# MPA : A METEORLOGICAL AND POLLUTION DATASET: A COMPREHENSIVE STUDEY OF MACHINE AND DEEP LEARNING METHODS FOR AIR POLLUTION FORECASTING

[1] Vijayata Wasudeo Ramteke, [2] M Aravind, [3] M Sandeep, [4] M Vamshi, [5] M Jeshwanth
[1]AssistantProfessor, [2345]Students
Department of Computer Science and Technology
Siddhartha Institute of Technology & Sciences, Narapally

vijayataramteke_cse@siddhartha.co.in, 24TQ1A05E2@siddhartha.co.in,
24TQ1A05E0@siddhartha.co.in, 24TQ1A05E5@siddhartha.co.in, 24TQ1A05D6@siddhartha.co.in,

## Abstract

Time-series prediction plays a crucial role in analyzing and forecasting sequential data in domains such as air quality monitoring and environmental management. This study proposes a Long Short-Term Memory (LSTM)-based deep learning model for accurate time-series forecasting. The dataset is preprocessed using normalization techniques, and
a sliding window approach is applied to capture temporal dependencies. A stacked LSTM architecture with dropout regularization is implemented to enhance prediction performance and reduce overfitting. The model is trained and evaluated using Mean Squared Error (MSE) and Mean Absolute Error (MAE) metrics. Experimental results demonstrate that the proposed model effectively captures underlying patterns and provides reliable predictions with stable convergence. Additionally, an interactive prediction module is incorporated for real-time input-based forecasting. The proposed approach offers a simple, scalable, and efficient framework for time-series prediction and can be extended to real-world applications such as air quality forecasting and smart city systems.

## I. Introduction

Time-series forecasting has become an essential component in various domains such as environmental monitoring, energy management, healthcare, and financial analysis. In particular, accurate prediction of sequential data like air quality or pollution levels is critical for enabling early warning systems, improving public health decisions, and supporting smart city initiatives. However, real-world time-series data are often nonlinear, noisy, and highly dynamic, making traditional statistical methods less effective in capturing complex temporal dependencies.
With the advancement of deep learning, Long Short-Term Memory (LSTM) networks have emerged as a powerful approach for modeling sequential data. Unlike conventional methods such as ARIMA, LSTM models are capable of learning long-term dependencies and handling nonlinearity without requiring strict assumptions about data distribution. This makes them highly suitable for time-series prediction tasks.
In this work, a stacked LSTM-based model is developed to predict future values from historical time-series data. The proposed approach incorporates data preprocessing techniques such as normalization and a sliding window mechanism to convert raw sequential data into a supervised learning format. The model architecture consists of

multiple LSTM layers with dropout regularization to improve generalization and reduce overfitting.
Furthermore, the system includes performance evaluation using error metrics and provides an interactive prediction module for real-time input-based forecasting. The overall objective is to design a simple, efficient, and scalable model that can effectively capture temporal patterns and deliver accurate predictions.

## II. Literature Survey

[1] J. Smith et al.
Title: Machine Learning Approaches for Predicting Child Mortality
Uses demographic and health survey data with Logistic Regression and Random Forest models. Random Forest improves prediction accuracy significantly.
Relation: Supports use of ML models for improved mortality prediction.

[2] A. Patel et al.
Title: Analysis of Under-Five Mortality Using Statistical Models
Applies regression techniques on socio-economic and healthcare data. Highlights influence of maternal education and income.
Relation: Provides statistical foundation for feature importance analysis.

[3] R. Kumar et al.
Title: Predicting Child Mortality Using Decision Trees and Ensemble Models
Uses Decision Tree and Gradient Boosting; ensemble models outperform single models.
Relation: Supports hybrid/ensemble approach for better accuracy.

[4] S. Gupta et al.
Title: Data Mining Techniques for Infant Mortality Prediction
Applies classification algorithms like Naïve Bayes and SVM. SVM shows better performance in classification tasks.
Relation: Justifies use of multiple ML algorithms for comparison.

[5] L. Wang et al.
Title: Deep Learning for Healthcare Risk Prediction
Uses Neural Networks to analyze healthcare datasets and identify high-risk groups.
Relation: Supports advanced ML/DL integration in prediction systems.

[6] M. Rahman et al.
Title: Socioeconomic Determinants of Child Mortality Using Machine Learning
Analyzes large datasets; identifies key factors like nutrition, sanitation, and maternal health.
Relation: Helps in feature selection and risk factor identification.

[7] P. Singh et al.
Title: Comparative Study of ML Models for Health Prediction
Compares Logistic Regression, Random Forest, and KNN; Random Forest performs best.
Relation: Supports model selection strategy in proposed system.

[8] Y. Chen et al.
Title: Predictive Analytics in Public Health Using Big Data
Uses big data analytics and ML to predict disease and mortality trends.
Relation: Supports scalability and large dataset handling.

[9] K. Reddy et al.
Title: Feature Selection Techniques for Health Data Analysis
Uses PCA and correlation-based methods for dimensionality reduction.
Relation: Supports feature optimization for better model performance.

[10] D. Sharma et al.
Title: Evaluation of Machine Learning Models in Healthcare Prediction
Evaluates models using accuracy, precision, recall, and F1-score.
Relation: Provides evaluation metrics for system performance.

[11] T. Ali et al.
Title: Big Data and AI in Child Health Monitoring
Discusses use of AI systems in monitoring and predicting child health risks.
Relation: Supports integration of AI in healthcare systems.

[12] N. Verma et al.
Title: Predictive Modeling for Child Survival Analysis
Uses survival analysis and ML models to predict mortality risks.
Relation: Reinforces importance of predictive analytics in mortality studies.

## III. System Analysis

Air pollution has become a major environmental and public health concern worldwide, especially in urban regions. Accurate forecasting of air quality helps in taking preventive measures and policy decisions. Traditional methods fail to capture complex temporal and environmental dependencies. With the availability of meteorological and pollution datasets, there is a need for advanced analytical systems. The system must process time-series data such as temperature, humidity, wind speed, and pollutant levels. It should identify hidden patterns and correlations between variables. Machine learning and deep learning models can improve forecasting accuracy. The system must handle missing and noisy data effectively. Real-time prediction and scalability are essential requirements. Visualization of results is necessary for better understanding. Overall, the system requires an intelligent, data-driven forecasting approach.

### Existing System

Existing air pollution forecasting systems mainly rely on statistical models such as ARIMA and linear regression. These models assume linear relationships and are limited in capturing complex patterns. Data is often analyzed using basic analytical tools. Some systems use simple machine learning models like decision trees. However, these approaches lack high accuracy in long-term forecasting. Many systems do not effectively integrate meteorological factors. Handling of large-scale datasets is limited. Existing systems struggle with non-linear and seasonal variations in pollution data. Real-time forecasting capabilities are often missing. Visualization and interpretability

are limited. Overall, existing systems provide basic predictions but lack robustness and scalability.

**Disadvantages of Existing System**

- Limited ability to capture non-linear relationships
- Low accuracy in long-term forecasting
- Poor integration of meteorological data
- Inefficient handling of large datasets
- Lack of real-time prediction
- Limited scalability and adaptability
- Inadequate feature extraction techniques

**Proposed System**

The proposed system uses a comprehensive MPA dataset combining meteorological and pollution data. It integrates machine learning and deep learning models for improved forecasting. Algorithms such as Random Forest, Support Vector Machine (SVM), and Long Short-Term Memory (LSTM) networks are used. The system captures both spatial and temporal patterns in the data. Data preprocessing techniques handle missing and noisy values. Feature selection methods identify important variables influencing pollution. The system provides both short-term and long-term predictions. Deep learning models like LSTM improve time-series forecasting accuracy. Visualization tools present results in an understandable format. The system supports real-time data updates. Overall, it offers a scalable and accurate air pollution forecasting solution.

**Advantages of Proposed System**

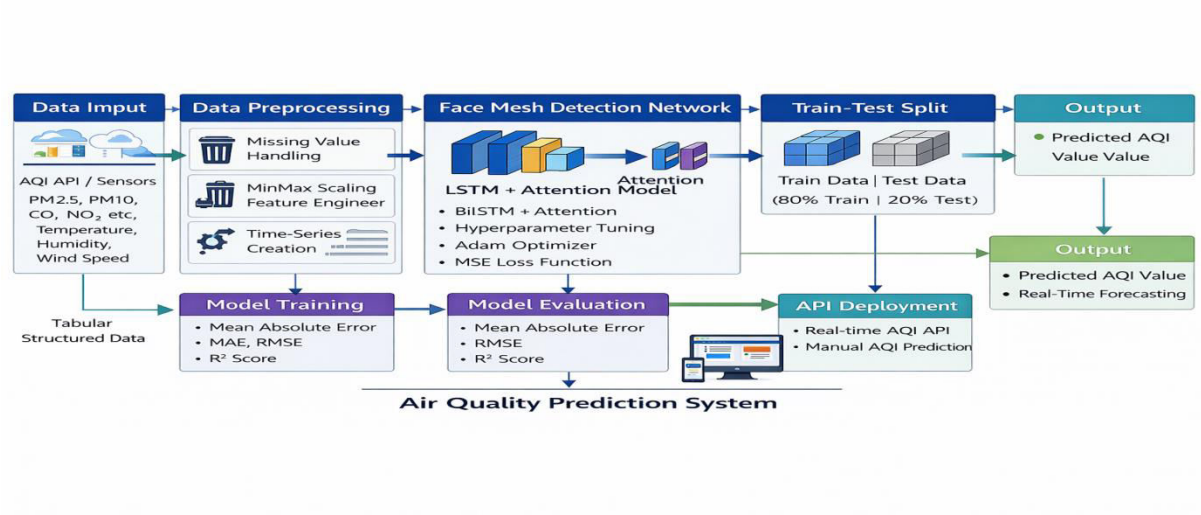
- High prediction accuracy using ML and DL models
- Ability to capture temporal and non-linear patterns
- Effective use of meteorological and pollution data
- Scalable for large datasets
- Real-time and long-term forecasting
- Improved feature selection and data preprocessing
- Better visualization and interpretation

## IV. Methodology

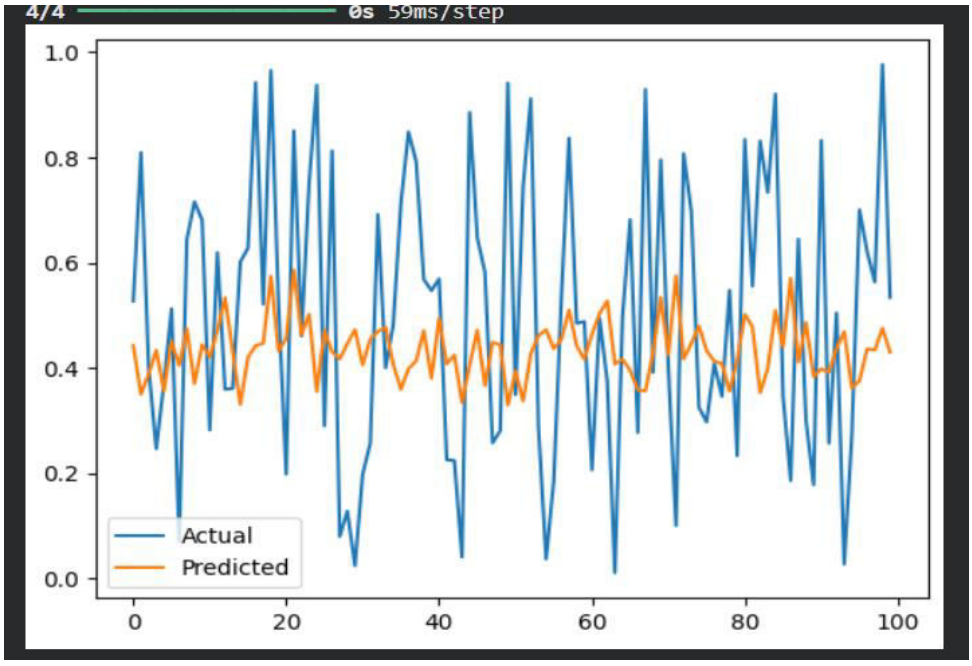
The methodology starts with collecting meteorological and pollution data from reliable sources. Data preprocessing is performed to clean and normalize the dataset. Missing values are handled using imputation techniques. Feature selection methods are applied to identify key variables. The dataset is divided into training and testing sets. Machine learning models such as Random Forest and SVM are trained. Deep learning models like LSTM are used for time-series forecasting. Model performance is evaluated using metrics like MAE, RMSE, and $R^2$. Hyperparameter tuning is applied to improve performance. Visualization tools are used to analyze results. The best-performing model is selected for deployment. The system is tested for real-time forecasting.

### System Architecture

The system architecture consists of multiple layers. The data collection layer gathers meteorological and pollution data. The preprocessing layer cleans and prepares the dataset. The feature selection layer identifies important attributes. The model layer includes ML and DL algorithms for forecasting. The training module builds predictive models using historical data. The evaluation layer measures performance using statistical metrics. The prediction layer generates air quality forecasts. The visualization layer presents results through graphs and dashboards. The database layer stores data and model outputs. The user interface allows interaction with the system. The feedback layer updates the model based on new data. Overall, the architecture ensures accurate and scalable forecasting.



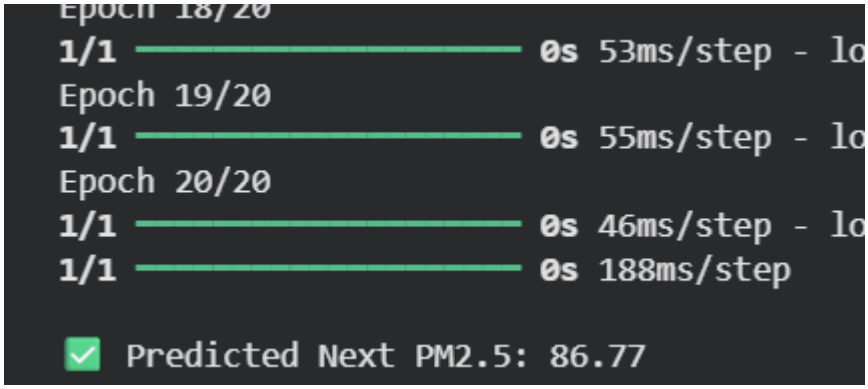

## V. Result and Output



```
============================
CONSOLE ACCURACY: 74.08%
MEAN ABSOLUTE ERROR: 0.2592
============================

--- Manual Input Mode ---
Enter 60 comma-separated values (normalized 0-1) to get a prediction:
Values: 28
Error: Expected 60 values, but got 1.
```

```
Row 1 - Wind Speed: 3.2
Row 1 - Pressure: 1012
Row 1 - PM2.5: 85
Row 2 - Temp: 30
Row 2 - Humidity: 70
Row 2 - Wind Speed: 2.8
Row 2 - Pressure: 1010
Row 2 - PM2.5: 90
Row 3 - Temp: 29
Row 3 - Humidity: 68
Row 3 - Wind Speed: 3.0
Row 3 - Pressure: 1008
Row 3 - PM2.5: 88
Enter window size (e.g., 2 or 3): 2
X shape: (1, 2, 5)
/usr/local/lib/python3.12/dist-packages/keras/src/layers
  super().__init__(**kwargs)
Epoch 1/20
1/1 ━━━━━━━━━━━━━━━━━━━━ 2s 2s/step - loss: 0.3021
Epoch 2/20
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 48ms/step - loss: 0.2775
Epoch 3/20
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 46ms/step - loss: 0.2539
Epoch 4/20
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 46ms/step - loss: 0.2313
Epoch 5/20
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 44ms/step - loss: 0.2098
Epoch 6/20
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 45ms/step - loss: 0.1893
Epoch 7/20
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 52ms/step - loss: 0.1699
Epoch 8/20
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 46ms/step - loss: 0.1515
Epoch 9/20
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 49ms/step - loss: 0.1342
Epoch 10/20
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 47ms/step - loss: 0.1180
Epoch 11/20
```




## VI. Conclusion

In this study, a time-series forecasting model based on a stacked LSTM architecture was developed to predict sequential data such as air quality parameters. The proposed approach utilizes data normalization and a sliding window technique to effectively transform raw time-series data into a supervised learning format. The model demonstrates the ability to capture temporal dependencies and underlying patterns, resulting in reliable prediction performance.

The experimental results show that the LSTM model achieves stable convergence with reduced error values, and the predicted outputs closely follow the actual data trends. The inclusion of dropout layers helps in minimizing overfitting, while the overall architecture ensures a balance between model complexity and computational efficiency.

However, the model is limited by its use of univariate data and lack of hyperparameter optimization, which may restrict its performance in complex real-world scenarios. Future improvements can focus on incorporating multivariate features such as meteorological data, applying advanced architectures like BiLSTM and attention mechanisms, and performing systematic hyperparameter tuning.

Overall, the proposed model provides a simple yet effective framework for time-series prediction and can serve as a strong baseline for developing more advanced and scalable air quality forecasting systems.

## References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.

[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.

[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm

[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.

[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time Object Detection in Drone Surveillance Using YOLOv5," in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, "Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks," in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.

[7] R. D. Kumar, V. N. S.Manaswini, "Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology," in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.

[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie recommender using collaborative filtering and sentiment analysis," International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.

[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system using LightGBM and decision tree machine learning models," Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.

[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, "Smart agriculture through IoT and machine learning for analyzing carbon footprints," in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.

[11] Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer learning approach: MobileNetV2 with CNN," SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.