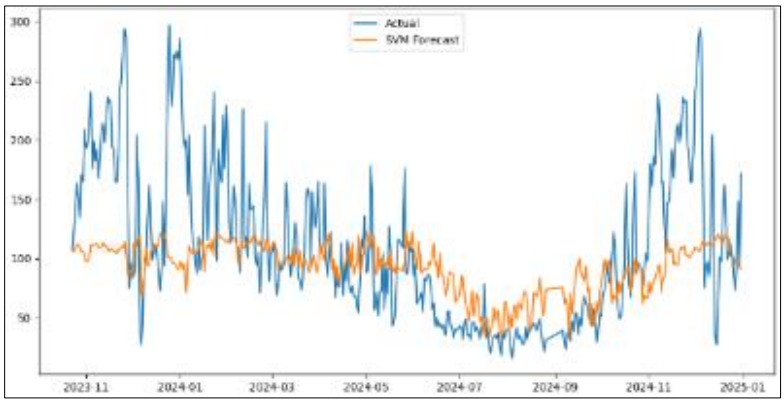


Figure 8 SVM Forecast vs. Actual

The Support Vector Machine (SVM) model predicts AQI with a moderate level of accuracy, but its forecast performance is limited. With an RMSE of 60.56 and an MAE of 43.47, the model exhibits a significant amount of error compared to previous models, such as Prophet, and deviates substantially from actual AQI values. Furthermore, the model shows that SVM explains only 15% of the variance in AQI, indicating that it does not adequately capture the underlying trends, as indicated by its low R-squared value of 0.15. These results suggest areas for improvement, such as adjusting hyperparameters, adding more pertinent features, or combining SVM with other machine learning algorithms to boost predictiveness.

3.5. LSTM

The Long Short-Term Memory (LSTM) model shows excellent predictive accuracy for AQI prediction, with an RMSE of 29.64 and an MAE of 20.55, both much lower than models such as SVM and Prophet. The very high R-squared value of 0.7751 means that the model explains around 77.5% of the variance in AQI, reflecting its capability to detect intricate patterns in sequential data. Moreover, the variance explained of 0.7800 also validates the model's effectiveness. Nevertheless, the MSE of 878.54 indicates residual error, which means there is scope for improvement. The LSTM model performs better than earlier methods and is a strong contender for time series forecasting.

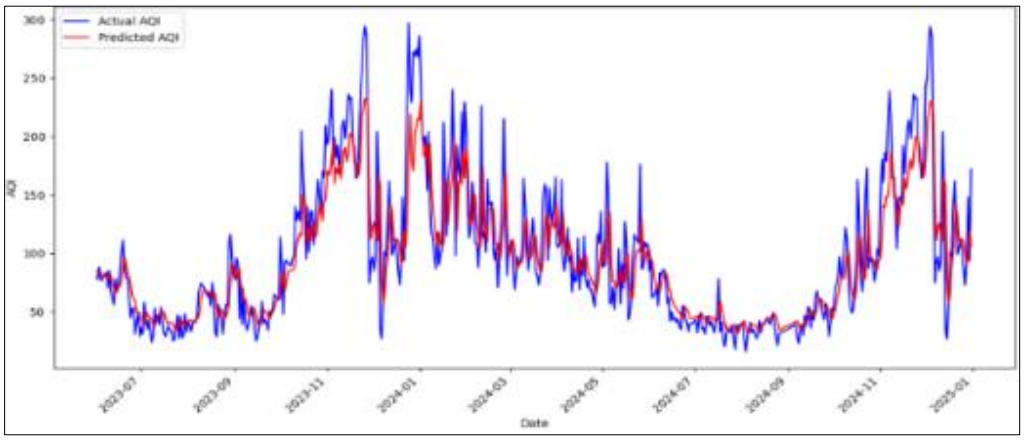


Figure 9 LSTM Forecast vs. Actual

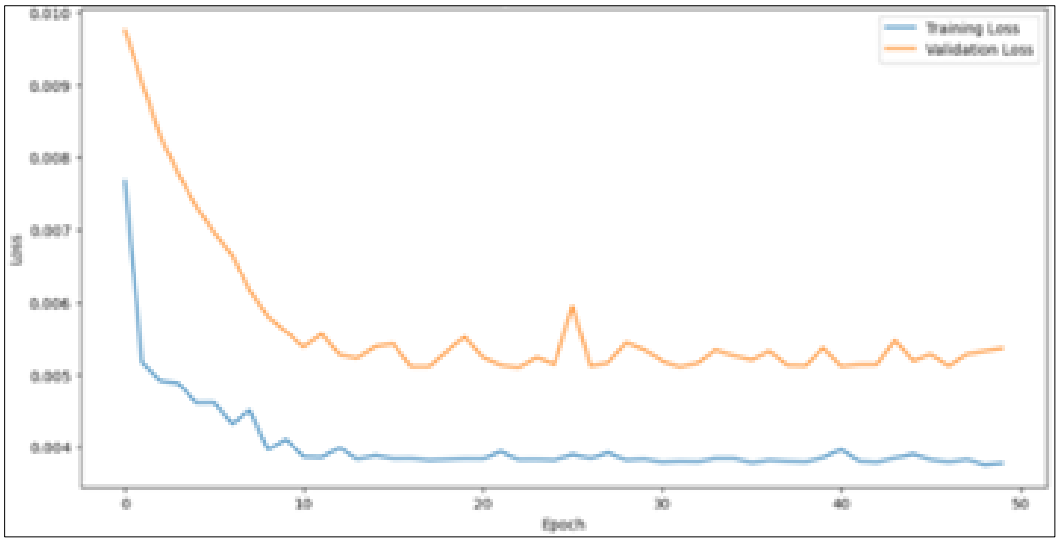


Figure 10 Training and Validation Loss

4. Conclusion

The comparison of AQI forecasting models shows that LSTM outperforms all other approaches with the lowest RMSE (29.64), MAE (20.55), and the highest R-squared (0.7751). This demonstrates its strong ability to capture non-linearity, long-term dependencies, and seasonal variations in air quality data. SARIMA emerges as the second-best model, effectively modelling both trend and seasonality, followed by ARIMA, which performs reasonably well but is limited in handling complex seasonal patterns. Exponential smoothing methods (SES, Holt's, Holt-Winters) show moderate performance, capturing basic trends but lacking adaptability to fluctuations. Prophet also shows limited accuracy, similar to Holt-Winters, and may benefit from incorporating additional variables. In contrast, SVM performs poorly with the highest error rates, indicating its inadequacy for time series forecasting in this context. Overall, deep learning models like LSTM are best suited for AQI prediction due to their capacity to handle sequential, non-linear data. In contrast, traditional models like SARIMA and ARIMA still hold value for capturing structured patterns. Future work could enhance model performance by integrating other environmental factors and developing hybrid approaches.

Future scope

This work can be expanded using various time series methods to forecast the air quality index or different pollutants for different places. Additionally, traditional and machine learning models can be combined to construct hybrid models. The efficiency of these hybrid models can be compared with that of conventional and machine learning models to inform a future course of action.

5. Compliance with ethical standards

Acknowledgments

The authors sincerely thank the Institute of Science, Nagpur, and its director for providing the necessary resources and support to conduct this research.

Conflict of Interest Statement

The authors declare that there is no conflict of interest.

References

[1] L. Liu *et al.*, "Spatial–Temporal Analysis of Air Pollution, Climate Change, and Total Mortality in 120 Cities of China, 2012–2013," *Front. Public Health*, vol. 4, Jul. 2016, doi: 10.3389/fpubh.2016.00143.

[2] S. Deepan and M. Saravanan, "Air quality index prediction using seasonal autoregressive integrated moving average transductive long short-term memory," *ETRI Journal*, vol. 46, no. 5, pp. 915–927, Oct. 2024, doi: 10.4218/etrij.2023-0283.

[3] S. Fan, D. Hao, Y. Feng, K. Xia, and W. Yang, "A Hybrid Model for Air Quality Prediction Based on Data Decomposition," *Information*, vol. 12, no. 5, p. 210, May 2021, doi: 10.3390/info12050210.

[4] A. Kumar and P. Goyal, "Forecasting of daily air quality index in Delhi," *Science of The Total Environment*, vol. 409, no. 24, pp. 5517–5523, Nov. 2011, doi: 10.1016/j.scitotenv.2011.08.069.

[5] V. Yadav, A. K. Yadav, V. Singh, and T. Singh, "Artificial neural network an innovative approach in air pollutant prediction for environmental applications: A review," *Results in Engineering*, vol. 22, p. 102305, Jun. 2024, doi: 10.1016/j.rineng.2024.102305.

[6] W. Abuouelezz, N. Ali, Z. Aung, A. Altunaiji, S. B. Shah, and D. Gliddon, "Exploring PM2.5 and PM10 ML forecasting models: a comparative study in the UAE," *Sci Rep*, vol. 15, no. 1, p. 9797, Mar. 2025, doi: 10.1038/s41598-025-94013-1.

[7] A. Hasnain *et al.*, "Time Series Analysis and Forecasting of Air Pollutants Based on Prophet Forecasting Model in Jiangsu Province, China," *Front. Environ. Sci.*, vol. 10, p. 945628, Jul. 2022, doi: 10.3389/fenvs.2022.945628.

[8] S.-C. Necula, I. Hauer, D. Fotache, and L. Hurbean, "Advanced Hybrid Models for Air Pollution Forecasting: Combining SARIMA and BiLSTM Architectures," *Electronics*, vol. 14, no. 3, p. 549, Jan. 2025, doi: 10.3390/electronics14030549.

[9] N. Sarkar, R. Gupta, P. K. Keserwani, and M. C. Govil, "Air Quality Index prediction using an effective hybrid deep learning model," *Environmental Pollution*, vol. 315, p. 120404, Dec. 2022, doi: 10.1016/j.envpol.2022.120404.

[10] G. Ravindiran, G. Hayder, K. Kanagarathinam, A. Alagumalai, and C. Sonne, "Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam," *Chemosphere*, vol. 338, p. 139518, Oct. 2023, doi: 10.1016/j.chemosphere.2023.139518.

[11] S. Al-Eidi, F. Amsaad, O. Darwish, Y. Tashtoush, A. Alqahtani, and N. Niveshitha, "Comparative Analysis Study for Air Quality Prediction in Smart Cities Using Regression Techniques," *IEEE Access*, vol. 11, pp. 115140–115149, 2023, doi: 10.1109/ACCESS.2023.3323447.

[12] A. Mohammadi, E. Pishgar, and J. Aguilera, "Spatial Prediction of High-Risk Areas for Asthma in Metropolitan Areas: A Machine Learning Approach Applied to Tehran, Iran," *IJGI*, vol. 14, no. 3, p. 105, Mar. 2025, doi: 10.3390/ijgi14030105.