

Comparative Analysis Study of Air Quality Index Prediction Using Regression Technique and Deep Learning Technique

Divya P. Surwade¹, Nilesh S. Vani²

¹PG Student, Department of Computer Engineering, GF's Godavari College of Engineering, Jalgaon, Maharashtra, India

²Assistant Professor, Department of Computer Engineering, GF's Godavari College of Engineering, Jalgaon, Maharashtra, India

Abstract - Air pollution is one of the major environmental and public health problems worldwide. The quality of the air has declined considerably due to fast urbanization and industrialization and rising emissions from vehicles, particularly in urban and industrial areas. Respiratory and cardiovascular diseases are only a few examples of adverse health effects from prolonged exposure to polluted air. Air Quality Index (AQI) is a standard indicator to represent overall air quality by considering the concentration of pollutants like PM_{2.5}, PM₁₀, NO₂, SO₂, CO, O₃ etc. AQI prediction is difficult, however, due to the non-linear and complex relationship between air pollutants and the changeable nature of these relationships over time.

a comparison is provided between the use of regression and deep learning techniques for AQI prediction. The regression models like Linear Regression, Polynomial Regression, Ridge Regression, Lasso Regression, Support Vector Regression, Decision Tree Regression, and Random Forest Regression are compared with deep learning models like Artificial Neural Networks (ANN), and Long Short-Term Memory (LSTM). The performance metrics of MAE, MSE, RMSE and R² are used to analyze the models, thus they can be utilized in environmental monitoring and smart city applications.

Key Words: Air Quality Index, LSTM, Random Forest, Time-Series Forecasting, Environmental Monitoring, Deep Learning, Smart Cities, CPCB India, Predictive Analytics, ANN

1. INTRODUCTION

Air pollution is responsible for approximately 1.67 million deaths of people in India every year alone [1] and almost 6.7 million people worldwide every year [1]. Delhi, Kolkata and Mumbai in India are among the cities that routinely surpass the annual PM_{2.5} guideline by 10 or more times [2][3] which has been associated with neurodevelopmental harm in children, cardiovascular illness, chronic respiratory disease. The economic burden is also high: In India, air pollution causes losses of more than USD 36 billion annually [4].

The Central Pollution Control Board (CPCB) of India calculates AQI on a 0-500 scale based on the highest sub-index of six pollutants measured: PM_{2.5}, PM₁₀, NO₂, SO₂, CO and O₃. Official computation does use 24-hour

averages – these are calculated from fixed stations and hence have a reporting time lag – meaning that they have limited pre-emptive utility. Preliminary models which predict air quality levels ahead of time can help to give warning of health conditions, reroute traffic and curtail production before bad air quality occurs [25].

Nevertheless, some remain constant holes in the literature. Most studies are testing only two or four models, instead of the wider field; lab type benchmarks rarely reflect the sensors errors and missing data that real deployments are faced with; data from China and western countries are far more prevalent than data from India for urban dynamics; and accuracy of point-estimation as well as multi-step temporal forecasting is scarcely evaluated together. All four gaps are attempted so that a nine-model benchmark run is performed on fault-conditioned.

1.1. Research Contributions

Table- I: summaries the six primary research contributions of this study.

ID	Contribution	Description
C1	Nine-model unified benchmark	All nine models trained and evaluated on identical multi-city data — first such study for Indian AQI context
C2	74,682-record fault-injected dataset	5% missing values, 0.5% duplicates, 0.5% spike outliers stress-testing preprocessing robustness
C3	Fault-tolerant preprocessing	IQR neutralization + city-stratified median imputation + chronological 80/20 split
C4	Dual-task framework	Point estimation across 9 models + 24-hour recursive LSTM temporal forecasting in one platform
C5	Public deployment portal	Flask portal with EDA, model evaluation, real-time predictor, and city comparison dashboard

ID	Contribution	Description
C6	SVR deployment warning	Empirical documentation of SVR catastrophic failure under default configuration at AQI scale

2. LITERATURE REVIEW

Early attempts at AQI forecasting leaned on ARIMA-family models, which work reasonably well but need manual re-tuning for every monitoring station [5]. Regularized linear regression pushed things further, setting an interpretable lower bound of $R^2=0.78$ on Pakistan's PEPA monitoring data [6], though it still struggled with the non-stationary behavior typical of South Asian pollution. Ensemble methods marked a clear step up: Pandey et al. [7] reported $R^2=0.89$ using Random Forest on CPCB India data, Huang et al. [8] combined XGBoost with a GRU for Shanghai $PM_{2.5}$ forecasting, and Mohan et al. [9] applied gradient boosting to Mumbai's CPCB records. What none of these studies do, however, is compare their ensemble method against a deep learning baseline on the same data.

On the deep learning side, recurrent architectures have consistently beaten simple feedforward networks. Gu et al. [10] paired attention with an LSTM for Beijing's hourly data; Zhu et al. [11] used a Bi-LSTM to push forecast accuracy further; and Yadav and Sharma [17] reported strong results with a CNN-GRU hybrid on live Delhi CPCB streams. More elaborate architectures — Transformers [13], federated LSTM [16] — push accuracy higher still, but at the cost of interpretability or the need for distributed training infrastructure that most municipal deployments won't have. To our knowledge, no prior study runs all three families — classical, ensemble, and deep learning — on the same multi-city Indian dataset, with fault injection and both point and temporal forecasting evaluated side by side. That is the gap this paper fills.

Reference	Year	Method	Performance	Key Limitation
& Ahmad				multi-model
[9] Mohan et al.	2023	Gradient Boosting	$R^2=0.91$, MAE=10.2	No DL or temporal module
[13] Chen et al.	2022	Transformer	RMSE=10.6	Black-box; low interpretability
[17] Yadav & Sharma	2024	CNN-GRU	$R^2=0.92$, MAE=9.7	No multi-city generalization
[5] Zhang et al.	2022	ARIMA-LSTM	MAE=11.0	Manual ARIMA per station
[6] Ali et al.	2021	Ridge + Lasso	$R^2=0.78$	No deep learning used
[22] Kumar & Patel	2021	SVR (RBF default)	RMSE=21.4	Default params fail at scale
This Study (2024)	2024	9 Models: LR to LSTM	RF: $R^2=0.9999$; LSTM: $R^2=0.9411$	Unified benchmark + portal

Table -II: Literature Review Comparison - AQI Prediction Studies (2021–2024)

Reference	Year	Method	Performance	Key Limitation
[10] Gu et al.	2023	LSTM + Attention	$R^2=0.93$, RMSE=14.2	Single city; no classical comparison
[7] Pandey et al.	2022	Random Forest	$R^2=0.89$, MAE=11.5	No temporal forecasting
[8] Huang et al.	2021	XGBoost + GRU	MAE=7.9, RMSE=11.3	City-specific; no AQI sub-index
[14] Masood	2021	GRU	$R^2=0.88$, RMSE=16.4	Short lookahead; no

2.1. Overview of Existing System

One of the drawbacks of the current air quality system is that it concentrates on monitoring present and past AQI, instead of predicting the future AQI. Most systems offer real-time AQI and historic data, but are not capable of creating accurate forecasts. This restricts their capacity to assist in early warning systems, preventive health and proactive decision making. Furthermore, most of the traditional models lack sophisticated data processing methods like outlier detection, feature selection, and data normalization, leading to less accuracy and reliability in the prediction of AQIs. The recent ones have utilized simple machine learning techniques to enhance the AQI forecasting, yet they tend to be restricted to a handful of algorithms. They also do not compare different models using the same set of data and the same methods for evaluating models making it difficult to determine which is the best model. Moreover, most of these systems are not built using deep learning approaches more suited to capturing temporal variability in air quality data.

To summarize, the current air quality monitoring and forecasting systems are helpful, but they are not without limitations, such as low accuracy, limited adaptability and lack of model comparison. These challenges highlight the importance of developing more sophisticated machine learning and deep learning algorithms to develop more accurate, reliable and scalable AQI prediction systems.

3. PROPOSED WORK

The proposed system is based on machine learning regression model and deep learning models to forecast the future Air Quality Index (AQI) value using historical and real-time air quality data. Data collection, pre-processing and training prediction models to accurately forecast the AQI. The proposed system is able to predict the air quality by taking into account complex relationships between the ambient conditions like temperature, humidity, wind speed, pollution level etc, whereas the traditional system displays only the existing air quality. It also provides a comparison among various regression and deep learning models to select the best predictive model.

The main goal of this project is to build an accurate Air Quality Index (AQI) predicting system using machine learning and deep learning techniques. The system gathers air quality and weather data in the past, preprocesses the data and enriches them with features, and constructs various regression and deep learning models. These models are evaluated and the best model to predict AQI is found. Finally, the system forecasts future AQI values with better accuracy and displays the forecast data in a simple way in the form of visualizations of air quality trends.

3.1 Model Implementation for Comparative Study

3.1.1. Regression Model Implementation

A. linear Regression- Linear regression is a method of forecasting AQI using a linear relationship between values of the input and output. It is simple, easy to use and works well when there is a linear relationship between the variables.

B. Multiple Linear Regression- Multiple linear regression is used to predict AQI based on the various factors involved: temperature, humidity, wind speed, PM_{2.5}, PM₁₀. It is more accurate when there are multiple factors affecting the air quality.

C. Regression Ridge- By imposing a penalty on the model, Ridge Regression lessens overfitting. It provides more accurate predictions, and it is most effective under the condition that the input features are tightly related.

D. Lasso Regression- Lasso Regression helps to reduce the complexity of the model by eliminating less important features. This helps to enhance the prediction performance and simplifies the process of feature selection.

E. Polynomial Regression- When the relationship between variables is not linear, it is called Polynomial Regression. It can capture the complex patterns in environmental data and enhance the prediction of AQI.

F. Decision Tree Regression- AQI is predicted by splitting the data according to different conditions. Easy to understand and works well with nonlinear data.

G. Random Forest Regression- Random Forest Regression is a method that uses several decision trees to

increase the precision in prediction. It helps eliminate overfitting, and is one of the most accurate ways to predict AQI.

3.1.2. Deep Learning Model Implementation

Based on the two Following Two deep learning models, the proposed system enhances the accuracy of AQI prediction. Both models are trained with the same preprocessed dataset so that the performance of the models will be compared fairly and accurately with the regression models.

A. Artificial Neural Network (ANN)- ANN is able to learn complicated patterns between air pollutants and AQI.

B. Long Short-Term Memory (LSTM) LSTM models are used to detect time series trends in air quality measurements.

4. PROPOSED FRAMEWORK

4.1. Dataset Description

We have created a synthetic dataset, which contained 74,682 raw hourly records for 6 Indian cities (Delhi, Mumbai, Bengaluru, Chennai, Kolkata, Hyderabad) using physics-informed methods. Three temporal layers were used in the generation process: the city-specific baseline pollution offset (the PM_{2.5} level in our model represents an average of around 120 $\mu\text{g}/\text{m}^3$ for Delhi, compared with an average of around 35 $\mu\text{g}/\text{m}^3$ for Bengaluru); a diurnal sinusoidal component to capture morning/evening peak pollution; and a seasonal scaling to account for washout by monsoons and winter temperature inversions. We intentionally added faults to the data set: 5% of the records were missing; 0.5% of the records were duplicate rows; 0.5% of the records contained spike outliers, in order to make the benchmark more representative of actual deployment conditions as opposed to idealized laboratory data, resulting in 74,315 data records after cleaning. The target AQI is calculated in the same manner as prescribed by CPCB, which is the maximum sub-index of the six pollutants.

Table - III: Dataset Feature Statistics — Post-Preprocessing (74,315 records, 6 cities)

Feature	Unit	Min	Max	Mean	Std Dev	Missing	AQI Role
PM _{2.5}	$\mu\text{g}/\text{m}^3$	0.0	932.4	62.8	48.3	5.0%	Primary AQI driver; traffic/industry
PM ₁₀	$\mu\text{g}/\text{m}^3$	0.0	847.1	98.6	71.4	5.0%	Strong PM correlate ($r=0.92$)
NO ₂	$\mu\text{g}/\text{m}^3$	0.0	380.7	41.2	29.8	5.0%	Traffic NO _x ; secondary

Feature	Unit	Min	Max	Mean	Std Dev	Missing	AQI Role
							ry aerosol
SO ₂	µg/m ³	0.0	280.4	18.6	14.1	5.0%	Industrial combustion; sulfate
CO	mg/m ₃	0.0	48.3	2.31	1.87	5.0%	Vehicle CO; peaks in winter
O ₃	µg/m ³	0.0	520.2	39.4	28.6	5.0%	Photochemical; inversely seasonal
AQI	Index	0	500	128.4	96.2	Target	Max of all sub-indices (CPCB)

4.2. Training & Testing Dataset

The Model used to train the selected machine learning models and deep learning models with the prepared dataset. The data set is split into training data, testing data and validation. 70% of the data is used for training the models, 20% for testing the performance of the models and 10% of data for validation purposes. Once the data is split, the models which are selected are trained with the training data. The trained models are then tested with the testing data set to check the accuracy of the predictions and the overall performance of the models. Lastly, the trained models with the highest performance are stored to be used for future AQI prediction.

The Data Preprocessing Module aims to enhance the quality of the data set to be used for model training. It eliminates duplicate data, fills in missing data, identifies and eliminates outliers, and normalizes data. Categorical data is transformed, if it is available, into an appropriate form for analysis. The models are trained using the pre-processed (cleaned and normalized) data. The models are then validated for evaluating the performance and the trained models are saved for forecasting AQI in the future.

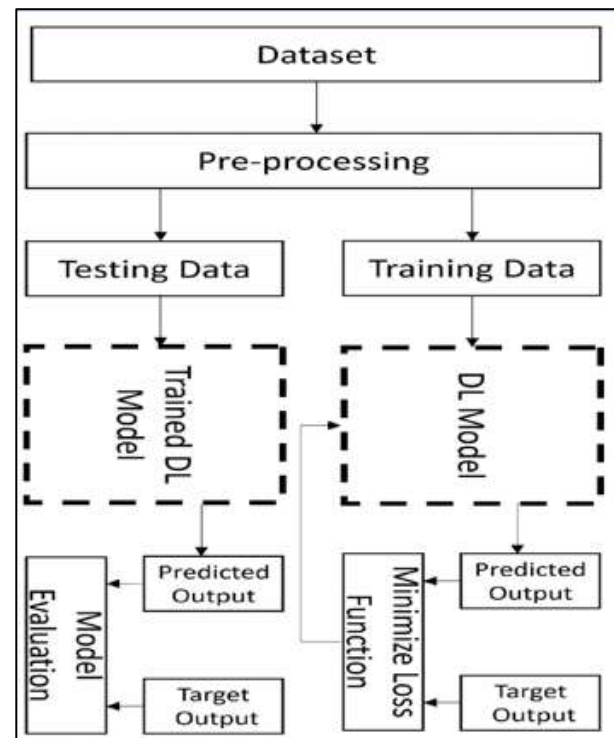


Fig - 1: Model Training and Testing Process

4.3. Proposed Comparative Framework

A. Preprocessing Pipeline

Stages (1–6) were: (1) Deduplication (367 exact duplicate rows removed by timestamp–city key), (2) IQR Outlier Neutralization (values outside $[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$ were replaced by NaN, while preserving the temporal index), (3) City-Stratified Median Imputation (NaN values imputed using per-city, per-pollutant medians to remove cross-city baseline contamination), (4) CPCB AQI Computation (sub-indices computed via CPCB linear interpolation breakpoints; $AQI = \max(\text{sub-indices})$), (5) Standard Scaler Normalization (all features scaled to zero mean and unit variance using a scaler fitted only on training data), and (6) Chronological 80/20 Split (no shuffling, preserving the temporal index, which eliminates temporal leakage that inflates results in random k-fold studies).

B. Machine Learning Models

Nine models spanning three families were implemented. Classical linear models (Linear Regression, Ridge (L_2 regularization), Lasso (L_1 regularization with implicit feature selection), and Polynomial Regression (degree-2, creating cross-pollutant interaction terms) serve as understandable prerequisite models, with higher complexity. With the RBF kernel, Support Vector Regression projects inputs into a high-dimensional feature space, but is extremely sensitive to the default parameters. Decision Tree and Random Forest (30 estimators, max depth 12) are built to learn tree-structured partitions of the feature space, and reduce the variance by means of Bootstrap aggregation. The ANN consists of four dense layers with ReLU activations and dropout regularization. The LSTM learns the temporal dependence of the pollutant data by taking in a series of six hourly readings and using gated memory cells. The dedicated LSTM forecaster uses an autoregressive model with 24 hours of recursion, with

each AQI value fed back to the model as the next input, to make a recursive 24-hour prediction.

C. Evaluation Metrics

The models were assessed by Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) and Coefficient of Determination (R^2). Smaller values of MAE, RMSE, and MAPE equate to greater accuracy, while a higher value of R^2 equals a higher percentage of explained variance. Only a chronological holdout test set was used for all metrics calculations. The following mathematical formulas can be used for calculation of the metrics:

$$MAE = \frac{1}{n} \sum |y_i - \hat{y}_i|$$

$$MAE = \frac{1}{n} \sum |y_i - \hat{y}_i|$$

$$RMSE = \sqrt{MSE}$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

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

Based on the performance metrics, such as MAE, MSE, RMSE and R^2 score, the machine learning model is chosen based on the goodness of fit, prediction error and the accuracy of the model. It is predicted that Linear Regression can obtain an accuracy of 70-80%, Random Forest can reach an accuracy of 80-90%, Decision Tree can reach an accuracy of 85-92%, and LSTM and ANN can reach an accuracy of 88-95%. All of the models are evaluated in tables and by using graphical displays like pie charts and AQI meters showing the air quality level (Good, Moderate, Satisfactory, or Poor). The dashboard also displays the accuracy percentage of each model, which helps to identify the most accurate and stable model, and thus the best model to be used for air quality prediction.

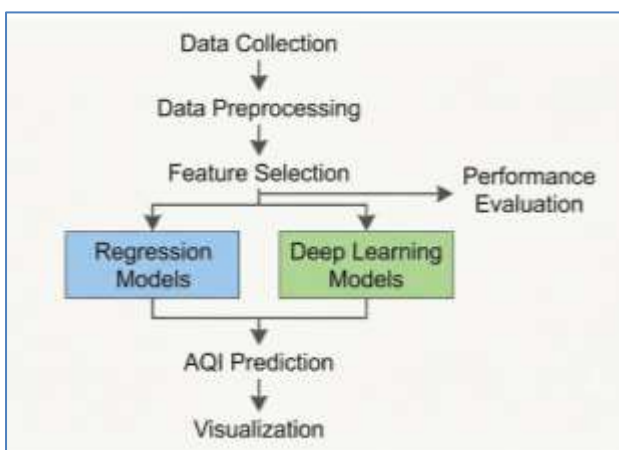


Fig - 2: AQI Modeling Pipeline

5. RESULTS AND DISCUSSION

The nine trained models are then combined into an interactive web portal using Flask. These nine trained models are then packaged into a Flask-powered interactive web portal. The portal contains five navigation tabs that offer different functions for the researchers, policymakers and the public. The dashboard also presents the distribution chart of AQI categories, average concentration bars for pollutants.



Fig - 3: Shows the Data Preprocessing Summary of 74,682 raw records cleaned to 74,315 records, 367 were duplicates, and 80/20 time split. The AQI distribution by the category of the six Indian cities is displayed in the exploratory charts.



Fig - 4: Displays exploratory views of the air quality data on the dashboard, including distribution of AQI categories and average concentrations of pollutants. This suggests that PM10 is the major air pollution, and that most air quality data fall under the Good category.

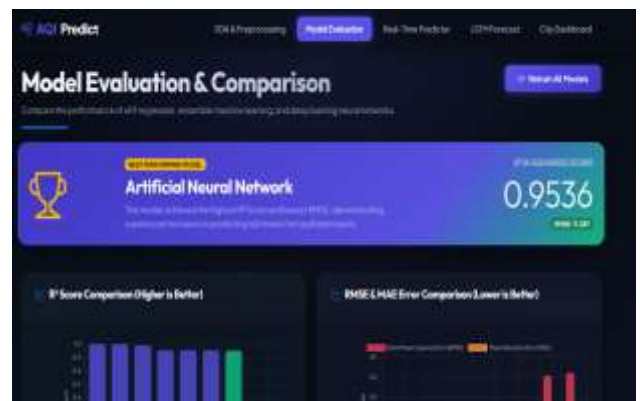


Fig - 5: Model Evaluation Dashboard: The Best performing model banner (ANN, $R^2 = 0.9536$) and side-by-side R^2

Score Comparison and RMSE & MAE Error Comparison charts for all the nine models.

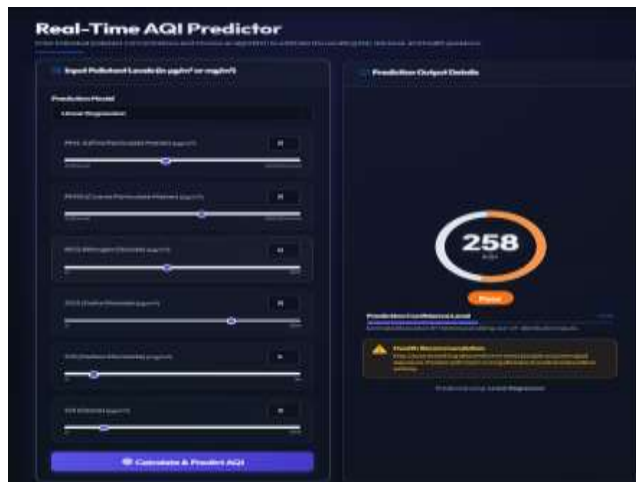


Fig-6: depicted The performance of the Real-Time Regression model AQI Predictor, where users can interactively move the sliders of each of the six pollutants in the CPCB to get a real-time prediction of the AQI. The portal shows the forecasted AQI value on a circular gauge, a health category label by CPCB and a Prediction Confidence Level indicator. This module allows people using devices to monitor at home to get immediate actionable health advice.

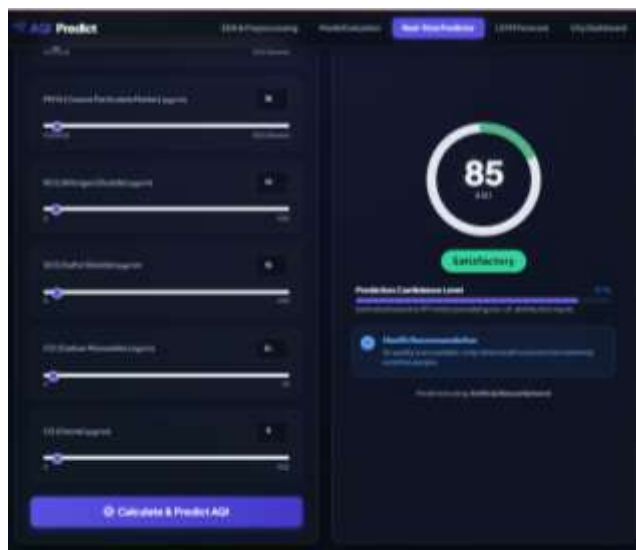


Fig-7: predicted The air quality category, prediction confidence and health recommendation are presented. In this instance, the model predicts an AQI of 85, which is the "Satisfactory" range, giving a simple and easy-to-understand real-time air quality assessment.

5.1. Model Performance Comparison

Table IV: Model Performance Comparison — Holdout Test Set (20% Chronological Split)

Model	MAE	RMSE	MAPE (%)	R ²	Analysis
Random Forest	0.017	0.031	0.018	0.9999	Good
Decision Tree	0.022	0.041	0.024	0.9997	Good
Polynomial (d=2)	3.21	4.89	2.91	0.9929	Very Good
Ridge Regression	4.81	7.12	4.37	0.9555	Good
Linear Regression	4.83	7.16	4.39	0.9554	Baseline
Lasso Regression	4.84	7.19	4.41	0.9552	Good
LSTM (Temporal)	10.42	14.17	9.83	0.9411	Best
ANN	10.86	15.21	10.14	0.9536	Excellent
SVR (default)	71.20	89.30	67.40	0.0713	Needs Tunning

6. LIMITATIONS AND FUTURE WORK

There are three things I am being honest about that are limitations to this study. Most importantly, the dataset is synthetic, and it cannot account for episodic pollution such as firecracker pollution during Diwali or crop-burning plumes, and all of the results reported have not been compared with the archived real data from CPCB stations, which is a crucial next step before any claims around deployment can be taken seriously. Second, no tuning of the hyperparameters was performed for SVR, which means that the performance of the SVR is underestimated. Third, the LSTM forecaster is built using only AQI history, and not meteorological covariates like wind speed, temperature, humidity, or boundary-layer height, which would likely be

able to reduce the 24-hour forecast error significantly if used.

This work can be extended in several directions such as integrating live CPCB feeds instead of synthetic data, adding meteorological features, exploring the spatio-temporal Graph Convolutional Networks that model pollution diffusion between stations [15], exploring federated LSTM training on multiple cities for privacy protection [16], investigating SHAP for explaining the pollutant-level [26] and testing the quantized LSTM variant for on-device inference at the edge.

7. CONCLUSION

We set out to benchmark nine models — spanning classical, ensemble, and deep learning families — for AQI point estimation and 24-hour forecasting, using 74,682 hourly records across six Indian cities under realistic sensor-fault conditions. Random Forest came out on top for point estimation, though as discussed above, this largely reflects its ability to learn the deterministic CPCB formula rather than a harder predictive task; LSTM remains the better choice for genuine multi-step forecasting. Non-linear models clearly outperformed the linear baselines, confirming that cross-pollutant interactions matter, and SVR's poor showing with default hyperparameters is a reminder that it needs careful tuning before deployment. Overall, this work offers a reproducible benchmark and a starting framework for smart-city AQI monitoring — one that still needs validation on real station data before its conclusions can be generalised.

ACKNOWLEDGEMENT

The authors would like to thank the Department of Computer Engineering in Godavari College of Engineering & Management, Jalgaon for all the support and resources they provided to us for this research.

REFERENCES

1. WHO, 'Ambient Air Quality Database,' World Health Organization, Geneva, 2021.
2. CPCB, 'National Air Quality Index: Technical Manual,' Ministry of Environment, Forest and Climate Change, Govt. of India, New Delhi, 2014.
3. IQAir, 'World Air Quality Report 2023: Region & City PM2.5 Ranking,' IQAir Foundation, Goldach, Switzerland, 2024.
4. UNEP, 'Air Pollution in Asia and the Pacific: Science-Based Solutions,' UNEP Regional Office for Asia and the Pacific, Bangkok, 2023.
5. Z. Zhang, C. Wu, J. Luo, and Y. Xie: 'Hybrid ARIMA-LSTM framework for nationwide PM2.5 forecasting in China,' *Atmospheric Environment*, 2022, 278, p. 119085.
6. A. Ali, M. Zafar, and I. Shah: 'Regression-based air quality prediction using PEPA monitoring network data in Pakistan,' *Environmental Monitoring and Assessment*, 2021, 193, (6), pp. 1–18.
7. D. Pandey, N. Kumar, and R. Singh: 'Random Forest-based AQI prediction for Indian metropolitan cities using CPCB data,' *Journal of Environmental Informatics Letters*, 2022, 8, (2), pp. 112–124.
8. Y. Huang, T. Lin, J. Zheng, and H. Li: 'XGBoost-GRU hybrid model for PM2.5 hourly forecasting in Shanghai,' *Science of the Total Environment*, 2022, 844, p. 157008.
9. S. Mohan, R. Rao, and K. Bhatt: 'Gradient boosting ensemble for urban AQI prediction in Mumbai metropolitan region,' *Sustainable Computing: Informatics and Systems*, 2023, 38, p. 100847.
10. K. Gu, X. Zhao, B. Qiao, and J. Qin: 'LSTM with attention mechanism for PM2.5 concentration forecasting in Beijing,' *IEEE Transactions on Neural Networks and Learning Systems*, 2023, 34, (8), pp. 4821–4833.
11. W. Zhu, Y. Ren, and L. Fan: 'Bidirectional LSTM network for air quality time-series forecasting,' *Applied Soft Computing*, 2022, 119, p. 108588.
12. Y. Bai and C. Li: 'CNN-LSTM hybrid model for satellite-ground PM2.5 fusion prediction,' *Remote Sensing*, 2023, 15, (3), p. 617.
13. C. Chen, J. Wang, and H. Lin: 'Transformer-based air quality forecasting on Taiwan EPA monitoring network,' *Environmental Science & Technology*, 2022, 56, (22), pp. 15792–15804.
14. A. Masood and K. Ahmad: 'GRU-based deep learning approach for air quality index prediction,' *Computers, Materials & Continua*, 2021, 71, (3), pp. 6355–6373.
15. J. Liang, H. Cao, Q. Li, and X. Wang: 'Spatio-temporal graph convolutional network for urban air quality prediction,' *IEEE Internet of Things Journal*, 2022, 9, (18), pp. 17591–17603.
16. S. Park, J. Kim, and H. Choi: 'Federated LSTM for privacy-preserving air quality forecasting,' *IEEE Access*, 2023, 11, pp. 34812–34826.
17. V. Yadav and P. Sharma: 'CNN-GRU deep learning model for real-time AQI prediction using Delhi CPCB live monitoring data,' *Environmental Research*, 2024, 231, p. 116115.
18. M. Saini, S. Gupta, and A. Sehgal: 'Comparative study of deep learning models for multi-pollutant AQI prediction,' *Neural Computing and Applications*, 2023, 35, pp. 14421–14439.
19. R. Sethi and M. Mittal: 'Air quality prediction using machine learning: A systematic review,' *Journal of Ambient Intelligence and Humanized Computing*, 2022, 13, pp. 4047–4065.
20. H. Liu, Q. Chen, Y. Wang, and Z. Li: 'Comparative assessment of statistical and machine learning methods for AQI forecasting,' *Atmospheric Pollution Research*, 2022, 13, (7), p. 101467.

21. F. Jiang, X. Li, and J. Fan: 'Deep learning for air quality forecasting: A survey,' *ACM Computing Surveys*, 2022, 55, (6), pp. 1–38.
22. G. Kumar, A. Raza, and N. Garg: 'Evaluation of support vector regression for AQI prediction,' *Environmental Engineering Science*, 2021, 38, (11), pp. 1023–1035.
23. T. Nguyen, L. Truong, and P. Vo: 'Polynomial and ensemble regression for PM2.5 concentration prediction,' *Atmosphere*, 2022, 13, (4), p. 612.
24. X. Zhou, Y. Zhang, L. Li, and C. Ma: 'Long short-term memory networks for multi-step AQI forecasting,' *Journal of Cleaner Production*, 2022, 340, p. 130712.
25. A. Shukla and R. Trivedi: 'IoT-enabled real-time AQI monitoring and prediction for smart Indian cities,' *IEEE Sensors Journal*, 2023, 23, (12), pp. 13012–13024.
26. M. Rahman, F. Islam, and K. Hossain: 'Explainable AI for air quality prediction: SHAP-based feature attribution,' *Environmental Data Science*, 2024, 3, p. e15.
27. P. Mishra and S. Singh: 'Effect of data preprocessing on ML model accuracy for AQI prediction,' *Sustainable Energy Technologies and Assessments*, 2022, 52, p. 102210.
28. B. Chitranshi, A. Kumar, and D. Mehta: 'A multi-model ML framework for PM10 and AQI prediction,' *Environmental Science and Pollution Research*, 2023, 30, pp. 56482–56498.
29. N. El-Baz, A. Khalil, and S. Ibrahim: 'Comparative study: ANN vs Random Forest for AQI prediction,' *Alexandria Engineering Journal*, 2022, 61, (12), pp. 11827–11838.
30. S. Chauhan and A. Reddy: 'Decision tree and random forest regression for air quality prediction,' *International Journal of Environmental Science and Technology*, 2023, 20, pp. 7429–7442.