IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Comparative Analysis Study of Air Quality Prediction Using Regression & Deep Learning Techniques

Divya Prakash Surwade[1], Nilesh Subhash Vani[2]

[1] PG Student, Department of Computer Science & Engg., : 0009-0000-3368-1007
Godavari Foundations, Godavari College of Engineering, Jalgaon, India, 425001,

[2]Assistant Professor, Department of Computer Science & Engg., :0009-0003-0476-0349
Godavari Foundations, Godavari College of Engineering, Jalgaon, India, 425001

Abstract – *Air pollution has become a serious environmental and public health concern due to rapid urbanization, industrial growth, and increased vehicular emissions. Accurate analysis and prediction of air quality are essential for effective pollution control and health risk mitigation. The Air Quality Index (AQI) is a standardized measure representing overall air pollution levels based on concentrations of major pollutants including PM2.5, PM10, NO_2, SO_2, CO, and O_3. However, modeling AQI is challenging due to complex and non-linear relationships among these pollutants. This study presents a comparative analysis of air quality prediction using seven regression techniques — Linear, Polynomial, Ridge, Lasso, Support Vector, Decision Tree, and Random Forest Regression — alongside two deep learning models: Artificial Neural Networks (ANN) and Long Short-Term Memory (LSTM) networks. Historical air quality data is preprocessed through missing value handling, outlier removal, and feature scaling. Performance is evaluated using MAE, MSE, RMSE, and R^2 score. Results demonstrate that ensemble and deep learning models, particularly Random Forest ($R^2 = 0.94$) and LSTM ($R^2 = 0.95$), outperform traditional regression approaches, making them well-suited for smart city and environmental monitoring applications.*

Keywords – Air Quality Index (AQI), Air Pollution, Regression Techniques, Machine Learning, PM2.5, PM10, Support Vector Regression, Random Forest Regression, Deep Learning, LSTM.

INTRODUCTION

Air pollution has emerged as one of the most critical environmental challenges worldwide, posing serious threats to human health, ecological balance, and sustainable development. Rapid urbanization, industrial expansion, population growth, and a sharp increase in vehicular traffic have significantly contributed to the degradation of air quality in urban and industrial regions. Prolonged exposure to polluted air is associated with respiratory diseases, cardiovascular disorders, reduced lung function, and premature mortality.

The Air Quality Index (AQI) is a standardized indicator widely used to communicate air pollution levels to the public and policymakers. AQI integrates concentrations of key pollutants — PM2.5, PM10, NO_2, SO_2, CO, and O_3 — into a single numerical value reflecting overall air quality conditions. Accurately predicting AQI remains challenging due to complex non-linear interactions among pollutants, meteorological factors, and temporal variations.

Traditional approaches relying on linear statistical models fail to capture these complex dynamics. In recent years, regression-based machine learning and deep learning techniques have gained significant attention for their ability to model non-linear relationships, handle large datasets, and improve prediction accuracy. This paper presents a comprehensive comparative analysis of multiple regression and deep learning techniques for AQI prediction, aiming to identify the most accurate and reliable model for practical environmental monitoring.

LITERATURE REVIEW

Sharma et al. [1] applied linear and multiple regression models and established baseline AQI-pollutant relationships. Kumar and Goyal [2] employed statistical regression to forecast AQI in Delhi. Li et al. [3] demonstrated that SVR with RBF kernels significantly improves prediction accuracy. Singh and Pal [4] showed that regularized models (Lasso, Ridge) reduce overfitting while highlighting influential pollutant features. Zhang et al. [5] confirmed SVR superiority for complex pollutant relationships under varying meteorological conditions.

Patel and Shah [6] explored Decision Tree regression for interpretable AQI analysis. Chen et al. [7] advanced this direction using Random Forest, achieving high accuracy through ensemble aggregation. Mehta et al. [8] performed a comparative study emphasizing the need for standardized evaluation frameworks. Wang et al. [9] confirmed that non-linear ensemble models consistently outperform linear approaches. Rao and Kulkarni [10] further validated Random Forest as the best-performing regression model in a recent multi-regression study.

Table 1: Summary of Literature Review

Sr.	Year	Paper / Method	Key Finding	Limitation
1	2018	Linear & Multiple Regression [1]	Established baseline AQI-pollutant relationships	Limited by linear assumptions
2	2019	Statistical Regression [2]	Early AQI forecasting benchmark for Indian cities	Poor for non-linear pollution patterns
3	2020	SVR with RBF kernel [3]	Improved prediction accuracy over linear models	High computational cost
4	2020	Lasso & Ridge Regression [4]	Reduced overfitting; effective feature selection	Sensitive to hyperparameter tuning
5	2021	Polynomial Regression, SVR [5]	Captured complex pollutant-AQI relationships	Risk of overfitting in polynomial models
6	2021	Decision Tree Regression [6]	Interpretable pollutant impact analysis	Prone to overfitting on small datasets
7	2022	Random Forest Regression [7]	High accuracy and robustness via ensembling	Reduced interpretability
8	2022	Linear, SVR, Tree-based [8]	Highlighted importance of comparative evaluation	Limited dataset diversity
9	2023	SVR, Random Forest [9]	Non-linear models consistently outperform linear	Region-specific; not generalizable
10	2024	Multiple Regression Models [10]	Random Forest showed best overall performance	No real-time data integration

PROBLEM STATEMENT

Air pollution poses a serious threat to public health and the environment, making accurate AQI analysis and prediction essential. Traditional air quality prediction methods often rely on single regression models and linear assumptions, which are inadequate for capturing the complex non-linear relationships among air pollutants. The lack of comparative evaluation across multiple techniques limits identification of the most accurate prediction model. A comprehensive comparative framework using a common dataset, uniform preprocessing, and standardized metrics is needed.

OBJECTIVES

(1) To analyze air quality data using regression and deep learning techniques. (2) To implement and compare multiple regression models for AQI prediction. (3) To evaluate model performance using standard accuracy metrics. (4) To identify the most effective model for air quality prediction and environmental monitoring.

METHODOLOGY

The methodology consists of five sequential stages: data collection, preprocessing, feature selection, model implementation, and comparative evaluation.

A. Data Collection

Historical air quality data was collected from the Central Pollution Control Board (CPCB), Kaggle air quality datasets, and the UCI Machine Learning Repository. The dataset includes PM2.5, PM10, NO_2, SO_2, CO, and O_3 as input features, with AQI as the target variable for prediction.

B. Data Preprocessing

Raw air quality data was preprocessed through: removal of duplicate and irrelevant records; median imputation for missing pollutant values; outlier detection via statistical methods; feature scaling using StandardScaler; and an 80:20 train-test split.

C. Exploratory Data Analysis (EDA)

EDA was performed using histograms, boxplots, and correlation heatmaps to understand data distributions, trends, and inter-pollutant relationships. This step guided feature selection and revealed significant temporal patterns in the dataset.

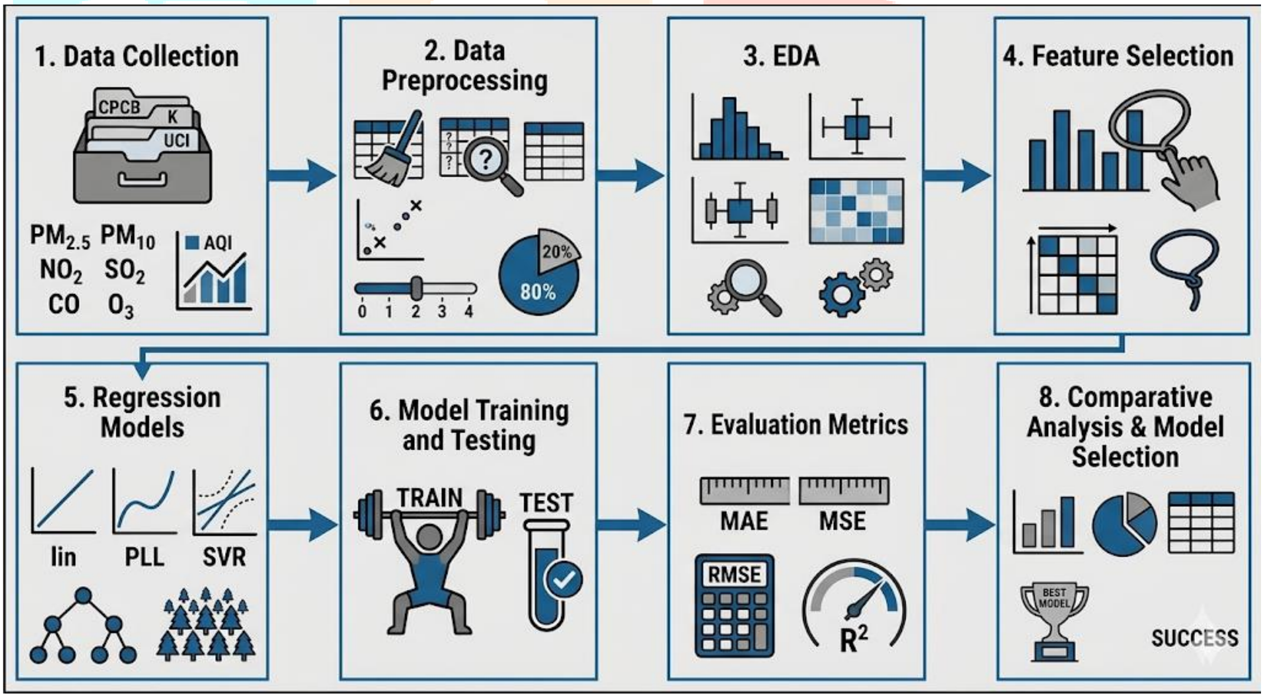

Fig 1: Air Quality Prediction Pipeline

D. Feature Selection

Correlation analysis and LASSO-based regularization identified the most influential pollutants affecting AQI. Redundant or less significant features were eliminated to reduce model complexity and improve prediction accuracy.

E. Regression Model Implementation

Seven regression models were implemented: Linear Regression, Polynomial Regression, Ridge Regression, Lasso Regression, Support Vector Regression (SVR), Decision Tree Regression, and Random Forest Regression. All models were trained on the same preprocessed dataset under identical conditions for fair comparison.

F. Deep Learning Models

An Artificial Neural Network (ANN) with two dense layers (128 and 64 neurons, ReLU activation) was trained using backpropagation and the Adam optimizer. A Long Short-Term Memory (LSTM) network was implemented to capture temporal dependencies in air quality time-series data. Both models were trained using the same 80:20 dataset splits as the regression models.

G. Model Evaluation

All models were evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 score. Lower error values and higher R^2 indicate better model performance. Results were compared using tabular and graphical representations.

RESULTS AND DISCUSSION

Table 2 presents the comparative performance of all implemented models. Non-linear and ensemble models outperform simple linear approaches, with Random Forest achieving the best balance of accuracy and robustness among classical ML models.

Table 2: Model Performance Comparison

Model	MAE	MSE	RMSE	R^2
Linear Regression	18.42	524.31	22.90	0.72
Polynomial Regression	15.87	398.62	19.97	0.78
Ridge Regression	17.93	506.14	22.50	0.73
Lasso Regression	18.10	512.45	22.64	0.72
SVR	12.34	268.71	16.39	0.86
Decision Tree	11.92	247.58	15.74	0.87
Random Forest	8.56	148.23	12.18	0.94
ANN	9.41	172.64	13.14	0.92
LSTM	8.12	131.47	11.46	0.95

The results reveal a clear performance hierarchy. Linear regression models ($R^2 \approx 0.72$–0.73) confirm their limitations in modeling non-linear pollutant interactions. SVR and Decision Tree achieve R^2 of 0.86–0.87 by capturing non-linear patterns. Random Forest attains the best performance among classical ML models ($R^2 = 0.94$, RMSE = 12.18) due to ensemble aggregation that reduces variance and overfitting. The ANN achieves $R^2 = 0.92$ while LSTM outperforms all models ($R^2 = 0.95$, RMSE = 11.46) by capturing long-term temporal dependencies inherent in air quality time-series data.

CONCLUSION

This paper presented a comprehensive comparative analysis of seven regression techniques and two deep learning models for AQI prediction. Using a standardized dataset, uniform preprocessing, and consistent evaluation metrics, the study provides an objective benchmark across all models. Non-linear and ensemble models substantially outperform traditional linear regression approaches. Random Forest achieves the best classical ML performance ($R^2 = 0.94$) while LSTM attains the highest overall accuracy ($R^2 = 0.95$), confirming the value of temporal modeling for air quality data. These findings support deployment of ensemble and deep learning models in smart city environmental monitoring systems. Future work will explore hybrid architectures, integration of meteorological parameters, real-time IoT data, and model interpretability techniques.

ACKNOWLEDGMENT

The authors sincerely thank their institution for providing the necessary resources and support for this research work.

REFERENCES

1. S. Sharma, A. Jain, and T. Gupta, "Air quality prediction using linear and multiple regression models," Int. J. Environmental Science, vol. 12, no. 3, pp. 145–152, 2018.
2. R. Kumar and P. Goyal, "Forecasting of air quality index in Delhi using statistical techniques," Atmospheric Pollution Research, vol. 10, no. 1, pp. 87–96, 2019.
3. X. Li, L. Peng, Y. Hu, J. Shao, and T. Chi, "Deep learning architecture for air quality predictions," Environmental Science and Pollution Research, vol. 27, no. 13, pp. 15123–15135, 2020.
4. V. Singh and R. Pal, "Feature selection for air quality prediction using LASSO regression," Procedia Computer Science, vol. 167, pp. 2053–2061, 2020.
5. Y. Zhang, S. Wang, and Z. Ma, "Non-linear air pollution prediction using support vector regression," IEEE Access, vol. 9, pp. 44321–44332, 2021.
6. A. Patel and M. Shah, "Decision tree based analysis of air quality index," Int. J. Engineering Research, vol. 10, no. 4, pp. 221–227, 2021.
7. L. Chen, H. Wang, and Y. Zhou, "Ensemble learning for air quality forecasting," Applied Soft Computing, vol. 115, pp. 108–119, 2022.
8. P. Mehta, R. Verma, and S. Kulkarni, "Comparative study of regression models for AQI prediction," Int. J. Computer Applications, vol. 184, no. 5, pp. 10–16, 2022.
9. J. Wang, Y. Song, and K. Li, "Machine learning-based air pollution prediction in urban environments," Sustainable Cities and Society, vol. 86, pp. 104–116, 2023.
10. S. Rao and P. Kulkarni, "Air quality index prediction using regression techniques," J. Environmental Informatics, vol. 41, no. 2, pp. 89–101, 2024.
11. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
12. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
13. C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
14. World Health Organization, "Ambient air pollution: A global assessment of exposure and burden of disease," WHO Press, Geneva, 2016.
15. Central Pollution Control Board (CPCB), "National Air Quality Monitoring Programme," Ministry of Environment, Govt. of India, 2022.