Machine Learning Approaches for Air Quality Prediction: A Review

Pawan Yadav[1], Dr. Arpit Solanki [2]
Dr. A.P.J. Abdul Kalam University, Indore
Corresponding Author Email: pawanyaduwanshi88@gmail.com

Abstract

Air pollution has become a major environmental concern as a result of growing urbanisation, industrial expansion, and increased traffic activity. Accurate air quality forecast is critical for pollution control, public health protection, and environmental sustainability. Traditional statistical tools sometimes struggle to handle the nonlinear and complicated correlations seen in air quality data. Machine learning (ML) algorithms are a viable alternative because they can learn patterns from existing data and forecast future pollution levels. This review article provides an overview of several machine learning methodologies used for air quality prediction, such as artificial neural networks, regression models, random forest, and other data-driven methods. The importance of data preparation, feature selection, and handling missing values in increasing prediction performance is also addressed. Existing research demonstrates that machine learning models may produce trustworthy prediction results; nevertheless, their success is dependent on dataset properties, ambient variables, and appropriate model selection. The study focuses on present obstacles and future research areas for creating efficient and effective machine learning-based air quality prediction systems.

Keywords: Air quality prediction, machine learning, artificial neural network, random forest, regression model, PM2.5 forecasting, environmental monitoring.

I. INTRODUCTION

Air pollution is one of the most serious environmental issues affecting human health, ecosystems, and climatic conditions. Pollution levels, including PM2.5 and PM10, NOx, SO_2, and CO, have raised global worries. Continuous exposure to these contaminants can cause respiratory illnesses, cardiovascular issues, and environmental deterioration [2]. Pollution concentrations are typically measured using air quality monitoring stations. However, setting up and maintaining a large number of monitoring stations is costly and complex, particularly in poor countries. As a result, prediction-based systems have gained popularity for projecting future air quality levels based on historical data and environmental characteristics.

Conventional statistical approaches have been used to forecast air pollution; however, these models often perform poorly when dealing with complicated nonlinear interactions between meteorological conditions, emission sources, and pollutant concentrations. Machine learning approaches bypass these restrictions by automatically learning patterns from data and making predictions without the need for predetermined mathematical equations [3]. Air quality predictions have been made using a variety of machine learning methods. Artificial neural networks (ANNs) have become widely employed due to their capacity to simulate nonlinear correlations between input variables and pollution levels [6]. Regression-based techniques give simple and efficient prediction models, but random forest algorithms perform better with complicated datasets and missing data [9], [16].

Recent research have shown that machine learning algorithms can accurately estimate PM2.5 levels, the air quality index (AQI), and other contaminants. However, the accuracy of these models is determined by a variety of factors, including data availability, preprocessing methods, feature selection, and environmental fluctuations. A decent machine learning technique must produce accurate predictions while remaining computationally efficient and practical. This study examines several machine learning algorithms used in air quality prediction, analyses existing research, and highlights obstacles and potential prospects for creating superior prediction frameworks.

II. LITERATURE REVIEW

Machine learning has been intensively researched for air quality prediction due to its capacity to analyse vast environmental datasets and discover complicated patterns. Researchers tested several methods to increase predicting accuracy and dependability. Sharma et al. studied artificial neural network-based techniques for vehicular pollution modelling and found that ANN models can accurately depict nonlinear correlations between vehicle emissions and air pollutant concentrations [6]. Cabaneros et al. evaluated various artificial neural network models for ambient air pollution prediction and concluded that neural network-based approaches outperform traditional methods for complicated pollution datasets [3].

Feng et al. created an artificial neural network model for PM2.5 forecasting by combining air mass trajectory data with wavelet processing techniques. Their findings demonstrated that integrating machine learning with appropriate feature extraction approaches can increase prediction performance [4]. Freeman et al. investigated air quality time-series forecasting approaches and concluded that data-driven models may accurately capture temporal fluctuations in pollutant concentration patterns [5]. Regression-based machine learning approaches have also been used to forecast air pollution. Rybarczyk and Zalakeviciute investigated regression models employing low-cost air quality monitoring data and found that sensor information may be efficiently used for pollution assessment [16].

Garg and Jindal investigated various forecasting models for PM2.5 prediction and concluded that choosing a suitable model based on dataset features is critical for improved performance [11]. Random forest algorithms have gained popularity because to their ability to handle complicated datasets while improving forecast reliability. Alsaber et al. used a random forest strategy to manage missing values in air quality monitoring datasets, demonstrating its usefulness in enhancing data quality prior to prediction modelling [9]. Zainuri et al. investigated several missing data imputation strategies and concluded that preprocessing procedures had a substantial impact on the performance of air quality prediction models [10].

Kumar and Pande used machine learning algorithms to forecast air pollution in Indian cities, demonstrating that local environmental factors influence prediction accuracy [12]. Gupta et al. conducted a comparative investigation of several machine learning methods for AQI prediction and found that algorithm selection has a substantial influence on predicting outcomes [17]. Shahbazi et al. looked at machine learning approaches for enhancing air quality forecasting and emphasised the need of dependable prediction systems for environmental monitoring and pollution control [1]. Overall, prior research suggests that machine learning algorithms have a high promise for air quality prediction, but obstacles such as data quality, model generalisation, and computing efficiency necessitate additional examination.

III. MACHINE LEARNING ALGORITHMS FOR AIR QUALITY PREDICTION

Machine learning algorithms are commonly employed to forecast air quality because they can learn complicated patterns from previous pollution and environmental data. Various methods have been used, based on the type of dataset, prediction purpose, and necessary precision. The most popular machine learning algorithms for air quality prediction are given here.

The Artificial Neural Network (ANN) is one of the most used machine learning algorithms for predicting air quality. ANN is made up of linked processing units known as neurones, which learn the correlations between input parameters and output values. Because of their capacity to represent nonlinear interactions, artificial neural networks (ANNs) have been effectively used to forecast pollutants such as PM2.5, PM10, and AQI [3-6]. However, ANN performance is dependent on the amount and quality of the training data, as well as the right selection of network parameters.

Regression models are simple and effective machine learning methods for estimating correlations between air contaminants and affecting variables including temperature, humidity, wind speed, and emission sources. In comparison to sophisticated algorithms, these models have a lower computing cost and are easier to comprehend. However, their effectiveness may suffer when air quality data exhibits highly nonlinear behaviour [16].

Random Forest is an ensemble-based machine learning system that mixes numerous decision trees to increase prediction accuracy while minimising overfitting. It handles huge datasets, many input variables, and missing data well. Because of its resilience, random forest has been used for air quality data analysis and preprocessing tasks, including dealing with insufficient monitoring data [9].

The Support Vector Machine (SVM) is a supervised learning method that separates data patterns by determining an appropriate decision boundary. It is appropriate for air quality prediction tasks that need limited datasets and high-dimensional characteristics. SVM may give strong prediction performance; nevertheless, choosing the right kernel functions and parameters is critical for getting better results.

Overall, machine learning methods improve air quality prediction by finding meaningful patterns from environmental data. The right algorithm is chosen based on data availability, pollutant characteristics, forecast needs, and processing limits. Because no single method performs optimally across all air quality datasets, algorithm selection and optimisation continue to be major research problems. [1], [3]

IV. COMPARATIVE ANALYSIS OF EXISTING WORK

Various machine learning methods have been used to estimate air quality, with each methodology having its own set of advantages and drawbacks. Artificial neural networks (ANNs) are one of the most widely utilised methodologies because they can learn complicated nonlinear correlations between input parameters and pollutant concentrations. To improve performance, ANN models typically require substantial datasets and correct parameter optimisation [3], [6]. Regression models are basic, computationally efficient, and straightforward to understand. They are beneficial for identifying the links between environmental factors and pollution levels, but their accuracy may suffer when the data contains complicated nonlinear patterns [16]. Random forest models are more capable of managing huge datasets, missing values, and various input variables, making them appropriate for realistic air quality monitoring applications [9].

Several research have attempted to improve prediction performance using improved preprocessing approaches. Missing data processing, normalisation, and feature selection all contribute significantly to machine learning model accuracy. Poor-quality input data can have a considerable impact on prediction outcomes, even when sophisticated algorithms are used [9], [10]. Table 1 summarises the performance comparisons of several machine learning algorithms described in prior publications.

Reference	**Machine Learning Method**	**Application Area**	**Major Findings**
Sharma et al. [6]	Artificial Neural Network	Vehicular pollution prediction	ANN effectively captured nonlinear pollution patterns
Feng et al. [4]	ANN with feature transformation	PM2.5 forecasting	Feature extraction improved prediction accuracy
Freeman et al. [5]	Machine learning-based forecasting	Air quality time series	Data-driven models successfully predicted temporal changes
Alsaber et al. [9]	Random Forest	Air quality dataset improvement	Random forest improved handling of missing data
Garg and Jindal [11]	Time-series forecasting models	PM2.5 prediction	Model selection strongly influenced prediction performance
Kumar and Pande [12]	Machine learning algorithms	Indian city air pollution	Environmental and location factors affected prediction accuracy
Gupta et al. [17]	Comparative ML analysis	AQI prediction	Different algorithms showed different prediction capabilities

The following analysis shows that no single machine learning approach performs well across all air quality datasets. ANN models are useful for nonlinear prediction, regression models are simple, and random forest models are resilient for complicated datasets. As a result, the selection of an appropriate algorithm is determined by data properties, prediction needs, and application conditions.

V. Conclusion and Future Scope

Machine learning approaches have demonstrated great promise for air quality prediction by providing effective solutions for analysing complicated environmental information. Several techniques, including artificial neural networks, regression models, and random forest algorithms, have been used successfully to forecast pollutants like PM2.5 and AQI. ANN-based models can capture nonlinear interactions between environmental factors, whereas regression and random forest approaches provide benefits in terms of simplicity, robustness, and computing efficiency [3], [6], [9], [16].

Machine learning models' success is determined by a variety of parameters, including data quality, feature selection, preprocessing methods, and environmental circumstances. Existing research indicates that no single method can deliver the most accurate prediction results for all air quality datasets. As a result, future research should focus on creating better machine learning frameworks that strike a compromise between prediction accuracy, computational efficiency, and practical application. The combination of real-time data, meteorological information, and modern data preprocessing algorithms can improve forecast accuracy.

REFERENCES

1. Z. Shahbazi, Za. Shahbazi, and S. Nowaczyk, "Enhancing Air Quality Forecasting Using Machine Learning Techniques," *IEEE Access*, vol. 12, pp. 197290–197299, 2024.
2. I. F. Gheorghe and B. Ion, "The effects of air pollutants on vegetation and the role of vegetation in reducing atmospheric pollution," in *The Impact of Air Pollution on Health, Economy, Environment and Agricultural Sources*, pp. 241–280, 2011, doi: 10.5772/17660.
3. S. M. Cabaneros, J. K. Calautit, and B. R. Hughes, "A review of artificial neural network models for ambient air pollution prediction," *Environmental Modelling & Software*, vol. 119, pp. 285–304, 2019, doi: 10.1016/j.envsoft.2019.06.014.
4. X. Feng, Q. Li, Y. Zhu, J. Hou, L. Jin, and J. Wang, "Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation," *Atmospheric Environment*, vol. 107, pp. 118–128, 2015, doi: 10.1016/j.atmosenv.2015.02.030.
5. B. S. Freeman, G. Taylor, B. Gharabaghi, and J. Thé, "Forecasting air quality time series using deep learning," *Journal of the Air & Waste Management Association*, vol. 68, no. 8, pp. 866–886, 2018, doi: 10.1080/10962247.2018.1459956.
6. N. Sharma, K. K. Chaudhry, and C. V. Rao, "Vehicular pollution modeling using artificial neural network technique: a review," *National Physical Laboratory Research Publication*, 2005.
7. H. T. Shahraiyni, S. Sodoudi, A. Kerschbaumer, and U. Cubasch, "New technique for ranking of air pollution monitoring stations in the urban areas based upon spatial representativity (Case study: PM monitoring stations in Berlin)," *Aerosol and Air Quality Research*, vol. 15, no. 2, pp. 743–748, 2015, doi: 10.4209/aaqr.2014.12.0317.

8. G. Narkhede, A. Hiwale, B. Tidke, and C. Khadse, "Novel MIA-LSTM Deep Learning Hybrid Model with Data Preprocessing for Forecasting of PM2.5," *Algorithms*, vol. 16, no. 1, p. 52, 2023, doi: 10.3390/a16010052.
9. A. R. Alsaber, J. Pan, and A. Al-Hurban, "Handling complex missing data using random forest approach for an air quality monitoring dataset: a case study of Kuwait environmental data (2012 to 2018)," *International Journal of Environmental Research and Public Health*, vol. 18, no. 3, p. 1333, 2021, doi: 10.3390/ijerph18031333.
10. N. A. Zainuri, A. A. Jemain, and N. Muda, "A comparison of various imputation methods for missing values in air quality data," *Sains Malaysiana*, vol. 44, no. 3, pp. 449–456, 2015, doi: 10.17576/jsm-2015-4403-17.
11. S. Garg and H. Jindal, "Evaluation of Time Series Forecasting Models for Estimation of PM2.5 Levels in Air," in *2021 6th International Conference for Convergence in Technology (I2CT)*, pp. 1–8, 2021.
12. K. Kumar and B. P. Pande, "Air Pollution Prediction with Machine Learning: A Case Study of Indian Cities," *International Journal of Environmental Science and Technology*, vol. 20, no. 5, pp. 5333–5348, 2023.
13. P. K. Maduri, P. Dhiman, C. Chaturvedi, and A. Rai, "Air Pollution Index Prediction: A Machine Learning Approach," in *Proceedings of Second International Conference in Mechanical and Energy Technology*, vol. 290, pp. 37–51, 2023.
14. L. L. S. Sachetti and V. F. S. Mota, "Time Series Prediction about Air Quality using LSTM-Based Models: A Systematic Mapping," *IEEE Xplore*, accessed Jan. 2025.
15. G. I. Drewil and R. J. Al-Bahadili, "Air pollution prediction using LSTM deep learning and metaheuristics algorithms," *Measurement: Sensors*, vol. 24, p. 100546, 2022, doi: 10.1016/j.measen.2022.100546.
16. Y. Rybarczyk and R. Zalakeviciute, "Regression Models to Predict Air Pollution from Affordable Data Collections," in *Machine Learning - Advanced Techniques and Emerging Applications*, 2017, doi: 10.5772/INTECHOPEN.71848.
17. N. S. Gupta, Y. Mohta, K. Heda, R. Armaan, B. Valarmathi, and G. Arulkumaran, "Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis," *Journal of Environmental and Public Health*, vol. 2023, Article ID 4916267, 2023, doi: 10.1155/2023/4916267.
18. G. S. Vidya and V. S. Hari, "Gold Price Prediction and Modelling using Deep Learning Techniques," in *2020 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, pp. 28–31, 2020, doi: 10.1109/RAICS51191.2020.9332471.
19. L. J. Stamenković, D. Z. Antanasijević, M. Đ. Ristić, A. A. Perić-Grujić, and V. V. Pocajt, "Prediction of nitrogen oxides emissions at the national level based on optimized artificial neural network model," *Air Quality, Atmosphere & Health*, vol. 10, no. 1, pp. 15–23, 2017.

20. G. N. Kouziokas, “The application of artificial intelligence in public administration for forecasting high crime risk transportation areas in urban environment,” *Transportation Research Procedia*, vol. 24, pp. 467–473, 2017.
21. B. Mahapatra and M. Irfan, “Asymmetric impacts of energy efficiency on carbon emissions: A comparative analysis between developed and developing economies,” *Energy*, vol. 227, Article 120485, 2021.
22. “Dinesh Bhardwaj and Dr. Sonawane Vijay Ramnath” A Comprehensive Review of Clustering Techniques: Methods, Applications, and Challenges, Journal of Innovative Engineering and Research (JIER), Vol. - 7, Issue -2, , pp. 1-8, October, 2024
23. C. Ramakrishna and Dr. Pramod Pandurang Jadhav” Machine Learning Based Sentiment Analysis Implementation for Multiple Reviews Classification, , Journal of Innovative Engineering and Research (JIER) Vol. - 7, Issue -2, pp. 21-25, October, 2024.
24. Sarkar Prasanjeet Jyotirmay and Dr. Santosh Pawar” Advancing Early Detection and Prediction of Diabetes Mellitus: A Robust Soft Voting Ensemble Machine Learning Approach Journal of Innovative Engineering and Research (JIER), Vol.- 7, Issue -1,pp. 12-17, April 2024.
25. Somya Dubey And Dr. Rajeev G. Vishwakarma” Data Sensitive Security Model using Transfer Learning for Healthcare Data using Biomedical Natural Language Processing, Journal of Innovative Engineering and Research (JIER), Vol.- 5, Issue - 2, pp. 1-4, October 2022.
26. Praveen Sharma and Dr. Deepika Pathak” A Novel Machine Learning Categorial Algorithm with Remote Detecting and GIS Based Decision Support System to Identify Disaster, Journal of Innovative Engineering and Research (JIER), Vol.- 5, Issue -2, pp. 25-29, October 2022.
27. “Mohammed Ali Shaik , Dhanraj Verma , Santosh Pawar and Ritesh Yadav” Predicting Diabetes through Machine Learning, Journal of Innovative Engineering and Research (JIER), Vol.- 4,Issue -2,pp. 6-10, October 2021.
28. “Hemlata Pate and Dhanraj Verma” Data Classification in Web Usage Mining using SVM, SVR and K-NN, Journal of Innovative Engineering and Research (JIER), Vol.- 4,Issue -2, pp. 18-22, October 2021.
29. “Kailash Patidar and Dr. Dhanraj Verma” A Survey of Evolutionary Algorithms for Data Mining, Journal of Innovative Engineering and Research (JIER), Vol.- 4,Issue -1,pp. 1-3, April 2021.
30. “Sami Abdul Qader Mohammed Al – Ademi and Dr. Sanjay Singh Bhadoriya” Study on Data Mining Tools Open Source, Journal of Innovative Engineering and Research (JIER), Vol.- 4,Issue -1,pp. 15-18, April 2021.
31. “Jaya Sharma and Dr. Sanjay Singh Bhadoriya” A Basic Study on Cloud Security Challenges, Journal of Innovative Engineering and Research (JIER), Vol.- 3,Issue - 2, pp. 12-15, October 2020.
32. Dhanraj Verma and Rameswar Singh Sikarwar” K-Nearest Neighbor Based Approach for Early Prediction of

Cardiovascular Disease, Journal of Innovative Engineering and Research (JIER), Vol.- 3,Issue - 1, pp. 7-11, April 2020.