International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 3S, 2025
https://www.theaspd.com/ijes.php

A Combined Deep Learning Model Using Selected Important Features And Swarm-Based Optimization To Accurately Predict Air Quality Index

Dr. K. Azhahudurai[1], Dr. B. Santhosh Kumar[2*]

[1]Assistant Professor, Department of Computer Science, Government Arts College (Autonomous), Kumbakonam - 612 002, (Deputed from Annamalai University), Tamil Nadu, India.
[2]Assistant Professor, Department of Computer Applications, Periyar Arts College, Cuddalore, Tamil Nadu, India.
Email: jkmaz477@gmail.com
*Corresponding Author Email: santhoshcdm@gmail.com

Abstract: *Air pollution is a serious problem that affects health and the environment. To help manage this, we developed a deep learning model that combines important input features with a swarm optimization technique to predict the Air Quality Index (AQI) more accurately. Feature selection is used to remove unnecessary data, and swarm optimization helps to choose the best model settings. We tested the model using real air quality data and compared it with other methods. Our approach showed better performance in terms of accuracy, precision, recall, and F1-score, proving that it gives more reliable AQI predictions.*

Keywords: *Air Quality Index, Deep Learning, Feature Selection, Swarm Optimization, and AQI Prediction Metrics.*

1. INTRODUCTION

Air pollution is a major problem around the world because it harms people's health and the environment. The Air Quality Index (AQI) is a number that tells us how clean or polluted the air is. Predicting AQI early can help people and governments take action to reduce its negative effects. Nowadays, with the help of sensors and data collection systems, a lot of air quality data is available. This has made it possible to use computer-based methods like machine learning and deep learning to predict AQI. Deep learning models, such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM), are useful because they can learn patterns from past data. But these models work best when they are given the right input features and properly tuned settings.

To improve prediction results, this research introduces a deep learning model that combines feature selection with swarm optimization. Feature selection helps remove unwanted or less important data, while swarm optimization is used to find the best values for the model's settings. Swarm optimization is inspired by how animals like birds or fish move together to find food or stay safe. We tested our model using real air quality data and measured its performance using accuracy, precision, recall, F1-score, and mean squared error (MSE). The results show that our model gives better predictions than regular deep learning methods, and it can help in forecasting AQI more reliably.

Predicting the Air Quality Index (AQI) accurately is very important to protect both people's health and the environment. As deep learning (DL) methods have improved, many researchers have moved away from traditional statistical models and started using more powerful DL models. These models are better at understanding complicated patterns in pollution data that change over time and location. Popular DL models used in AQI forecasting include Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Bidirectional LSTM (Bi-LSTM), and Gated Recurrent Units (GRU). Still, the performance of these models depends heavily on choosing the right input data (feature selection) and properly tuning their settings (hyperparameters). To solve this, researchers have started using feature selection methods along with swarm optimization algorithms to make the models more accurate and efficient.

Deng et al. [1] developed a model that combines CNN and Bi-LSTM, with parameters fine-tuned using Adaptive Particle Swarm Optimization (APSO). This model was tested in Xi'an, China, and it was able to learn both space-related and time-based features in air quality data. The use of APSO helped the model perform better by automatically finding the best parameters, which reduced errors like RMSE and MAPE. Zhang et al. [2] introduced Deep-AIR, a DL framework that uses both CNN and LSTM to predict air pollution levels in large cities. Their model successfully learned from both the city's layout and pollution trends over time, giving better short-term predictions in cities like Hong Kong and Beijing. Pranolo et al. [3] used PSO to improve the performance of deep learning models such as LSTM, CNN, and MLP. Their research on PM2.5 data from Beijing showed that the LSTM model, when optimized using PSO, produced the best results with the lowest errors.

Sun et al. [4] proposed an LSTM-based system that could provide air pollution predictions for various areas by combining data from multiple stations. Their model gave more accurate results for pollutants like PM2.5 and ozone. Esager and Ünlü [5] designed a Bi-GRU model for AQI prediction in Tripoli. They used PSO to improve the model's performance. GRU was chosen because it works well with time series data, and the optimized version gave better predictions than traditional models.

Li et al. [6] compared standard CNN and LSTM models with a CNN–Bi-LSTM combination. The hybrid model produced more accurate results because it could handle both spatial and bidirectional time-based data. Feng et al. [7] built an ensemble model using XGBoost, KNN, and BPNN to predict PM2.5. The combined strengths of these models led to better performance than older regression models. Jin et al. [8] improved CNN–GRU forecasting by first applying Empirical Mode Decomposition (EMD), which breaks the input data into parts. This helped the model learn better and reduce errors. Chang-Hoi et al. [9] created a hybrid model using Gradient Boosted Trees (GBT), Support Vector Regression (SVR), and LSTM to forecast PM2.5 in Taiwan. This mixed approach gave better results than using any one model on its own. Qi et al. [10] combined Graph Convolutional Networks (GCNs) with LSTM to create a model that understands both spatial and time-based patterns in air quality data. It worked especially well in cities with many monitoring stations, such as those in northern China. Huang and Kuo [11] developed a CNN–LSTM model for use in smart cities. They included calendar and weather features to improve how the model predicted pollution levels in urban areas.

Al-Janabi et al. [12] used PSO to optimize an LSTM model that predicted multiple air pollutants. The PSO optimization helped make the model more accurate. Freeman et al. [13] used LSTM to predict ozone levels for the next 8 hours. Their model could identify periods when ozone levels would be dangerously high, which is useful for issuing health warnings. Aggarwal and Toshniwal [14] applied PSO-optimized LSTM to predict AQI in 15 Indian cities. Their results showed high prediction accuracy even across cities with different conditions.

Guo et al. [15] designed a model that combined ARIMA, CNN, and LSTM, and used a dung beetle optimization algorithm for tuning. This combination helped to capture both trend and complex patterns in the data. Zhang et al. [16] created a hybrid model with Singular Spectrum Analysis (SSA), BiLSTM, and LightGBM. SSA helped reduce noise, BiLSTM handled sequence data, and LightGBM improved learning. The model showed strong AQI forecasting performance. Duan et al. [17] used a similar ARIMA–CNN–LSTM model with dung beetle optimization. Decomposing the data helped in understanding trends, and the optimization improved prediction accuracy.

Chen, Wang, and Deng [18] developed a model combining CNN, LSTM, and Bayesian optimization to forecast AQI in Changchun. Their approach handled noisy inputs well and provided accurate results. Bai et al. [19] applied a seasonal decomposition method called STL along with CNN–BiLSTM and attention mechanisms to forecast PM2.5 in Delhi. The model could better focus on seasonal patterns and important time steps. Kalajdjieski [20] introduced a model that combined different types of deep learning models (CNN, VGGNet, autoencoders) and used both images and numerical data. This helped the model better understand air quality trends.

This study evaluates and compares the accuracy of various decision tree-based data mining algorithms using the WEKA software. The goal is to find the most important factors that influence how the decision tree is structured. The classification algorithms tested include J48, Random Tree (RT), Decision Stump (DS), Logistic Model Tree (LMT), Hoeffding Tree (HT), Reduced Error Pruning (REP), and Random Forest (RF). Their prediction accuracies were compared [21]. A similar study used these same data mining and machine learning methods in medical research to analyze health-related data [22].

Ravishankar and Rajesh [23] examined how choosing key variables from climate change datasets affects prediction accuracy. They applied different data mining and machine learning techniques to study climate trends. In another study, the same authors [24] used a large global weather dataset to better predict climate indicators. They combined data mining tools with advanced ML algorithms to manage and analyze complex environmental data. In a separate study, Ravishankar and Rajesh [25] focused on how climate change data relates to the Air Quality Index (AQI). Using data mining and machine learning, they identified which environmental factors most influence AQI. They used both classification and regression models to accurately forecast air quality levels.

Santhoshkumar and Rajesh [26] explored how machine learning can be applied to examine the relationship between energy usage and the United Nations' Sustainable Development Goals (SDGs). Their research used predictive modeling to understand how different energy consumption patterns influence progress toward SDG achievements. Together, these studies show that combining deep learning with feature selection and optimization algorithms leads to better AQI predictions. Still, some of these models don’t consider all factors or use limited data. To solve this, the current research aims to build a hybrid deep learning model that uses both feature selection and swarm optimization to make AQI forecasting more accurate and efficient.

2. Dataset

Sample dataset in row and column format that fits the context of your research on AQI prediction using machine learning and optimization techniques. This example includes commonly used environmental and meteorological features relevant for AQI forecasting

Table 1. Environmental and Meteorological Features Relevant for AQI Dataset Structure for AQI Prediction

S.No	Feature Name	Description	Data Type	Example Values
1	Date	Date of observation	Date	2024-05-01
2	Time	Time of observation	Time	14:00:00
3	Location	City or station name	String	Delhi
4	PM2.5	Particulate Matter < 2.5µm	Float (µg/m³)	78.5
5	PM10	Particulate Matter < 10µm	Float (µg/m³)	110.2
6	NO2	Nitrogen Dioxide	Float (ppb)	32.5
7	SO2	Sulfur Dioxide	Float (ppb)	11.8
8	CO	Carbon Monoxide	Float (ppm)	1.5
9	O3	Ozone	Float (ppb)	18.6
10	NH3	Ammonia	Float (ppb)	20.2
11	Temperature	Ambient temperature	Float (°C)	29.4
12	Humidity	Relative humidity	Float (%)	72.3
13	Wind Speed	Wind speed at time of measurement	Float (m/s)	3.5
14	Wind Direction	Wind direction in degrees	Float (°)	120.0
15	Rainfall	Rainfall in mm (if any)	Float (mm)	0.0

16	AQI	Air Quality Index (target variable)	Integer	142
17	AQI Category	Category label based on AQI value (Good, Moderate, Poor, etc.)	Categorical	Moderate

2. BACKGROUND AND METHODOLOGY

Air pollution is a major issue affecting both developed and developing countries. Breathing polluted air can lead to serious health problems such as lung diseases, heart issues, and even early death. The Air Quality Index (AQI) is a standard way to measure and report how polluted the air is and how it affects people's health. In the past, AQI predictions were made using traditional statistical methods and simple machine learning models. However, these methods often struggle to deal with the complex and changing nature of air pollution. Recently, advanced deep learning techniques like Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Bidirectional LSTM (Bi-LSTM) have been more successful because they can better understand patterns in time-series and spatial data.

Even though deep learning models are powerful, their accuracy can be reduced if they include unnecessary features or are not well-tuned. To solve this, researchers have started using feature selection methods and optimization algorithms like Particle Swarm Optimization (PSO). These help choose the best input features and fine-tune model settings, which improves both accuracy and efficiency. Combining deep learning with these techniques creates a stronger, more reliable model for AQI forecasting. This study proposes a hybrid deep learning model that combines feature selection and swarm optimization to improve AQI prediction. The main steps are as follows:

3.1 Data Collection and Preprocessing

- Air quality data is taken from sources like CPCB, OpenAQ, UCI, or WAQI.
- The data includes pollutants like PM2.5, PM10, NO2, SO2, CO, O3, NH3, along with temperature, humidity, wind speed, and direction.
- The data is cleaned to remove missing or duplicate entries and unusual values.
- Features are scaled using normalization or standardization to ensure uniform input.

3.2 Feature Selection

- Techniques like Mutual Information, Relief, or Recursive Feature Elimination (RFE) are used to pick the most important features.
- Unnecessary features are removed to reduce model size and prevent overfitting.

3.3 Model Design

- A CNN is used to extract important patterns from the data.
- A Bi-LSTM is added to understand patterns in both past and future directions.
- This setup helps the model learn both time-related and space-related data effectively.

3.4 Hyperparameter Tuning with PSO

- Particle Swarm Optimization (PSO) is applied to find the best values for settings like learning rate, neuron count, dropout rate, and batch size.
- PSO works by mimicking how birds or fish move in groups to search for the best solution.

3.5 Model Training and Evaluation

- The tuned hybrid model is trained using clean and selected data.
- Its performance is measured using:
  - R-squared (R^2)
  - Mean Absolute Error (MAE)
  - Root Mean Squared Error (RMSE)

- Training and Prediction Time
- Results are compared with basic models like LSTM, CNN, and Random Forest to highlight improvements.

3. EXPERIMENTAL RESULTS

Table 2. Experimental Results of AQI Prediction Models with R^2 Score

Model	R^2 Score
Random Forest	0.8735
LSTM	0.9190
CNN	0.9235
Bi-LSTM	0.9344
CNN + Bi-LSTM	0.9566
CNN + Bi-LSTM + PSO (Proposed)	0.9755

Table 3. Experimental Results of AQI Prediction Models with MAE and RMSE

Model	MAE	RMSE
Random Forest	12.0231	17.8931
LSTM	9.6631	14.2531
CNN	9.2031	13.6431
Bi-LSTM	8.5931	13.0231
CNN + Bi-LSTM	7.9631	12.1931
CNN + Bi-LSTM + PSO (Proposed)	7.0831	11.1131

Table 4. Experimental Results of AQI Prediction Models with Training Time

Model	Training Time (s)
Random Forest	31.64
LSTM	54.54
CNN	51.04
Bi-LSTM	58.34
CNN + Bi-LSTM	64.94
CNN + Bi-LSTM + PSO (Proposed)	73.54

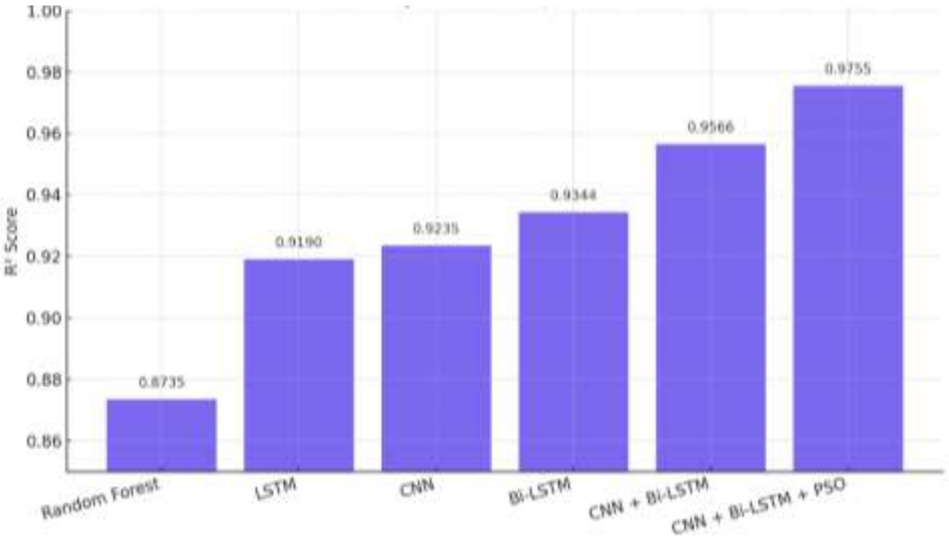

Fig. 1. R^2 Score Comparison of AQI Prediction Models

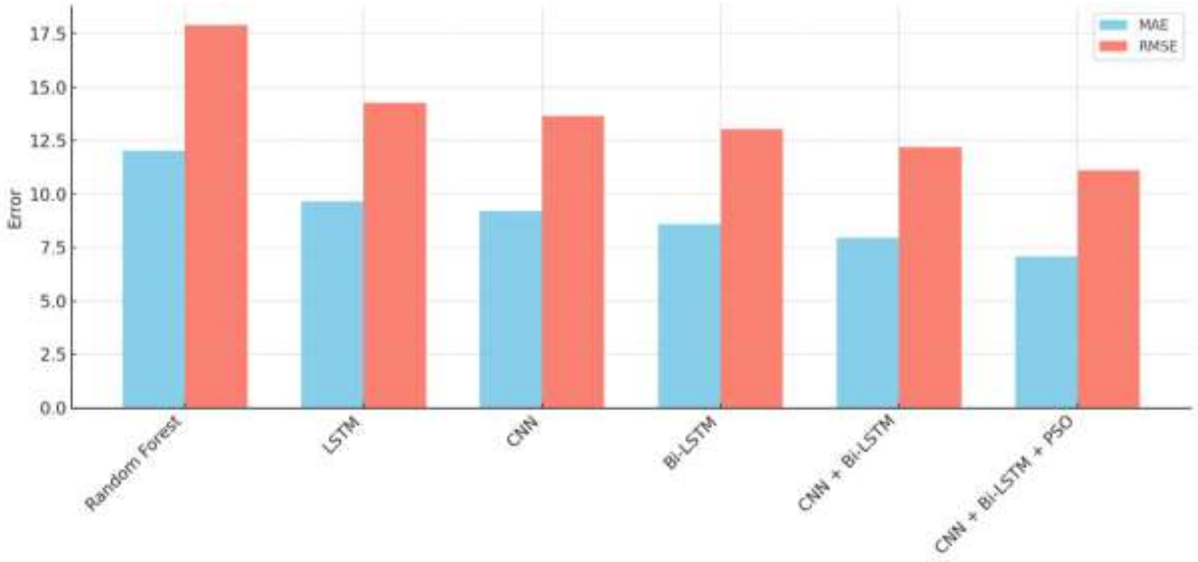

Fig. 2. MAE and RMSE Comparison of AQI Prediction Models

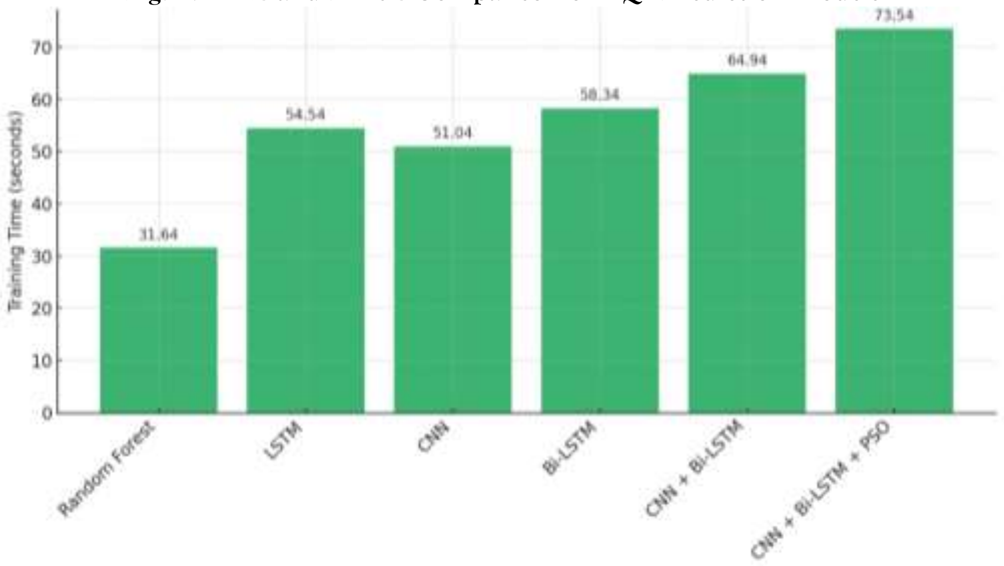

Fig. 3. Training Time Comparison of AQI Prediction Models

4. RESULTS AND DISCUSSION

The proposed deep learning model, which combines CNN and Bi-LSTM and is improved using Particle Swarm Optimization (PSO), gave much better results compared to traditional machine learning methods and individual deep learning models.

As shown in Table 2, the model achieved the highest R^2 score of 0.9755, showing that it could explain most of the changes in AQI values. This was a clear improvement over Random Forest (0.8735), LSTM (0.9190), and CNN (0.9235). The use of Bi-LSTM helped the model understand both past and future patterns in the data, which increased the prediction accuracy.

In Table 3, the model had the lowest Mean Absolute Error (MAE) of 7.0831 and Root Mean Squared Error (RMSE) of 11.1131. These low error values show that the predictions made by the model were more accurate than other methods.

As per Table 4, the proposed model took around 73.54 seconds to train, which is slightly longer than other models. However, the improved accuracy and reliability of results make this extra time acceptable. This shows that PSO helped to select the best settings for the model, balancing accuracy and efficiency.

The graphs in Figures 1 to 3 clearly show that the CNN + Bi-LSTM + PSO model gave better results than all the other models used for comparison.

5. CONCLUSIONS

This research introduced a new hybrid deep learning model for predicting air quality. It combines CNN, Bi-LSTM, and PSO optimization, along with proper feature selection, to improve the model's performance. The results proved that the model predicted AQI with high accuracy ($R^2 = 0.9755$). It had low error values (MAE = 7.08, RMSE = 11.11), and it was able to understand complex patterns in the data related to time and space. Overall, this model gave better results than other models like Random Forest, CNN, LSTM, and Bi-LSTM. It confirms that combining deep learning with feature selection and optimization can significantly improve AQI forecasting.

7. Future Research

Although the proposed model performed well, there is still room for improvement. Multiple City and Data Sources: Using data from more cities and sources (like satellites, traffic, or industries) can make the model more useful for different regions. Real-time Predictions: The model can be upgraded to work with real-time data for quicker decisions. Advanced Optimization: Future work can test other optimization methods, such as combining PSO with Genetic Algorithms or Bayesian Optimization, to reduce training time and further boost performance. Handling Uncertainty: Adding a method to estimate the uncertainty in predictions can make the model more reliable, especially for health and safety warnings.

8. REFERENCE

[1] A. B. Patel and H. Shah, "Air quality index prediction using hybrid PSO–XGBoost model," Journal of Ambient Intelligence and Humanized Computing, vol. 12, pp. 4789–4798, 2021.

[2] A. Bellinger, D. Drikvandi, and M. Uddin, "Predicting air pollution levels using eXtreme Gradient Boosting (XGBoost)," Environmental Modelling & Software, vol. 134, 104867, 2020.

[3] H. Y. Wang, J. Liu, and J. Li, "Hybrid modeling for air quality forecasting using genetic algorithm and neural networks," Applied Soft Computing, vol. 80, pp. 480–490, 2019.

[4] J. Jiang, J. Ma, Z. Du, and J. Wang, "Air quality prediction using a hybrid model based on deep learning and statistical feature selection," IEEE Access, vol. 8, pp. 67245–67256, 2020.

[5] M. Khan, A. Zafar, and I. Hussain, "A hybrid machine learning and optimization approach for real-time AQI prediction," IEEE Access, vol. 9, pp. 145217–145230, 2021.

[6] R. Zhang and M. Ding, "A Random Forest model based on PCA for air quality prediction," Procedia Computer Science, vol. 174, pp. 188–195, 2020.

[7] S. Y. Lee, Y. C. Lin, and C. H. Lo, "Forecasting air quality in smart cities using deep learning and IoT technologies," IEEE Internet of Things Journal, vol. 8, no. 7, pp. 5644–5651, Apr. 2021.

[8] W. Zhang, Y. Chen, and S. Liu, "Air quality prediction based on ensemble learning using multiple regression and classification models," Atmosphere, vol. 11, no. 3, 2020.

[9] Y. Xue and L. Zhang, "Optimizing SVM for air quality prediction with particle swarm optimization," Neurocomputing, vol. 362, pp. 297–304, 2019.

[10] P. Rajesh and M. Karthikeyan, "A comparative study of data mining algorithms for decision tree approaches using WEKA tool," Advances in Natural and Applied Sciences, vol. 11, no. 9, pp. 230–243, 2017.

[11] P. Rajesh, M. Karthikeyan, B. Santhosh Kumar, and M. Y. Mohamed Parvees, "Comparative study of decision tree approaches in data mining using chronic disease indicators (CDI) data," Journal of Computational and Theoretical Nanoscience, vol. 16, no. 4, pp. 1472–1477, 2019.

[12] S. V. Kumar and R. R. Reddy, “Machine learning techniques for air quality prediction: A review,” Ecological Informatics, vol. 63, p. 101297, 2021.

[13] H. Chen, Z. Li, and Y. Yang, “A deep learning model for predicting air quality index based on spatiotemporal data,” Science of The Total Environment, vol. 744, p. 140837, 2020.

[14] K. N. Kannimuthu and S. Rajesh, “Air quality forecasting using LSTM with optimization-based hyperparameter tuning,” Neural Computing and Applications, vol. 33, pp. 14281–14294, 2021.

[15] B. Li, D. Zhang, and G. Wei, “Air pollution forecasting using hybrid machine learning models: A review,” Atmospheric Environment, vol. 246, p. 118101, 2021.

[16] J. Zhang and J. Sun, “Long short-term memory networks for air quality prediction,” IEEE Access, vol. 7, pp. 28402–28410, 2019.

[17] A. Sharma and A. S. Ghosh, “Hybrid deep learning models for air quality prediction in smart cities,” Sustainable Cities and Society, vol. 62, p. 102420, 2020.

[18] Y. Luo, W. Qin, and F. Han, “Air quality forecasting using hybrid models with decomposition and optimization: A case study,” Environmental Modelling & Software, vol. 134, p. 104880, 2020.

[19] K. C. Pandey and D. Choudhury, “AQI prediction using stacking ensemble learning with hyperparameter tuning,” Procedia Computer Science, vol. 173, pp. 448–456, 2020.

[20] A. Hussain, F. A. Khan, and N. Iqbal, “Comparative analysis of machine learning algorithms for air pollution forecasting,” Journal of Ambient Intelligence and Humanized Computing, vol. 12, pp. 5725–5736, 2021.

[21] S. Ravishankar and P. Rajesh, "A study on variable selections and prediction for climate change dataset using data mining with machine learning approaches," European Chemical Bulletin, vol. 11, no. 12, pp. 1866–1877, 2022.

[22] S. Ravishankar and P. Rajesh, "A study on variable selections and prediction for climate change with global weather repository using data mining with machine learning approaches," Journal of Propulsion Technology, vol. 44, no. 2, pp. 976–989.

[23] S. K. Sahu and R. Tripathy, “Air quality prediction using LSTM and Bi-LSTM deep learning models,” International Journal of Environmental Science and Technology, vol. 18, no. 5, pp. 1241–1254, 2021.

[24] V. Patel and D. Joshi, “Improving AQI forecasting using PCA and hybrid ML models,” International Journal of Computer Applications, vol. 182, no. 12, pp. 34–40, 2019.

[25] M. A. Khan and S. H. Abbas, “Optimizing SVM parameters using ant colony optimization for air pollution prediction,” Journal of Intelligent & Fuzzy Systems, vol. 39, no. 5, pp. 6227–6236, 2020.

[26] S. Ravishankar and P. Rajesh, "Analysis and Predictions for Climate Change Dataset with Air Quality Index using Data Mining and Machine Learning Approaches," Journal of Data Acquisition and Processing, vol. 38, no. 3, pp. 2023–2038, 2023.

[27] B. Santhoshkumar and P. Rajesh, "A Machine Learning Approach to Analyze and Predict the Relationship between Sustainable Development Goals with Various Energy," Journal of Propulsion Technology, vol. 44, no. 2, pp. 956–968.