INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ORIGINAL ARTICLE

OPEN ACCESS

Received: 25/11/2025

Accepted: 03/12/2025

Published: 31/12/2025

Citation: Dr A Angelpreethi , Subash K, Lekka P (2025) A Hybrid Predictive Modelling of Environmental Pollutant Concentrations using Machine and Deep Learning Techniques. Indian Journal of Science and Technology 18(48): 3851-3862. https://doi.org/10.17485/IJST/v18i48.1889

* **Corresponding author**.

angelpreethi_cs1@mail.sjctni.edu; angelpreethi.mca@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

A Hybrid Predictive Modelling of Environmental Pollutant Concentrations using Machine and Deep Learning Techniques

Dr A Angelpreethi[1]*, K Subash[1], P Lekka[2]

1 Department of Computer Science, St. Joseph's College, Affiliated to Bharathidasan University Tiruchirappalli, Tamil Nadu, India
2 Department of Commerce Computer Application, St. Joseph's College, Affiliated to Bharathidasan University Tiruchirappalli, Tamil Nadu, India

Abstract

Background: Diminishing air quality leads to serious risks to human beings, animals and surrounding environment. The pollution level indicator is a measure, which measure the cleanliness of the air, and it provides information on the level of polluted air and its potential impact on human health. **Objectives:** To propose a hybrid model using automated learning systems for predicting the AQI using historical air quality data. **Methods:** This approach involves cleaning the raw data, and model training using different machine learning and deep learning algorithms such as linear regression, decision tree, random forest, and convolutional neural network. **Findings:** Among all models, the RF_CNN algorithm gives the best results with a MAE of 19.26**,** MSE of 620.67**,** RMSE of 24.92**,** and R^2 of 0.56**.** **Novelty:** This system predicts AQI by using automated learning algorithms and predict which learning algorithm produce better results. The results show that the deep learning models outperform existing machine learning models in predicting the index with best performance. The proposed approach can be used by environmental and zoological informatics, government agencies and policymakers to make data driven decisions regarding polluted air control, wild life conservation, eco system and health monitoring.

Keywords: Wild life conservation; Environmental pollution; Model evaluation; Deep Neural Networks; Decision making

1 Introduction

Atmospheric contamination is caused by a variety of pollutants that can be injurious to humans as well as environment. Some of the major air pollutants include CO, CO_2, NO_2, NO, SO_2, NH_3, particulate matter (PM 2.5 and PM 10), ozone (O_3), acrolein, asbestos, benzene, CS_2, polycyclic aromatic hydrocarbons (PAHs), synthetic vitreous fibers, and total petroleum hydrocarbons [(1)]. Monitoring the polluted air has become important due to its impact on human as well as environment [(1)]. The polluted air can cause a severe problem including respiratory, cancer, heart diseases and breathing problems. Additionally, this polluted air can harm our ecosystem, sudden climate change and agricultural production problems. Quality air is important since it helps governments, environmental sectors to understand how safe or unsafe the air is on any given day. Hence, accurately forecasting air quality is crucial for managing natural habitats, preserving wildlife, and monitoring environmental changes, in addition to safeguarding human health.

Over the years, Machine Learning (ML) and Deep Learning (DL) techniques have been widely used by many researchers for predicting the quality of the air.ML models such as Random Forest (RF)[2], Support Vector Regression[3], K-Nearest Neighbours,[4] are work well for simple patterns but they struggle with complex non-linear data. DL models such as LSTM[4] ANN[5] and CNN[6] learn complex patterns, but they depend on time series data and computationally high process. Additionally, DL models require more training data for producing better accuracy. Earlier studies use only one type of model and do not compare several ML and DL algorithms together. Some studies use hybrid model on small region-specific dataset to predict the air quality making their results less reliable. This proposed work is different from the existing work since it combines ML and DL model together and uses complex benchmark dataset, applies clear data cleaning methodologies and produce higher accuracy than the previous studies.

To overcome these research gaps, this proposed model develops a prediction framework for predicting the air quality using multiple ML and DL models and also combine them to form hybrid model. The present study used multi-year dataset from Beijing, china containing more than 43,824 records. The dataset includes pollutant values, meteorological data such as humidity, temperature, wind speed and direction of the wind. The dataset is cleaned using data pre-processing techniques.

To assess the performance of the air quality forecast models with the testing dataset, performance measures such as R-squared (R^2), Root Mean Square Error (RMSE), Mean Absolute Error (MAE) are created to combine the results of both ML and DL models[4]. The proposed results show that the hybrid models provide better predictions than the individual models. The main advancements presented in this research are outlined below:

1. Predicting air quality using computational algorithms which combines ML and DL techniques and compare its performance
2. Implementing and evaluating the models for instance LR, DT, RF, SVM, ANN and CNN using air quality datasets[7].
3. A structured data cleaning strategy that improves the performance.
4. Analysing and comparing the models using the evaluation metrics such as MAE, MSE, RMSE and R^2 score[8].

In India, industries are one of the major sources of polluted air, accounting for 51% of the total pollution[1]. Table 1 shows the average annual concentrations of some common pollution levels in Indian cities[9].

Table 1. Pollution Concentration in different regions of India

Year	City	Amount of PM2.5 ($\mu g/m^3$)	Amount of PM10 ($\mu g/m^3$)	Concentration of NO_2 ($\mu g/m^3$)	Concentration of SO_2 ($\mu g/m^3$)	CO Amount (mg/m^3)	Ozone Measure (O_3) ($\mu g/m^3$)
2024	Delhi	122	292	80	19	1.7	23
2024	Mumbai	63	128	53	14	1.2	29
2024	Kolkata	90	163	56	16	1.4	39
2024	Chennai	46	98	27	8	1.1	29
2024	Hyderabad	57	106	36	8	1.1	41
2024	Bengaluru	39	97	30	6	1.2	27
2024	Ahmedabad	89	173	41	23	1.5	41

In Table 1, PM2.5 refers to fine particulate matter with a diameter of less than 2.5 micrometres, PM10 refers to particulate matter with a diameter of less than 10 micrometres, NO_2 refers to nitrogen dioxide, SO_2 refers to sulfur dioxide, CO refers to carbon monoxide, and O_3 refers to ozone[1]. The data is collected for the year 2014 and it consists of few cities in India[10]. The AQI serves as a pointer of how clean the air is and helps measure its possible effects on human as well as health.

Accurate predictions of AQI can help the people to take necessary steps to protect their health, environment, governments and organizations to make policies to reduce polluted air. Many industries and businesses can also use AQI predictions

Table 2. Index value and its category

Index Value of the Air	Category
0-50	Acceptable
51 – 100	Suitable
101-200	Modest
201-300	Unhealthy
301-400	Highly unhealthy
401-500	Severe Needs Attention

to maintains their operations and reduce their contribution to polluted air[11]. Existing studies have proposed ML and DL approaches for AQI prediction, each approach differ in terms of data, model design, and its performance. A comparative overview of earlier works is presented in Table 3.

Table 3. Comparative Study of Existing System

Model	Author (year)	Advantages	Limitations
Deep recurrent neural networks	Singh et al. (2022)[5]	Effective in handling temporal dependencies and incorporating outside conditions	Massive data needed to process, may not capture complex non-linear relationships
Hybrid attention-based CNN-LSTM model	Karkuzhali et al. (2024)[6]	Can location, time related data and incorporate attention mechanisms for improved performance	Requires huge data for training
AI based Prediction Framework	Malhotra *et al.* (2024)[12]	Provides an extensive overview of AI-based AQI prediction techniques and highlights emerging hybrid approaches.	Does not include implementation or performance evaluation on real-world datasets.
Optimized Machine Learning model	Natarajan *et al.* (2021)[13]	Can integrate data from multiple sources for improved accuracy	May require extensive pre-processing of data
Graph convolutional neural networks	Abbasi et al. (2025)[14]	Effective in handling complex spatial relationships in the data	Requires huge data for training, may not work well with non-spatial data
Ensemble model with fuzzy logic	Ida Seraphim et al. (2024)[15]	Can combine multiple models for improved performance and incorporate uncertainty	Requires tuning of parameters, may not work well with highly correlated predictors
Time-aware graph neural networks	Pan et al. (2025)[16]	Effective in handling temporal and spatial dependencies and incorporating external factors	Often depend on large datasets and impose high computational cost

Table 3 shows that a different ML and DL models have been implemented for predicting the Air Quality Index (AQI).

2 Methodology

2.1 Data collection

The data collected from the Beijing Municipal Environmental Monitoring center which consists of 35 air quality stations which monitors across the city [https://doi.org/10.24432/C5JS49 data collected from 02.01.2010 to 31.12.2014] in UCI Machine Learning Repository. These stations are fully connected with sensors that gather hourly data on a range of air pollutants, such as PM2.5, PM10, SO2, NO2, CO, and O3. The stations additionally collect meteorological data which includes temperature, air humidity, direction of the wind and speed of the wind. The dataset consists of data from 2010 to 2014. The dataset consists of 43,824 entries each consists of 13 features such as hourly readings of six air pollutants and seven meteorological factors. This data collected for the purpose of predicting the AQI in Beijing city and prevent the general public from the air related health issues.

2.2 Data Cleaning

One of the essential processes in preparing the data for the use of ML models is data pre-processing. This step consists of transforming the unprocessed data into processed (Cleaned) data. The data cleaning phase outlines the techniques employed in this study to prepare and predict the AQI with the help of computational learning models which use ML and DL techniques. The cleaning strategies used in this study includes outlier processing, impute the missing values and selecting the features.

2.2.1. Outlier processing:

Outlier processing involves identification of data and handling of data which are dissimilar from other data in the dataset. Generally, outliers in the dataset can occur false values, faulty readings from the sensors, error transactions etc. Detecting outliers is essential so as to avoid faulty values in the cleaning phase itself. The z-score technique is used here to identify the outliers. Z-score is a statistical technique that uses mean(μ) and standard deviation (σ) to transform each input into output [17]. Here in this study, identifies how far the individual observation deviates from the average value. The outliers are identified using

$$z = \sigma^{-1}(x - \mu) \tag{1}$$

2.2.2. Missing value handling

Missing values processing refers to the identification and handling of missing data points in a dataset. Missing data can be addressed through different strategies like discarding incomplete records, estimating the absent values, or using predictive modelling [17]. A frequently applied imputation approach is to substitute the missing entries with a representative statistic typically the mean or median calculated from the available observations within that particular attribute.

$$\text{X\_imputed} = \text{X.fillna}\left(\text{X.mean}()\right) \tag{2}$$

The output will show the first few rows of the resulting DataFrame, which will give an idea of what the modified dataset looks like.

Table 4. Sample Pre-processed data

No	year	month	day	hour	pm2.5	DEWP	TEMP	PRES	cbwd	Iws	Is	Ir
1	2010	1	2	0	129.0	-16	-4.0	1020.0	SE	1.79	0	0
2	2010	1	2	1	148.0	-15	-4.0	1020.0	SE	2.68	0	0
3	2010	1	2	2	159.0	-11	-5.0	1021.0	SE	3.57	0	0
4	2010	1	2	3	181.0	-7	-5.0	1022.0	SE	5.36	1	0
5	2010	1	2	4	138.0	-7	-5.0	1022.0	SE	6.25	2	0

2.3 Classification

Selecting proper ML and DL models is essential for accurate predictions of AQI, which helps the users to make decisions about their health and animal health [2]. In this study some supervised ML techniques can be used to forecast the AQI.

2.3.1. Random Forest (RF):

This approach belongs to the family of ensemble methods. The RF can be applied to regression and classification techniques. RF is a powerful algorithm that can handle huge datasets, nonlinear relationships, and missing values [2]. RF has been widely applied in AQI forecasting research because it can efficiently manage datasets with many input features. [2].

The prediction for a sample x is obtained by taking average of the outputs of all decision trees:

$$\widehat{(y)}(x) = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \tag{3}$$

In the above equation T is the number of trees and h_t(x) is the prediction from the tth tree [2].

2.3.2. Support Vector Regression (SVR):

SVR is a supervised ML technique used for regression problems. SVR is effective when the relationship between the predictors and the target variable is nonlinear[3],[18]. This ability comes from its kernel functions. Those functions can allow the model to map input into higher-dimensional spaces for better prediction. The regression model can be written as:

$$f(x) = \sum_{i=1}^{n} \alpha_i \, K(x, x_i) + b \tag{4}$$

Here K represents the kernel function and α_i are Lagrange multipliers[3],[18].

2.3.3. Gradient Boosting Machines (GBMs):

GBMs comes under a family of ML methods that builds a sequence of weak models, and slightly improving its prediction accuracy with each iteration.[2],[19]. Every new tree focuses on reducing the residual errors produced by earlier trees, gradually improving performance. The prediction is computed as:

$$\widehat{y}(x) = y_0 + \eta S(x) \text{ where } S(x) = \sum_{m=1}^{M} h_m(x) \tag{5}$$

Where y_0 is the initial prediction, η is the learning rate, and $h_m(x)$ denotes the prediction of the m^{th} weak learner[19].

2.3.4. K-Nearest Neighbors (KNN):

The KNN method is a straightforward, instance-based learning technique that can be applied to both classification and regression problems[4]. For regression, the algorithm estimates the value of a new data point by taking the mean of the target values of the *K* closest samples, while for classification it selects the most common class among those neighbors. The prediction can be expressed as:

$$\widehat{y}(x) = \frac{1}{K} \sum_{i=1}^{K} y_i \tag{6}$$

where y_i denotes the output values associated with the *K* nearest points[4].

2.3.5. Artificial Neural Networks (ANNs):

Artificial Neural Network is designed to process similarly on how our human brain process information[5],[20]. This ANN have neurons that are used to connect and help to think and learn. These neurons also used to process the data and solve real time problems. ANN generally used for classification and regression related problems[5],[20]. In this study ANN is used for predicting the AQI since the data non-linear in nature and can have many complex dependencies. Each individual neuron in the system can be expressed as

$$y = \phi\left(\sum_{i=1}^{n} w_i x_i + m\right) \tag{7}$$

In this equation x_i are inputs, w_i are weights, m represents the bias parameter, and ϕ denotes the activation function[5],[20].

2.3.6. Convolutional Neural Networks (CNNs):

This CNN is a special type of model in deep learning framework used to process grid-based data. Generally, it is used to process images and videos. This CNN consists of convolutional layers which apply filters to extract the patterns or features[6],[20]. The convolution process is used to filter the input using dot product. The activation functions are non-linear functions specially applied to non-linear data[6]. In this study, to process the non-linear data the CNN is used. The pooling layers are used to reduce the dimensionality and increase the accuracy. The process denoted by

$$Y = g(K \otimes X + b) \tag{8}$$

In this expression, **X** denotes the input data, **K** refers to the convolutional kernels (filters), and $\otimes$ represents the convolution operation. The term **b** is the added bias, while **g** is the activation function applied to the result[5].

2.3.7. Long Short-Term Memory (LSTM) networks

LSTM networks are commonly used for time related data and can be used for AQI prediction as well [4], [20]. They include special cells connected with memory that control information using gates [4], [21]. The key operations are:

$$i_t = \sigma\left(U_i\left[h_{t-1}, x_t\right] + b_i\right. \quad (9)$$

$$f_t = \sigma\left(U_f\left[h_{t-1}, x_t\right] + b_f\right. \quad (10)$$

$$c_t = f_t \odot C_{t-1} + i_t \odot tanh\left(U_c\left[h_{t-1}, x_t\right] + b_c\right. \quad (11)$$

[4]The input gate, forget gate, and cell state at time step t are expressed by i_t, f_t, and c_t respectively. Element wise product is indicated by $\odot$ operator [4].

2.4 Predicting the Quality of the Air:

The proposed method consists of hybrid ML and DL models which are strongly used to predict the AQI. ML models can find the patterns and its relations in air quality data whereas DL models can learn the complex patterns. In this study both models are combined for accurate prediction of air quality data.

1. Initially the process starts with collecting the data from sensors and metrological stations in government sources. The data are cleaned by removing unwanted values, imputing missing values and identifying the outlier values. Finally, the data converting into the usable format.
2. After data collection, the features are identified such as humidity of the air, temperature of the air, speed and pollution levels. Those features are selected because they have the huge effect on AQI.
3. The third stage is to divide the data into training and testing so as to given as an input to the ML and DL models. Those models can learn and improve. The combination of ML and DL (hybrid model) created by combining the results of ML and DL. The hybrid model is fine-tuned by adjusting the settings of the system to improve the accuracy [20], [22]. The model is tested with new data to evaluate its accuracy. The prediction is computed by using

$$\mathrm{AQI\ (x)} = \mathrm{w}_1 * \mathrm{ML\ (Pol,\ Temp,\ Hum)} + \mathrm{w}_2 * \mathrm{DL\ (I,\ O,\ T)} \quad (12)$$

   In this equation AQI (x) is the predicted AQI. Pol, Temp, Hum are used as inputs to the ML models, corresponding to concentration of the pollutant, air temperature, and air humidity. I, O, and T indicates inputs to the DL models, corresponding to the levels of NO_2, O_3, and air temperature. The weights given to the ML and DL algorithms are indicated using w_1 and w_2 respectively.
4. Linear Regression (LR), Random Forest (RF), Support Vector Regression (SVR), K Nearest Neighbours (KNN) [18] ML algorithms are used to validate the AQI data. Similarly, ANN, LSTM and CNN are DL models are used to test the AQI data [22].

By merging their outputs and adjusting the weights, the hybrid model achieves more accurate and reliable AQI predictions.

3 Results and Discussion

This section elaborates the experimental results obtained from the ML algorithms and DL algorithms to the air quality dataset. Before training the model, the data was pre-processed and divided into training and testing datasets. Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Coefficient of Determination (R^2) [23] are the quality measures used to compute the AQI. The comparative results proves that the proposed method outperforms the existing techniques. The KNN algorithm achieves moderate performance compared to SVR which produces lower accuracy. Therefore, the hybrid model is identified as the most effective approach for predicting air quality.

Totally, the dataset can have **43,824** entries, each entry containing **17** features. The dataset features provide detailed information for classifying the air quality and patterns of weather in Beijing during a five-year period.

Table 5. Dataset Characteristics

Feature	Description
Number	Serial number assigned to each data entry.
Year	The year in which the observation was recorded.
Month	The month corresponding to the observation.
Day	The day on which the measurement was taken.
Hour	The time of observation, recorded in 24-hour format.
PM2.5	Concentration of fine particulate matter (PM2.5) in micrograms per cubic meter ($\mu g/m^3$).
PM10	Concentration of coarse particulate matter (PM10) in micrograms per cubic meter ($\mu g/m^3$).
SO_2	Level of sulfur dioxide in micrograms per cubic meter ($\mu g/m^3$).
NO_2	Level of nitrogen dioxide in micrograms per cubic meter ($\mu g/m^3$).
CO	Concentration of carbon monoxide in milligrams per cubic meter (mg/m^3).
O_3	Concentration of ozone in micrograms per cubic meter ($\mu g/m^3$).
TEMP	Air temperature in degrees Celsius (°C).
PRES	Atmospheric pressure measured in hectopascals (hPa).
DEWP	Dew point temperature in degrees Celsius (°C).
Rain	Amount of rainfall recorded (mm).
wd	Direction of the wind in degrees.
WSPM	Speed of the wind (m/s).

Figure 1 indicates how the pm2.5 values are fluctuated in the training dataset. The values are changes over time. In Figure 1 the x axis represents the index values and y axis represents the pm2.5 values. The pm2.5 values are ranging from 0 to 900 with more number of spikes which indicates outlier values. This diagrammatic representation helps the distribution and overall trend of pm2.5 levels. This representation provides clear understanding of the data patterns and better decision making for further analysis.

Figure 2 represents the pm2.5 values used in the training dataset. The values in the figure clearly states how the measurements changes over time. Even though the appearance of the plot generated by the training dataset, some points of time the testing dataset is differ. Here the x axis represents the index values and y axis represents the pm2.5 values. The variations are observed similar to training dataset. Analysing the PM2.5 values in the dataset helps to measure how well the model performs on unseen data. The test dataset values of the PM2.5 closely matches to training dataset. This indicates that the model gathered useful patterns and predict well on new data.

3.1 Evaluation Metrics

To evaluate the performance the ML and DL models, some statistical metrics were used. The following metrics were used to measure the performance of the predictive models.

3.1.1. Mean Absolute Error (MAE):

MAE is the dimension used to find the average number of errors between the actual values and predicted values [19]. This can be measured using

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_{i-} \widehat{y_i}|$$

3.1.2. Mean Squared Error (MSE):

MSE used to calculate the average of the squared differences between the acutal values and predicted values by giving higher weights to larger errors [19].

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_{i-} \widehat{y_{i\ )}}^{2}$$

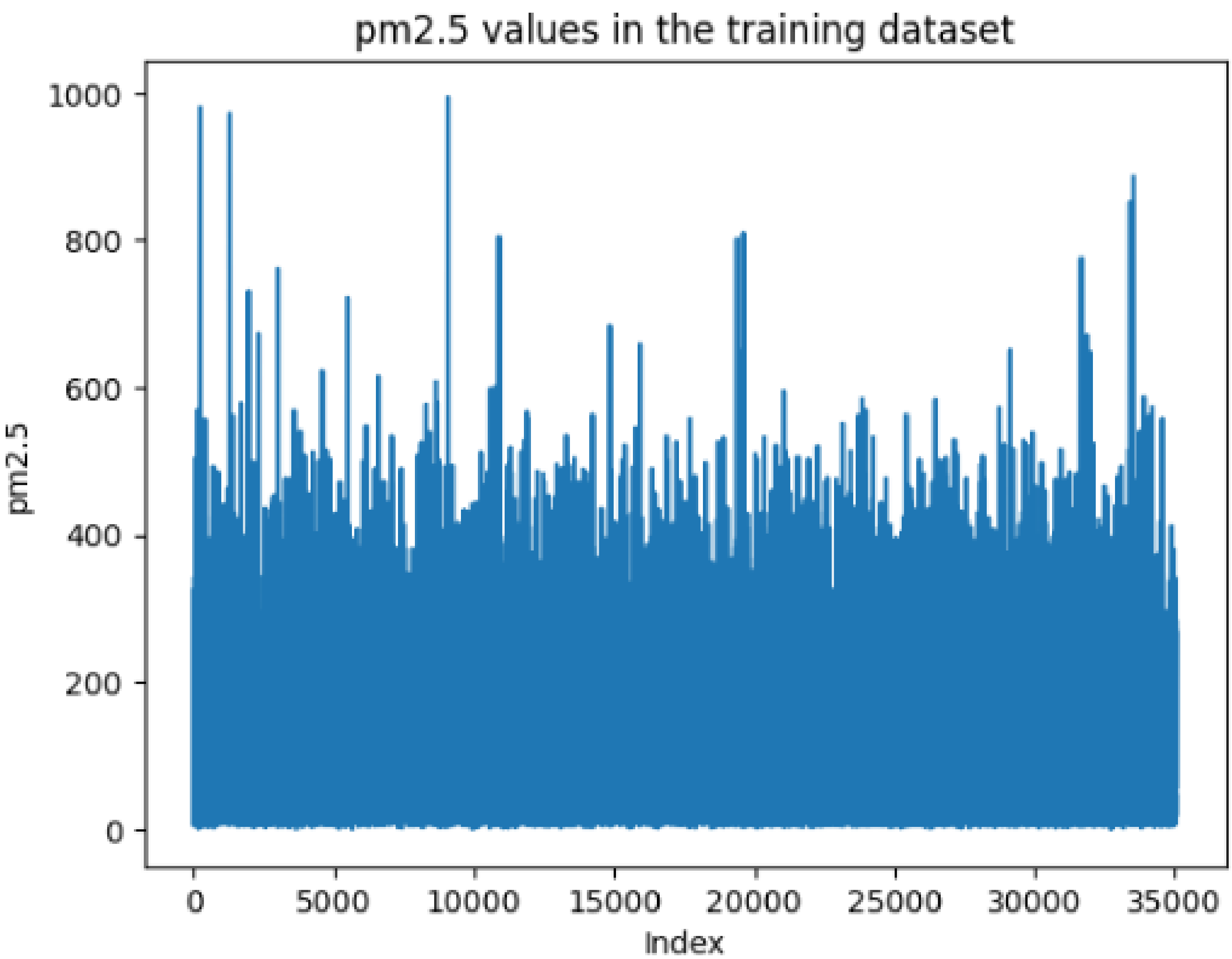

Fig 1. pm2.5 values (training dataset)

3.1.3. Root Mean Squared Error (RMSE):

RMSE is obtained by taking the square root of the MSE and reflects the typical magnitude of the prediction errors, similar to a standard deviation of the residuals [19]. $RMSE = \sqrt{MSE}$

3.1.4. Coefficient of Determination (R^2):

The R^2 statistic indicates how effectively the predictive system accounts for the variability present in the actual data [19]. In the equation 1 suggest a strong association between the predicted outcomes and the actual values.

$$R^2 = 1 - \frac{Sum\,of S_{res}}{VS_{tot}}$$

where $Sumof S_{\text{res}}$represents the total of the squared residuals, and VS_{tot}denotes the total variation in the input data. These evaluations were applied to measure the quality of prediction in each learning technique and to identify the most appropriate technique for AQI estimation.

Based on the performance indicators presented in the table, the Random Forest technique emerges as the most effective option for AQI prediction using the Beijing PM2.5 dataset [19]. It achieves the smallest error values—MAE of 21.65, MSE of 838.7, and RMSE of 28.96—along with the highest R^2 score of 0.24, outperforming the other approaches evaluated in this study.

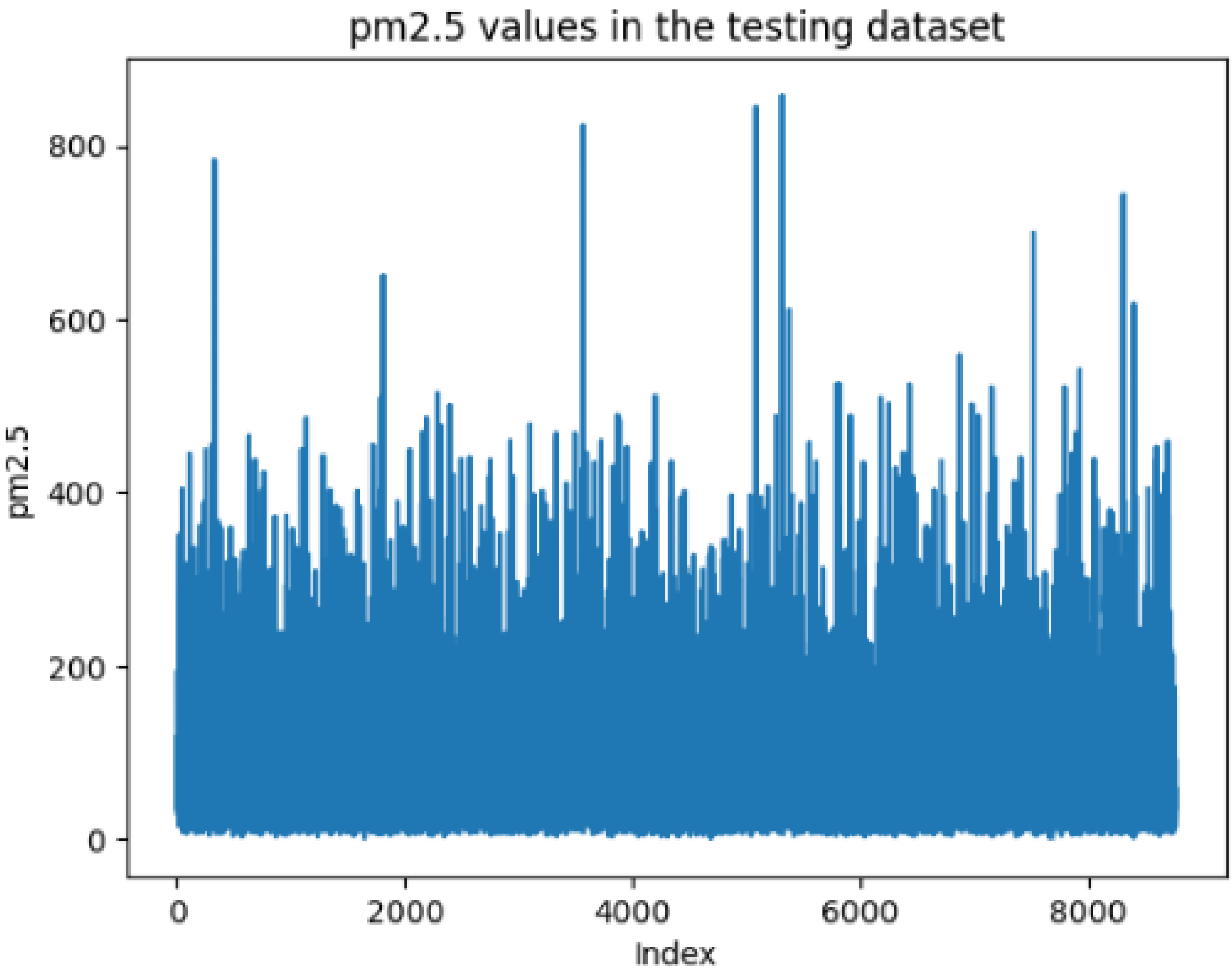

Fig 2. pm2.5 values (testing dataset)

Table 6. Performance Evaluation with single Algorithms

Model	MAE	MSE	RMSE	R^2
LR	24.13	981.69	31.31	0.24
RF	21.65	838.7	28.96	0.39
SVR	26.02	1174.85	34.26	0.11
KNN	22.98	886.66	29.78	0.34
ANN	22.67	879.33	29.64	0.35
LSTM	21.96	869.68	29.49	0.36
CNN	21.8	863.06	29.36	0.37

3.2 Comparative Analysis

These metrics indicate that the RF model provides the most accurate predictions of AQI values compared to the other models.

Table 7. Comparative Assessment with ML_CNN Algorithms

Model	MAE	MSE	RMSE	R^2
LR _CNN	19.82	652.97	25.59	0.53
RF_CNN	19.26	620.67	24.92	0.56
SVR_CNN	20.71	746.11	27.32	0.45
KNN_CNN	20.5	731.33	27.03	0.47
ANN_CNN	19.58	633.82	25.17	0.54

Based on the assessment metrics in the Table 7, the RF+ CNN algorithm is the suitable algorithm for predicting AQI using the Beijing PM2.5 Data among the evaluated algorithms. The RF_CNN algorithm produces MAE of 19.26, MSE of 620.67, RMSE of 24.92 and R^2 of 0.56. While comparing these values with other values, RF_CNN produce lower MAE, MSE, RMSE and higher R^2. This means that the RF_CNN accurately predicts the AQI than the other models.

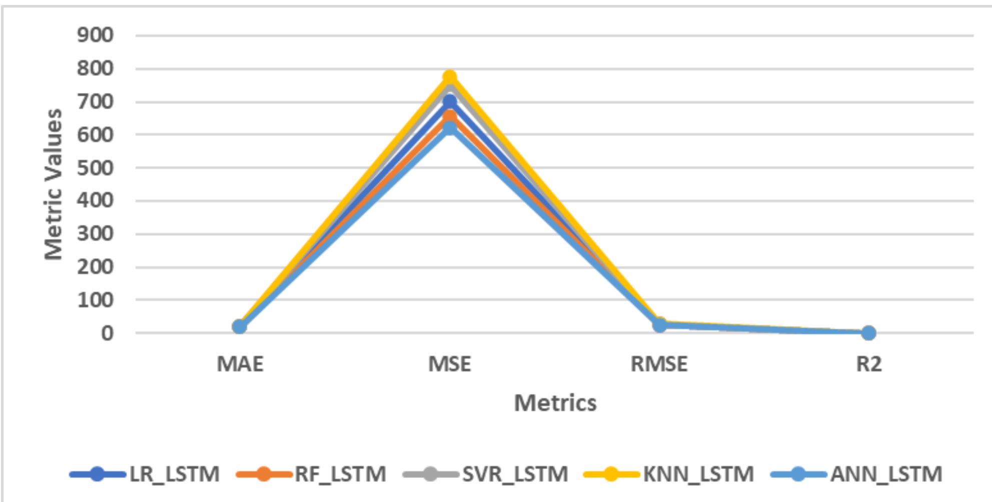

Fig 3. Performance Analysis of ML and LSTM Algorithm

Figure 3 represents the performance analysis of ML and LSTM algorithm. There are LR_LSTM, RF_LSTM, SVR_LSTM, KNN_LSTM and ANN_LSTM algorithms are present in the analysis. Among these ANN_LSTM performs best among other algorithms. The ANN_LSTM algorithm has lowest MAE, MSE, RMSE which indicates the model accurately predicts the AQI. The model also has highest R^2 value of 56% which makes the model more suitable for prediction.

Compared to LR and KNN_LSTM models have lowest value of R2 value which indicates that it provides poor performance while prediction. The RF_LSTM model is also performing well but compared with ANN_LSTM, slightly lower R^2 value. Based on the results of this study the ANN_LSTM model is best for predicting air quality index in accurate manner [24].

To analyse the effectiveness of the proposed system, the results are compared with the existing system. Since the existing studies are used different type of datasets, places, pollutants and evaluation techniques their results cannot be compared with ours. To make a reasonable comparison, we applied the techniques used in the earlier studies to our Beijing air quality dataset. By testing all models, using the same conditions and evaluation metrics, we obtained results which are presented in Figure 4.

Figure 4 clearly indicates that our proposed hybrid model performs better than the existing methods in terms of MAE, MSE, RMSE and R^2 [20]. While Analysing Figure 4 it has lowest error rates of 19.26 of MAE, 620.67 of MSE, 24.92 of RMSE and highest R^2 value of 0.56 compared with existing systems. These results prove that our model is accurate and demonstrates the novelty and strength of our approach.

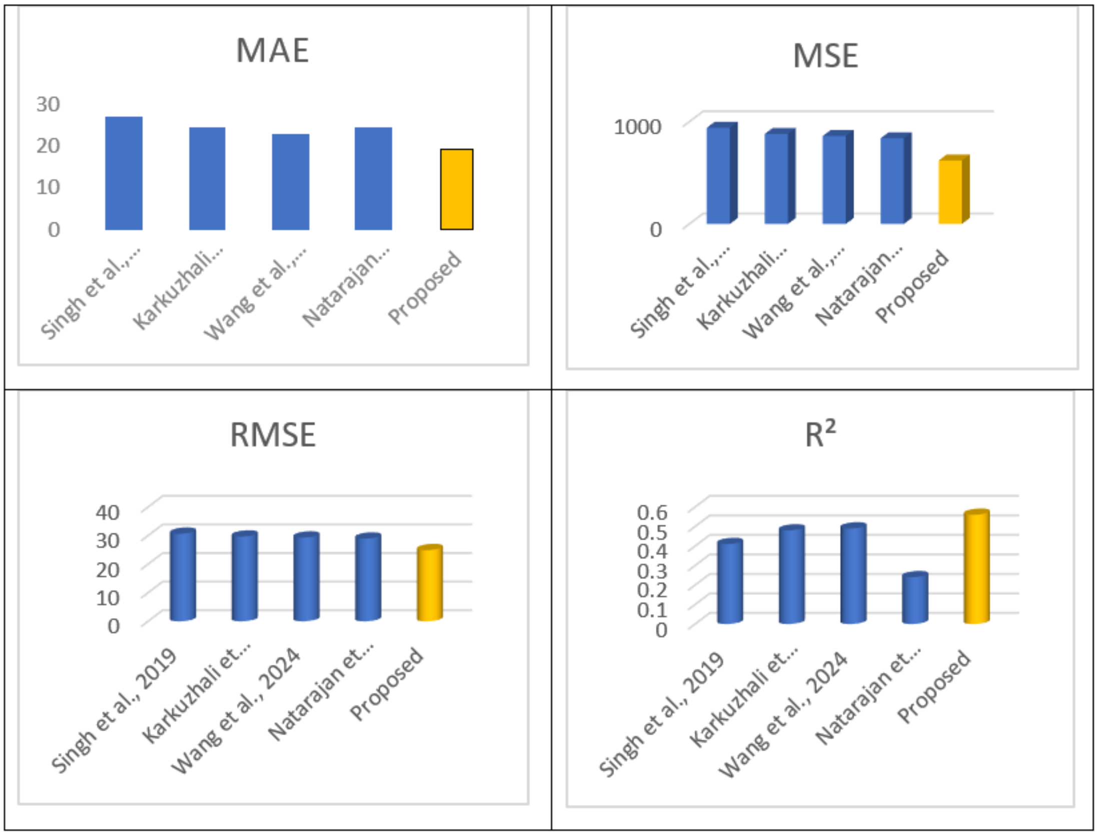

Fig 4. Comparative Results of Proposed vs Existing System

4 Conclusion

This study has developed an accurate and reliable air quality prediction model by focusing the strengths of ML and DL models. The proposed a hybrid framework was evaluated for its performance using the Beijing air quality dataset. The model also compared with several existing datasets and was subjected to the same evaluating techniques. The results clearly indicate the novelty and strength of this work. The proposed hybrid model achieves MAE of 19.26, MSE of 620.67, RMSE of 24.92 and R^2 of 0.56 which are better than the existing models tested. When compared with existing models, the proposed model continued to show higher accuracy and less error rate. This proves the efficiency of the proposed system. Hybrid Integration helps to new contribution to predicting the air quality index as previous research studies generally relied on only one type of model. The strength of this research include:

1. A hybrid integration of ML and DL models that improves the air quality prediction by combining complimentary learning abilities
2. A reliable comparison framework that measures the existing methods using the same datasets under same conditions.
3. Use of real index values of air quality which makes the results more meaningful for environmental decision making
4. Strong generalization among pollutants due to multi feature Beijing dataset.

However, this study also certain limitations. The model was trained on data from one region. In future this study implemented with multiple regions and other climatic conditions. Additionally, AI features and advanced architectures can be integrated.

5 Acknowledgement

The authors thank, DST-FIST, Government of India for funding towards infrastructure facilities at St.Joseph's College (Autonomous), Tiruchirappalli- 620 002.

References

1) Maltare NN, Vahora S. Air Quality Index prediction using machine learning for Ahmedabad city. *Digital Chemical Engineering*. 2023;7:100093. 10.1016/j.dche.2023.100093.
2) Wang Z, de Lima RP, Crooks JL, et al. Increasing the Spatial Coverage of Atmospheric Aerosol Depth Measurements Using Random Forest and Mean Filters. *arXiv*. 2023. 2306.07242.
3) Searle R. Investigation into Machine Learning and Emotional and Engagement Tracking Tools to Support and Enable At-Home Immersive Virtual Therapies. University of Kent (United Kingdom). 2024. 10.22024/UniKent/01.02.108791.
4) Kalantari E, Gholami H, Malakooti H, Nafarzadegan AR, Moosavi V. Machine learning for air quality index (AQI) forecasting: shallow learning or deep learning? *Environmental Science and Pollution Research International*. 2024;31(54):62962–62982. 10.1007/s11356-024-35404-1.
5) Singh P, Narasimhan TL, Lakshminarayanan CS. DeepAir: Air Quality Prediction using Deep Neural Network. Kochi, India. 2019. 10.1109/TENCON.2019.8929470.
6) Karkuzhali, Puyalnithi T, Nirmalan. Air quality prediction using stacked bi-long short-term memory and convolutional neural network in India. *Environmental Protection*. 2024;50(4):9–21. 10.24425/aep.2024.152891.
7) Tripathi SL, Kumar OP, Stalin AD, Ali T. .
8) Patel K, Oza AD, Sharma A, Mantri A, Kishore H. .
9) Ryalls JMW, Bishop J, Mofikoya AO, Bromfield LM, Nakagawa S, Girling RD. Air pollution disproportionately impairs beneficial invertebrates: a meta-analysis. *Nature Communications*. 2024;15(1):5447. 10.1038/s41467-024-49729-5.
10) Belsky J, Joshi NK. Impact of Biotic and Abiotic Stressors on Managed and Feral Bees. *Insects*. 2019;10(8):233. 10.3390/insects10080233.
11) Özüpak Y, Alpsalaz F, Aslan E. Air Quality Forecasting Using Machine Learning: Comparative Analysis and Ensemble Strategies for Enhanced Prediction. *Water, Air, & Soil Pollution*. 2025;236:464. 10.1007/s11270-025-08122-8.
12) Malhotra M, Walia S, Lin CC, et al. A systematic scrutiny of artificial intelligence-based air pollution prediction techniques, challenges, and viable solutions. *Journal of Big Data*. 2024;11:142. 10.1186/s40537-024-01002-8.
13) Natarajan SK, Shanmurthy P, Arockiam D, et al. Optimized machine learning model for air quality index prediction in major cities in India. *Scientific Reports*. 2024;14:6795. 10.1038/s41598-024-54807-1.
14) Abbasi MT, Alesheikh AA, Lotfata A, et al. Hybrid graph convolutional networks for air quality prediction: a systematic review of foundations, challenges, and opportunities. *International Journal of Environmental Science and Technology*. 2025;22:13327–13362. 10.1007/s13762-025-06599-8.
15) Seraphim BI, Nagarajan E, Pathak S. Predicting Air Quality: Ensemble Learning Approach. Hassan, India. 2024. 10.1109/ICDSIS61070.2024.10594204.
16) Pan Z, Chen H, Chen W, Cai F, Liu X. Time-Aware Graph Learning for Link Prediction on Temporal Networks. *IEEE Transactions on Neural Networks and Learning Systems*. 2025;36(6):10619–10634. 10.1109/TNNLS.2025.3545021.
17) Lee JY, Han SH, Kang JG, Lee CY, Lee JB, Kim HS, et al. Comparison of Models for Missing Data Imputation in PM-2.5 Measurement Data. *Atmosphere*. 2025;16(4):438. 10.3390/atmos16040438.
18) Wu H, Yamal JM, Yaseen A, Maroufy V. .
19) Ayus I, Natarajan N, Gupta D. Comparison of machine learning and deep learning techniques for the prediction of air pollution: a case study from China. *Asian Journal of Atmospheric Environment*. 2023;17:4. 10.1007/s44273-023-00005-w.
20) Jafari S, Byun YC. AI-driven state of power prediction in battery systems: A PSO-optimized deep learning approach with XAI. *Energy*. 2025;331:136764. 10.1016/j.energy.2025.136764.
21) Wu Q, Hong S, Yang L, Mu H, Huang C, Niu X, et al. Coupling coordination relationships between air pollutant concentrations and emissions in China. *Atmospheric Pollution Research*. 2023;14(3):101678. 10.1016/j.apr.2023.101678.
22) Dagur A, Agarwal S, Shukla DK, Ali S, Sharma S. .
23) Hooman R, Samer A, Robertson S, Tran NK, Hu B. Common statistical concepts in the supervised Machine Learning arena. *Frontiers in Oncology*. 2023;13. 10.3389/fonc.2023.1130229.
24) Mohapatra E, Das M, Das CP, Rath S. Predictive environmental modelling for air quality assessment via machine learning techniques. *Toxicological & Environmental Chemistry*. 2025;107(7):1509–1546. 10.1080/02772248.2025.2532741.