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

Predicting Air Pollution Levels Using LSTM Networks And Particle Swarm Optimization

Hardik Umed Bhai Patel[1] , Dr. Shivendra Kumar Jha[2] , Prof. Prashant Lalwani[3] , Prof. Sudhanshu Dixit[4] , Dr. Utkarsh Nigam[5] , Dr. Ankit.J. Patel[6]

[1]Assistant Professor, Civil Engineering, Government Engineering College Palanpur, Gujarat, 385001.hardikpatel.gec@gmail.com
[2]Assistant Professor, Civil Engineering, L. D. College of Engineering Ahmedabad, Gujarat, 380015. shivendra.jha@ldce.ac.in.
[3]Assistant Professor, Civil Engineering, L. D. College of Engineering Ahmedabad, Gujarat, 380015. prashant.lalwani86@gmail.com
[4]Assistant Professor, Civil Engineering, L. D. College of Engineering Ahmedabad, Gujarat, 380015.sudhanshudixit@ldce.ac.in
[5]Assistant Professor, Civil Engineering, L. D. College of Engineering Ahmedabad, Gujarat, 380015. utkarsh.nigam99@gmail.com
[6]Assistant Professor, Applied mechanics, Government Engineering College Modasa, Gujarat, 383315. er.ankitjpatel@gmail.com

ABSTRACT:
Air pollution is a major contributor to both public health issues and climate change, representing one of the most pressing challenges faced by humanity. Consequently, accurate forecasting of air pollution has become increasingly important. In this study, we propose a predictive model that integrates the Particle Swarm Optimization (PSO) algorithm with a Long Short-Term Memory (LSTM) deep learning framework. The model is designed to optimize the hyperparameters of the LSTM network and forecast the Particulate matter 2.5-micron concentration for the following day using historical Particulate matter 2.5-micron data. The optimization-enhanced model demonstrates superior performance, yielding a lower Root Mean Square Error (RMSE) compared to traditional machine learning approaches. Notably, the proposed model achieves an RMSE of 2.42 in Particulate matter 2.5-micron prediction.

Key words: *LSTM, Particulate matter, RSME.*

1.0 INTRODUCTION:

Air pollution refers to the presence of harmful substances or pollutants in the Earth's atmosphere, leading to the degradation of air quality. These pollutants—comprising gases, particulate matter, chemicals, or biological agents—accumulate beyond natural concentrations and pose significant threats to human health, ecosystems, and the environment. Air is essential for all living organisms on Earth. However, despite considerable advancements over the past five decades, pollution levels have continued to rise. This increase is largely driven by urbanization, industrialization, vehicular emissions, power generation, chemical activities, and certain natural phenomena such as agricultural burning, volcanic eruptions, and wildfires.

Air pollution is one of the most pressing environmental issues globally, contributing to respiratory diseases, environmental degradation, and climate change. PM2.5, fine particulate matter with a diameter less than 2.5 micrometers, is especially harmful due to its ability to penetrate deep into the lungs. Predicting PM2.5 concentrations is essential for health alerts.

The Long Short-Term Memory (LSTM) network is a type of Recurrent Neural Network (RNN) designed to effectively capture temporal dependencies in sequential data. Unlike standard RNNs, which utilize feedback loops to retain information from recent inputs, LSTM networks incorporate specialized memory cells that enable the retention of information over longer time intervals. These memory cells include

internal mechanisms called gates—specifically, the input gate, forget gate, and output gate—which regulate the flow of information by selectively updating, retaining, or discarding data based on its relevance.

A standard LSTM architecture consists of three primary layers. The first is the input layer, which receives the sequence data. This is followed by the recurrent layer, which contains the LSTM cells responsible for processing and preserving temporal information. Finally, the output layer generates the model's predictions. The interactions among the gates within each LSTM cell play a critical role in controlling how past information influences the current output and future learning.

Particle Swarm Optimization (PSO) is a population-based, stochastic optimization technique inspired by the social behaviour of birds flocking or fish schooling. It was introduced by Kennedy and Eberhart in 1995. In PSO, each solution to the optimization problem is represented as a "particle" within a swarm. These particles explore the search space by adjusting their positions and velocities based on their own experience and the experiences of neighbouring particles.

Traditional statistical models often struggle with the non-linear and temporal nature of air pollution data. Deep learning techniques, particularly LSTM networks, are capable of modelling long-term dependencies in time series data. However, selecting appropriate hyperparameters for LSTM models can be challenging and significantly impacts model performance.

2.0 LITERATURE REVIEW

Numerous studies have explored the use of machine learning and deep learning for air quality prediction. Approaches like Support Vector Machines (SVM), Random Forests (RF), and Artificial Neural Networks (ANN) have shown promise. More recently, LSTM networks have gained attention for their ability to capture temporal patterns. Optimization algorithms such as Genetic Algorithms (GA) and PSO have been used to enhance model accuracy.

An ensembled method for air quality monitoring and control using machine learning. The study presented by author S John Livingston (2023) focuses on air quality monitoring and control by leveraging machine learning algorithms such as Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM). It aims to forecast air quality levels by analyzing various environmental parameters, including NO_2, CO, O_3, PM2.5, PM10, SO_2, temperature (TEMP), pressure (PRES), dew point (DEWP), rainfall (RAIN), wind direction (WD), and wind speed (WSPM). Highlighting the significance of understanding pollutant sources, their effects, and concentration levels, the research aspires to improve pollution control strategies. Using eight selected parameters from the Beijing air quality dataset, the proposed model predicts the Air Quality Index (AQI) and is evaluated with regional data. The integration of machine learning techniques in air quality forecasting is emphasized as a promising and efficient solution for tackling environmental challenges.

Air Quality Index prediction using machine learning for Ahmedabad city. The study presented by author Nilesh N. Maltare (2023) The study presented in [2] compares various machine learning techniques—including SARIMA, Support Vector Machine (SVM), and Long Short-Term Memory (LSTM)—for predicting the Air Quality Index (AQI) in Ahmedabad city. To ensure optimal model performance, several data preprocessing techniques are applied prior to feeding the data into the machine learning models. Emphasizing the Support Vector Machine algorithm with a Radial Basis Function (RBF) kernel, the research demonstrates that this approach yields comparatively superior results for AQI prediction in the region. Utilizing data sourced from the Central Pollution Control Board (CPCB) of India, the study analyzes and forecasts air quality trends. This work contributes to the domain of digital chemical engineering by applying advanced machine learning methods to environmental monitoring and air quality prediction.

3.0 METHODOLOGY

Air pollution prediction using LSTM deep learning and particle swarm optimization algorithm

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 3s, 2025
https://www.theaspd.com/ijes.php
The Existing System LSTM deep learning and particle swarm optimization algorithm.

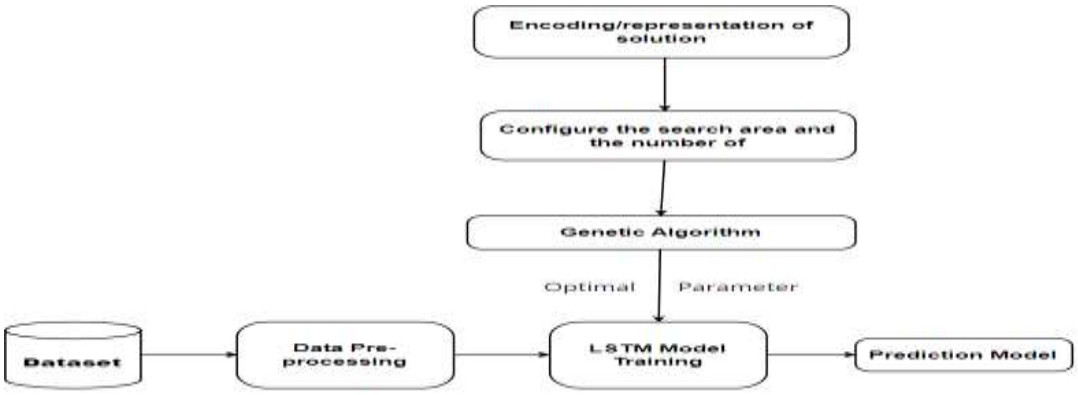


Fig. 1.1 The Existing GA-LSTM System

- The Existing methodology combines the Genetic Algorithm (GA) with the Long Short-Term Memory (LSTM) deep learning model to forecast air pollution.
- GA is employed to optimize crucial LSTM hyperparameters, including window size and unit count.
- This model forecasts pollution levels for the subsequent day, focusing on four pollutant types: PM10, PM2.5, CO, and NOX.

PSO typically exhibits higher computational efficiency, demanding less computational resources and time for execution compared to GAs in specific problem areas. Its streamlined nature and fewer parameters often lead to quicker convergence and decreased computational load.

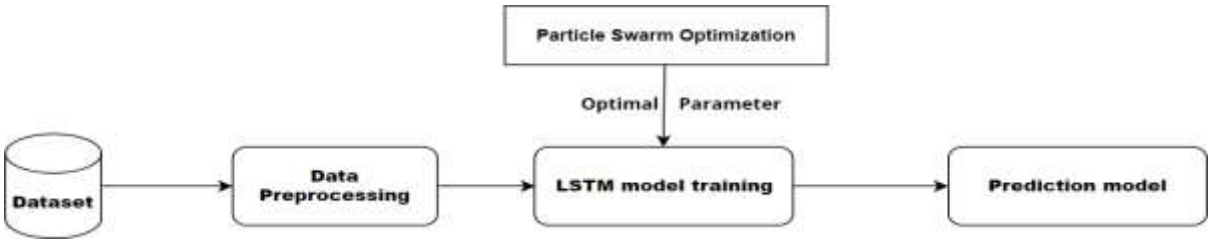


Fig. 1.2. The Proposed System

4.0 RESULTS AND DISCUSSIONS

Dataset

As we are comparing our result with existing work, we are using the same dataset as the existing work has used.

- URL: https://www.kaggle.com/datasets/fedesoriano/air-quality-data-in-india/
- Dataset contains air quality data of India (2017-2022).
- It contains different pollutant data which are PM2.5, PM10, CO, NOx, NO, NO2, NH3, SO2, O3, Benzene, Toluene, Xylene.
- We used PM2.5 time series data.

The dataset contains the 25787 records of air pollution data from year 2017 to 2022.Here we use this dataset to train the model. The training set consists of 85% of records, while the test set consists of 15% records. Here we can see the image of the dataset in figure 4.1.

City	Datetime	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI	AQI_Bucket
Ahmedab	01-01-2015 01:00			1	40.01	36.37		1	122.07		0	0	0		
Ahmedab	01-01-2015 02:00			0.02	27.75	19.73		0.02	85.9		0	0	0		
Ahmedab	01-01-2015 03:00			0.08	19.32	11.08		0.08	52.83		0	0	0		
Ahmedab	01-01-2015 04:00			0.3	16.45	9.2		0.3	39.53	153.58	0	0	0		
Ahmedab	01-01-2015 05:00			0.12	14.9	7.85		0.12	32.63		0	0	0		
Ahmedab	01-01-2015 06:00			0.33	15.95	10.82		0.33	29.87	64.25	0	0	0		
Ahmedab	01-01-2015 07:00			0.45	15.94	12.47		0.45	27.41	191.96	0	0	0		
Ahmedab	01-01-2015 08:00			1.03	16.66	16.48		1.03	20.92	177.21	0	0	0		
Ahmedab	01-01-2015 09:00			1.47	16.25	18.02		1.47	16.45	122.08	0	0	0		
Ahmedab	01-01-2015 10:00			2.05	13.78	16.08		2.05	15.14		0	0	0		
Ahmedab	01-01-2015 11:00			2.27	13.87	16.73		2.27	14.12	99.17	0	0	0		
Ahmedab	01-01-2015 12:00			1.73	12.87	14.63		1.73	13.26	91.67	0	0	0		
Ahmedab	01-01-2015 13:00			1.72	14.15	15.55		1.72	17.2	95.92	0	0	0		
Ahmedab	01-01-2015 14:00			1.85	15.74	17.62		1.85	18.78		0	0	0		
Ahmedab	01-01-2015 15:00			0.95	15.94	16.18		0.95	19.16		0	0	0		
Ahmedab	01-01-2015 16:00			0.87	17.28	18.32		0.87	17.83		0	0	0		
Ahmedab	01-01-2015 17:00			0.8	19.04	20		0.8	16.14	187.62	0	0	0		
Ahmedab	01-01-2015 18:00			0.47	21.24	22.7		0.47	11.93		0	0	0		
Ahmedab	01-01-2015 19:00			0.53	25.63	27.42		0.53	14.99		0	0.33	0		
Ahmedab	01-01-2015 20:00			0.47	16.22	16		0.47	13.66	187.42	0	0.23	0		

Fig. 1.3. The Dataset Image

As an initial step, the data from each monitoring station was separated into individual CSV files named after their respective stations. Unnecessary features, such as SO_2, NO, O_3, AQI, and AQI-Bucket, were removed using the Drop function. To handle missing values, the fill method was applied to replace all missing entries with zeros.

Next, the MinMaxScaler was employed to normalize the dataset's feature values. This scaling technique transforms the values to a standardized range between 0 and 1, preserving the relative differences between data points while ensuring consistent scaling across all features for model training.

Implementation Details:

Software and Hardware Used for Implementation:

1. Google Colab
2. Python libraries.

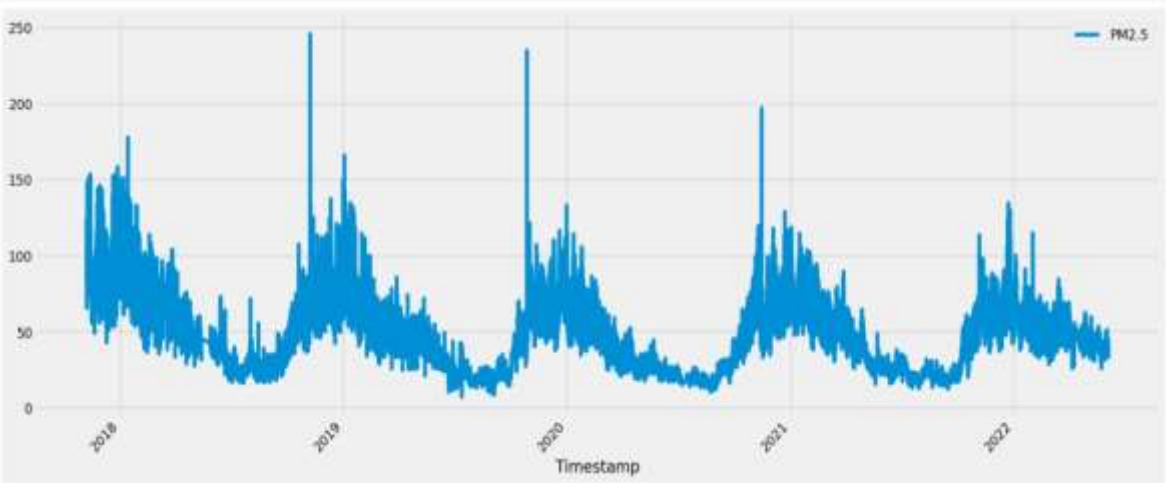
The implementation was carried out using Google Colab, a cloud-based platform that offers an interactive Python environment with built-in support for a wide range of machine learning libraries. Google Colab also provides significant computational resources, making it well-suited for training and evaluating machine learning algorithms efficiently.

Results:

Initially, the air pollution data was visualized to understand underlying patterns and trends. Following this, data preprocessing steps were applied to prepare the dataset for analysis. The dataset includes air pollution trends recorded on a daily, monthly, and hourly basis, covering the period from 2018 to 2020

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

Fig. 1.4 : PM2.5 trend with the year



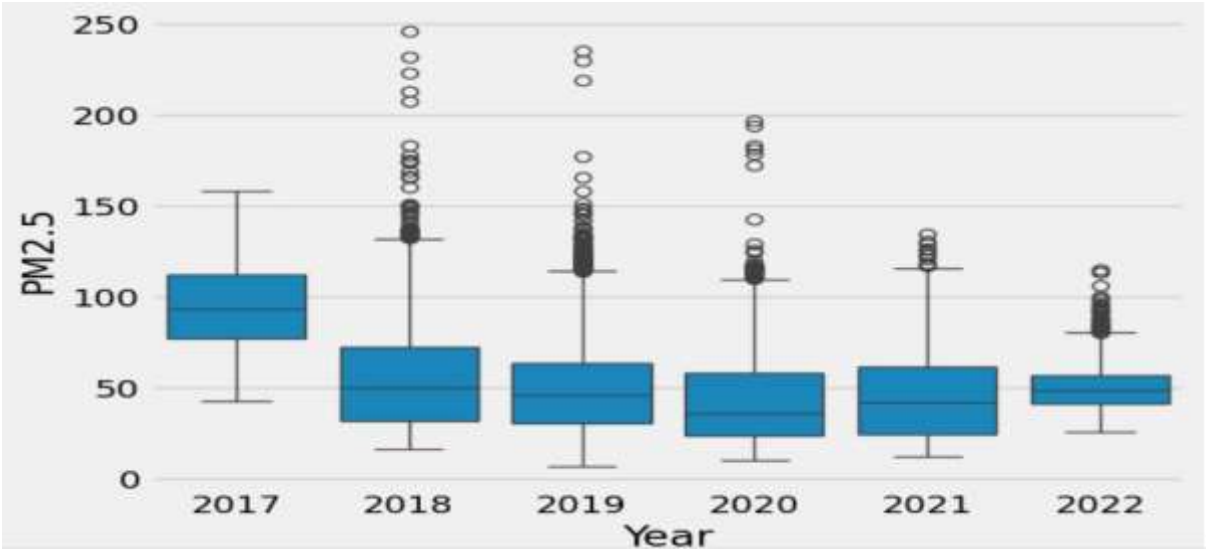


Fig. 1.5 : PM2.5 trend with the Month

The data undergoes preprocessing before being utilized in our model. During this step, we calculate the mean of the PM2.5 value for the specific day. After the data preprocessing step, our data resembles the figure below.

To prepare the data for modelling, Minmax scaling was applied—a widely used normalization technique in machine learning. This method scales numerical features to a uniform range between 0 and 1. By subtracting the minimum value of each feature and dividing by its range, this approach ensures a proportional transformation while preserving the original distribution of the data.

This normalization is particularly beneficial when dealing with features that have varying scales, as it prevents any single feature from disproportionately influencing the learning process.

Fig. 1.6. Data after MinMax Scaling

```
array([[0.80924161],
       [0.91297267],
       [0.94951492],
       ...,
       [0.27497376],
       [0.24870153],
       [0.24014526]])
```

The dataset was divided into training and testing sets, with 75% of the data used for training and the remaining 25% reserved for testing. This pre-processed data was then fed into the model. Upon completion of the training phase, the model delivered highly accurate results, which were evaluated using two key performance metrics: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

RMSE and MAE are critical for evaluating the predictive accuracy of the model. They measure the deviation between the predicted values and the actual observed values in the dataset. Lower values of RMSE and MAE indicate a closer alignment between the predicted and actual outcomes, reflecting the model's precision and reliability. Analyzing these metrics offers meaningful insights into the model's effectiveness in identifying and capturing the underlying trends and patterns within the data.

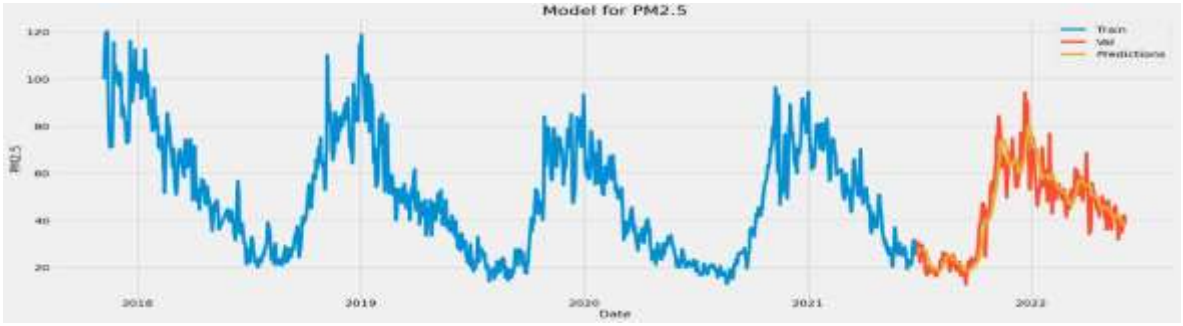


Fig. 1.7. Model performance

	Proposed Model (PSO LSTM)
RMSE	2.42
MAE	8.50

The table below illustrates the predictions made by the proposed model for the PM2.5 value.

Date	PM2.5	Predictions
2021-06-28	28.667917	25.768383
2021-06-29	30.903750	26.204222
2021-06-30	29.017917	26.725386
2021-07-01	27.166667	27.201784
2021-07-02	27.132917	27.648120
...	...	...
2022-05-31	41.465417	38.162060
2022-06-01	41.645833	37.987030
2022-06-02	42.255417	38.018085
2022-06-03	39.425417	38.224083
2022-06-04	38.503750	38.411067

323 rows × 2 columns

Table 1.1. Predicted Value for the PM2.5

CONCLUSION:

In this research, we propose a novel approach for air pollution prediction by integrating Particle Swarm Optimization (PSO) with a Long Short-Term Memory (LSTM) neural network. This hybrid model, referred to as PSO-LSTM, demonstrated outstanding performance, achieving a Root Mean Square Error (RMSE) of 2.42—significantly outperforming the Genetic Algorithm (GA)-based LSTM model.

In conclusion, the PSO-LSTM algorithm has shown superior predictive accuracy compared to the GA-LSTM model, particularly in forecasting PM2.5 concentrations. By combining PSO with the LSTM deep learning framework, the model benefits from optimized hyperparameter tuning, leading to more accurate and dependable predictions. PSO's efficient exploration of the parameter space enables the identification of optimal configurations, enhancing the overall performance of the LSTM model.

This improved accuracy underscores the potential of PSO as a powerful optimization technique for enhancing LSTM-based models in the domain of air quality forecasting. The findings emphasize the value of the PSO-LSTM model as a robust tool for environmental monitoring and air pollution control, offering improved predictive capabilities to support informed decision-making and mitigate the harmful effects of air pollution on public health and the environment.

REFERENCES:

1.) An ensembled method for air quality monitoring and control using machine learning [Published on Elsevier in Measurement: Sensors journal, 2023] https://doi.org/10.1016/j.measen.2023.100914
2.) Air Quality Index prediction using machine learning for Ahmedabad city [Published on Elsevier in Digital Chemical Engineering 2023] https://doi.org/10.1016/j.dche.2023.100093
3.) Fuzzy Inference System Tree with Particle Swarm Optimization and Genetic Algorithm: A novel approach for PM10 forecasting. [Published on Elsevier in Expert Systems With Applications , 2021] https://doi.org/10.1016/j.eswa.2021.115376
4.) Forecasting PM 2.5 concentration based on integrating of CEEMDAN decomposition method with SVM and LSTM. [Published on Elsevier in Ecotoxicology and Environmental Safety, 2023] https://doi.org/10.1016/j.ecoenv.2023.115572
5.) Air pollution prediction using LSTM deep learning and metaheuristics algorithms. [Published on Elsevier in Measurement: Sensors journal, 2022] https://doi.org/10.1016/j.measen.2022.100546
6.) A model for particulate matter (PM2.5) prediction for Delhi based on machine learning approaches. [Published on Elsevier in Procedia Computer Science 167 (2020) 2101–2110.] https://doi.org/10.1016/j.envsoft.2022.105529
7.) Methods used for handling and quantifying model uncertainty of artificial neural network models for air pollution forecasting [Published on Elsevier in Environmental Modelling and Software, 2022.] https://doi.org/10.1016/j.envsoft.2022.105529
8.) Prediction of Air Pollutants using supervised Machine Learning. [Proceedings of the Fifth International Conference on Intelligent Computing and Control Systems (ICICCS 2021)] https://doi.org/10.1016/j.eswa.2021.3212176