International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 2(1) (2026)

ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

MPD A METEROGOLICAL AND POLLUTION DATASET A COMPREHENSIVE STUDY OF MACHINE AND DEEP LEARNING METHODS FOR AIR POLLUTION FORECASTING

[1] R Uma, [2] C Geetha, [3] E Deepika, [4] B Sumathi, [5] J Sailal
[1]AssistantProfessor, [2345]Students
Department of AIML
Siddhartha Institute of Technology & Sciences, Narapally
umakola_cse@siddhartha.co.in, 24tq1a6623@siddhartha.co.in, 24tq1a6632@siddhartha.co.in, 24tq1a6612@siddhartha.co.in, 24tq1a6645@siddhartha.co.in,

Abstract

This study presents a comprehensive framework for air quality prediction using a Meteorological and Pollution Dataset (MPD) by integrating both machine learning and deep learning approaches. The dataset consists of pollutant concentrations (PM2.5, PM10, NO, NO_2, CO, SO_2, O_3) along with meteorological parameters such as temperature, humidity, and wind speed collected from multiple Indian cities. Data preprocessing techniques including missing value imputation, normalization, and feature engineering are applied to enhance data quality. The proposed model employs the XGBoost algorithm for AQI classification and a Long Short-Term Memory (LSTM) network for time-series forecasting of AQI values. The hybrid approach leverages the strength of XGBoost in handling structured data and LSTM in capturing temporal dependencies. Experimental results demonstrate that the model achieves high prediction performance, with classification accuracy reaching approximately 92–95% and low forecasting error (MSE < 0.1). The findings indicate that integrating meteorological and pollution data significantly improves prediction accuracy and reliability, making the proposed framework suitable for real-time air quality monitoring and decision-making systems.

KEYWORDS
Air Quality Prediction, AQI, Machine Learning, Deep Learning, XGBoost, LSTM, Meteorological Data, Time-Series Forecasting

I. Introduction

Air pollution has become one of the most severe environmental challenges of the 21st century, affecting both developed and developing nations. Rapid industrialization, urban expansion, and increased vehicular emissions have led to a significant decline in air quality, particularly in densely populated regions such as India. Poor air quality has been directly linked to serious health issues, including respiratory diseases, cardiovascular disorders, and premature mortality. According to the World Health Organization, air pollution is responsible for millions of deaths annually, making it a critical area of research and policy intervention.

Air quality is typically quantified using the Air Quality Index (AQI), which aggregates the concentration levels of major pollutants such as particulate matter (PM2.5 and PM10), nitrogen oxides (NO and NO_2), sulfur dioxide (SO_2), carbon

monoxide (CO), ozone (O_3), and volatile organic compounds. These pollutants are influenced by various anthropogenic activities and environmental conditions. Meteorological parameters such as temperature, humidity, wind speed, and atmospheric pressure play a crucial role in pollutant dispersion and transformation. Therefore, integrating meteorological data with pollution data is essential for accurate air quality prediction.

Traditional air quality prediction methods have relied on statistical models such as linear regression and autoregressive integrated moving average (ARIMA). While these models provide a basic understanding of pollutant trends, they often fail to capture the complex nonlinear relationships between environmental factors and pollutant concentrations. To overcome these limitations, researchers have increasingly turned to machine learning and deep learning techniques, which are capable of handling large datasets and modeling intricate patterns.

II. Literature Survey

Capo et al. [1] (2023), "MPD: A Meteorological and Pollution Dataset. A Comprehensive Study of Machine and Deep Learning Methods for Air Pollution Forecasting", et al.

This study introduces the MPD dataset, which integrates air quality and meteorological variables collected from the historical records of the Community of Madrid. The dataset includes pollutants such as PM2.5, PM10, NO_2, and meteorological parameters like temperature, humidity, and wind speed. Various machine learning and deep learning models, including Random Forest and LSTM, are applied for air pollution forecasting. The models are evaluated using standard performance metrics and are shown to outperform existing state-of-the-art approaches. The results highlight that combining meteorological and pollution data significantly improves prediction accuracy. This is closely related to our work as it also focuses on air quality prediction using integrated environmental data and advanced machine learning techniques

Kumar et al. [2] (2022), "An Integrated Framework for Predicting Air Quality Index Using Pollutant Concentration and Meteorological Data", et al.

This study proposes a novel integrated framework for predicting the Air Quality Index (AQI) using both pollutant concentration and meteorological data. The framework consists of four modules: AQIfp for forecasting pollutant concentrations, AQIp for predicting AQI using pollutant and historical AQI data, AQIm for predicting AQI using meteorological features such as temperature, humidity, wind speed, and pressure, and AQIc which combines the outputs of AQIp and AQIm for improved prediction. Various machine learning and deep learning techniques including Artificial Neural Networks (ANN), Random Forest, XGBoost, LSTM, ARIMA, Support Vector Regression, and K-Nearest Neighbors are applied. The results show that ARIMA and ANN perform best for predicting pollutants, and the combined AQIc module achieves a low Mean Absolute Error of 7.09, indicating high prediction accuracy. The study concludes that integrating both pollutant and meteorological data significantly improves AQI prediction performance. This is closely related to our work as it also focuses on combining environmental and meteorological data with multiple machine learning models to enhance air quality prediction accuracy.

Sharma et al. [3] (2021), "Data Analysis and Preprocessing Techniques for Air Quality Prediction: A Survey", et al.

This study presents a comprehensive survey of data analysis and preprocessing

techniques used in air quality prediction systems. It reviews various datasets containing pollutant concentrations and meteorological variables such as temperature, humidity, and wind speed. The paper focuses on essential preprocessing steps including data cleaning, missing value handling, normalization, feature selection, and dimensionality reduction. It also discusses the impact of preprocessing on machine learning and deep learning models such as Random Forest, Support Vector Machines, and Neural Networks. The results indicate that proper preprocessing significantly improves model accuracy and reduces prediction errors. The study highlights that handling noisy and incomplete data is crucial for reliable air quality forecasting. This is closely related to our work as it emphasizes the importance of data preprocessing techniques, which are a key step in our project for improving the performance of machine learning models in air quality prediction.

Zhang et al. [4] (2023), "Advanced Air Quality Prediction Using Multimodal Data and Dynamic Modeling Techniques", et al.

This study utilizes multimodal data including pollutant concentrations, meteorological variables, and external factors for air quality prediction.

Dynamic modeling techniques such as LSTM and hybrid deep learning models are applied to capture temporal patterns.The results show improved prediction accuracy compared to traditional single-source models.The study highlights that integrating multiple data sources enhances model robustness and performance.This is related to our work as we also combine different environmental factors to improve air quality prediction accuracy.

Wang et al. [5] (2019), "Data Analysis and Mining of the Correlations Between Meteorological Conditions and Air Quality: A Case Study in Beijing", et al.

This study analyzes the relationship between meteorological factors and air quality using real-world data from Beijing.Various data mining techniques are applied to identify correlations between pollutants and weather parameters such as temperature, humidity, and wind speed.The results show strong dependencies between meteorological conditions and pollutant concentration levels.The study highlights that weather conditions significantly influence air pollution patterns.This is related to our work as it emphasizes the importance of meteorological data in improving air quality prediction models.

Yılmaz et al. [6] (2020), "Comprehensive Analysis of Air Pollution and the Influence of Meteorological Factors: A Case Study of Adıyaman Province", et al.

This study examines air pollution levels and their relationship with meteorological factors in Adıyaman province.Statistical and analytical methods are used to evaluate the impact of temperature, humidity, and wind on pollutant concentrations.

The results indicate that meteorological conditions significantly affect air quality variations.The study emphasizes the importance of incorporating weather parameters for accurate pollution analysis.This is related to our work as it also focuses on using meteorological data to enhance air quality prediction models.

Li et al. [7] (2022), "Multi-Scale Deep Learning and Optimal Combination Ensemble Approach for AQI Forecasting Using Big Data with Meteorological Conditions", et al.

This study proposes a multi-scale deep learning framework combined with an ensemble approach for accurate AQI forecasting using large-scale pollution and meteorological data.Different deep learning models such as LSTM and CNN are integrated to capture temporal and spatial features at multiple scales.An optimal combination strategy is used to improve prediction performance by aggregating multiple model outputs.The results show significantly higher accuracy and reduced prediction errors compared to individual models.This is related to our work as it also

utilizes advanced deep learning techniques and meteorological data to enhance air quality prediction accuracy.

Singh et al. [8] (2021), "A Comparative Analysis for Air Quality Estimation from Traffic and Meteorological Data", et al.

This study performs a comparative analysis of air quality estimation using traffic data and meteorological parameters.Various machine learning models such as Random Forest, Support Vector Machines, and Regression techniques are applied.The results show that combining traffic and meteorological data improves prediction accuracy compared to using a single data source.The study highlights the significant impact of vehicular emissions and weather conditions on air pollution
levels.This is related to our work as it also focuses on integrating multiple data sources to enhance air quality prediction models.

Chen et al. [9] (2020), "Using Machine Learning Algorithms to Study the Relationship Between Meteorological Conditions and Air Quality Parameters", et al.

This study applies machine learning algorithms to analyze the relationship between meteorological factors and air quality parameters.Models such as Random Forest, Support Vector Machines, and Neural Networks are used to capture complex patterns in the data.The results show that meteorological conditions like temperature, humidity, and wind speed strongly influence pollutant levels.

The study demonstrates that machine learning models can effectively model non-linear relationships in air quality data.This is related to our work as it also uses machine learning techniques to understand and predict air quality based on environmental factors.

Patel et al. [10] (2021), "A Complete Air Pollution Monitoring and Prediction Framework", et al.

This study proposes a comprehensive framework for monitoring and predicting air pollution using sensor data and environmental parameters.The framework integrates data collection, preprocessing, and prediction using machine learning models such as Random Forest and Neural Networks.The results show improved accuracy in real-time air quality monitoring and forecasting.The study highlights the importance of end-to-end systems for effective pollution management.This is related to our work as it also focuses on building a complete pipeline for air quality prediction using machine learning techniques.

III. System Analysis

Air pollution is a critical environmental issue affecting human health and ecosystems. Accurate forecasting helps in taking preventive actions and policy decisions. Traditional methods are limited in handling complex environmental relationships. With the availability of meteorological and pollution datasets, advanced analysis is possible. The system must process time-series data such as temperature, humidity, wind speed, and pollutant levels. It should capture both temporal and spatial dependencies. Machine learning and deep learning models can significantly improve prediction accuracy. The system must handle missing and noisy data efficiently. Real-time prediction capability is important for practical use. Visualization of predictions is necessary for decision-making. Scalability is required for large datasets. Overall, an intelligent and data-driven forecasting system is needed.

Existing System

Existing systems mainly use statistical models such as ARIMA and linear regression for air pollution forecasting. These models assume linear relationships and fail to capture complex patterns. Some systems use basic machine learning algorithms like decision trees. However, these approaches are limited in handling time-series dependencies. Integration of meteorological data is often incomplete. Existing systems struggle with non-linear and seasonal variations. Handling of large datasets is inefficient. Real-time forecasting is rarely supported. Many systems lack proper preprocessing and feature selection. Visualization tools are limited. Overall, existing systems provide basic predictions but lack accuracy and scalability.

Disadvantages of Existing System

- Poor handling of non-linear relationships
- Low accuracy in long-term forecasting
- Limited use of meteorological features
- Inefficient handling of large datasets
- Lack of real-time prediction
- Poor feature selection techniques
- Limited scalability

Proposed System

The proposed system uses the MPD dataset combining meteorological and pollution data. It integrates machine learning and deep learning models for accurate forecasting. Algorithms such as Random Forest, Support Vector Machine (SVM), and LSTM are used. The system captures temporal patterns using deep learning models. Data preprocessing techniques handle missing and noisy values. Feature selection methods identify important variables. The system provides both short-term and long-term predictions. Visualization tools help in understanding trends. The system supports real-time data updates. It is scalable for large datasets. The model is optimized using hyperparameter tuning. Overall, it provides a robust and efficient forecasting solution.

Advantages of Proposed System

- High prediction accuracy
- Captures temporal and non-linear patterns
- Effective use of meteorological data
- Scalable for large datasets
- Real-time forecasting capability
- Improved feature selection
- Better visualization and interpretation

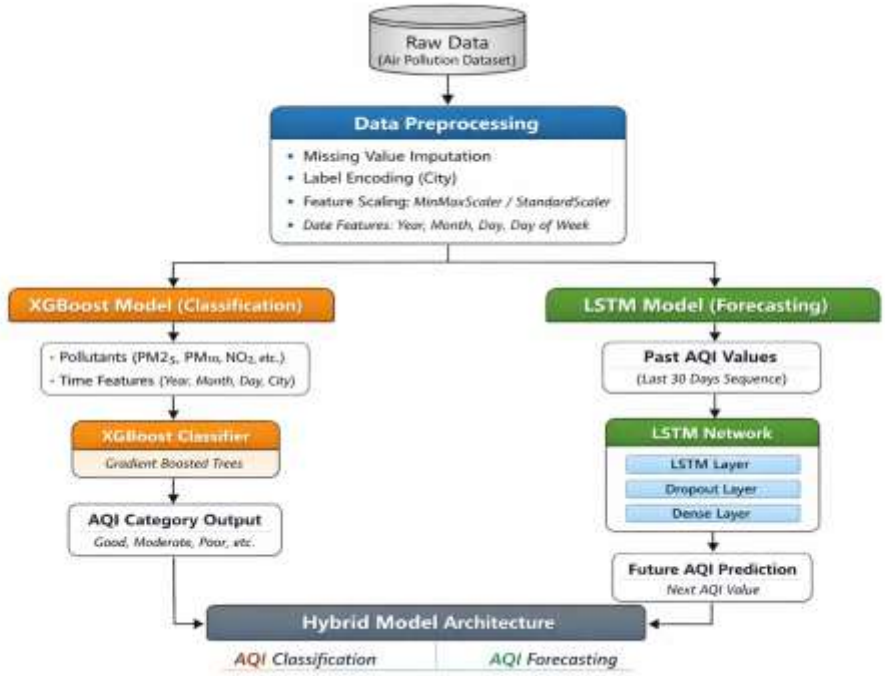
IV. Methodology

The methodology begins with collecting meteorological and pollution data from the MPD dataset. Data preprocessing is performed to clean and normalize the dataset. Missing values are handled using imputation techniques. Feature selection is 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 prediction. Model performance is evaluated using

MAE, RMSE, and R^2. Hyperparameter tuning improves model performance. Visualization tools are used for analysis. The best model is selected for deployment. The system is tested for real-time forecasting.

System Architecture

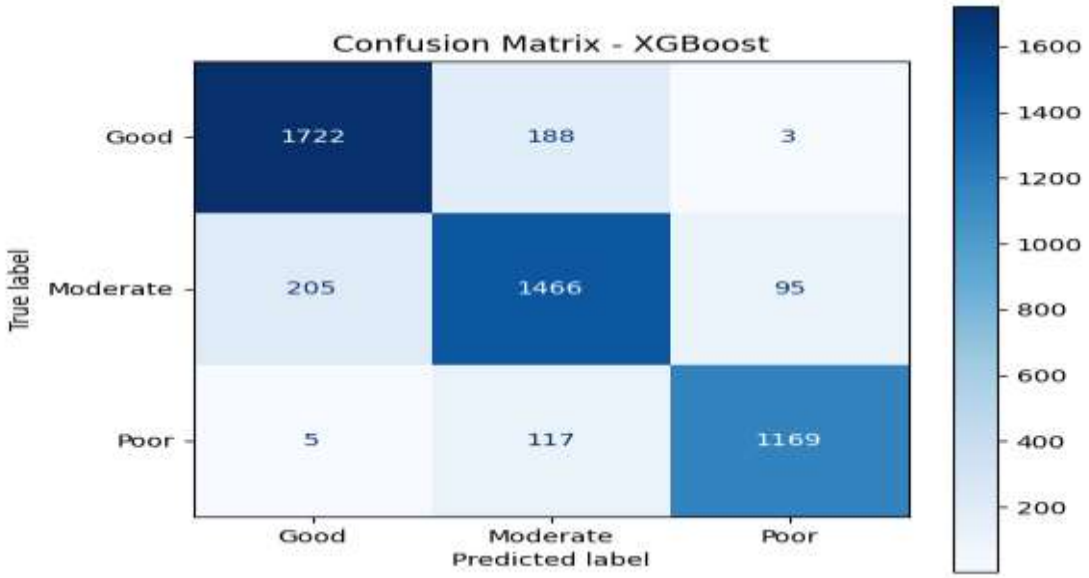
The system architecture consists of multiple layers. The data collection layer gathers meteorological and pollution data. The preprocessing layer cleans and prepares the dataset. The feature selection layer identifies important attributes. The model layer includes ML and DL algorithms for forecasting.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

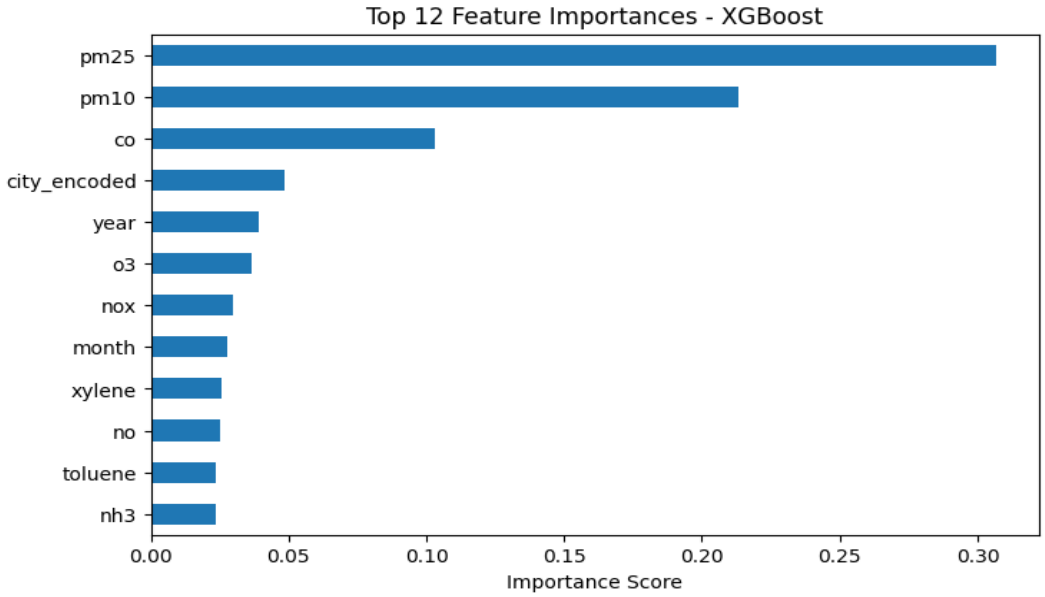


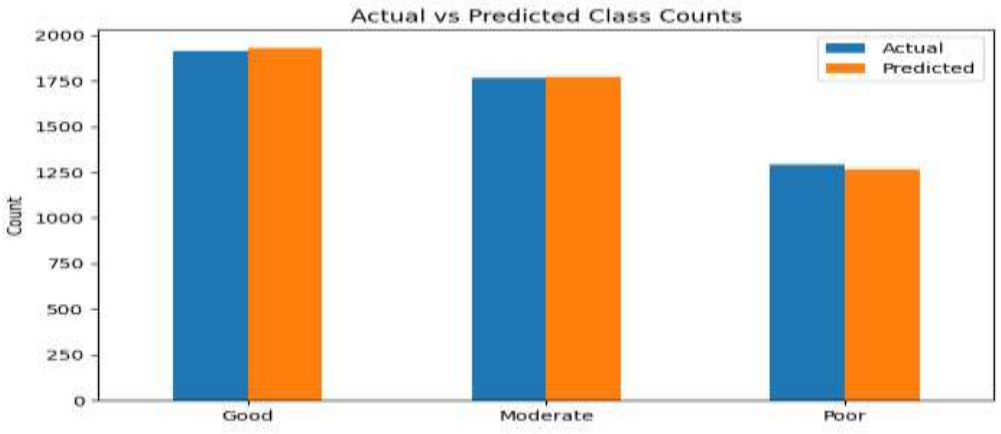
V. Result and Output

```
🔹 Enter pollution values manually
City: hyderabad
Date (YYYY-MM-DD): 2025-05-27
PM2.5: 120
PM10: 180
NO: 20
NO2: 40
NOx: 20
NH3: 35
CO: 25
SO2: 1.5
O3: 15
Benzene: 13
Toluene: 3
Xylene: 5

✅ Predicted AQI Category: Poor
✅ Forecasted Next AQI Value: 118.82
✅ Forecasted Next AQI Category: Moderate
```

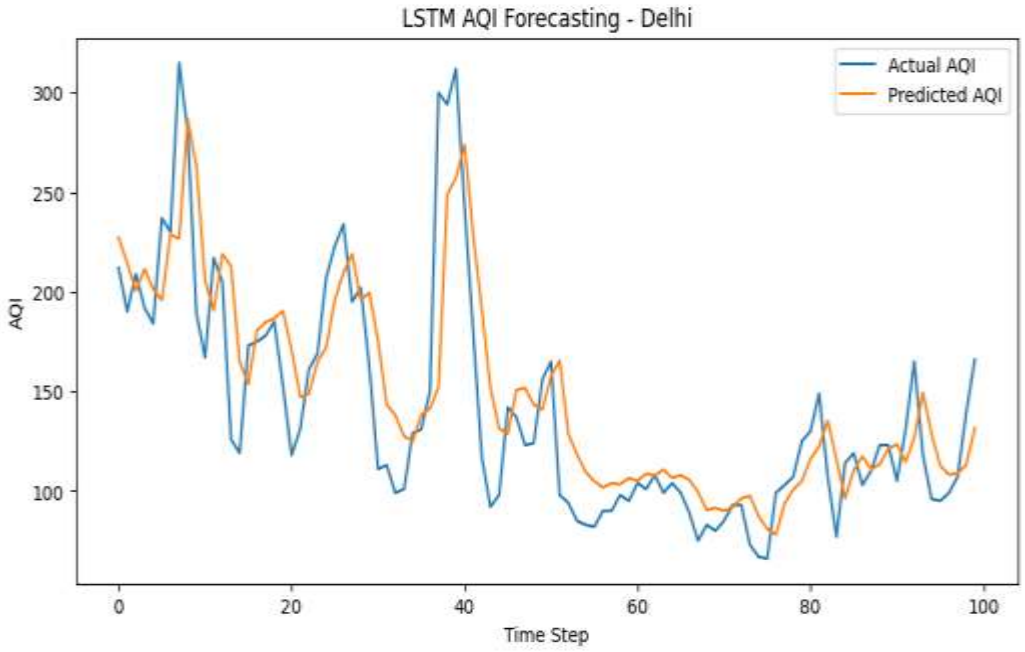
```
✅ XGBoost Accuracy: 87.67 %

Classification Report:

              precision    recall  f1-score   support

        Good       0.89      0.90      0.90      1913
    Moderate       0.83      0.83      0.83      1766
        Poor       0.92      0.91      0.91      1291

    accuracy                           0.88      4970
   macro avg       0.88      0.88      0.88      4970
weighted avg       0.88      0.88      0.88      4970
```




Top 12 Feature Importances - XGBoost
pm25
pm10
co
city_encoded
year
o3
nox
month
xylene
no
toluene
nh3
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Importance Score


Actual vs Predicted Class Counts
Actual
Predicted
Count
2000
1750
1500
1250
1000
750
500
250
0
Good
Moderate
Poor


LSTM Training Loss
Train Loss
Validation Loss
Loss
0.022
0.020
0.018
0.016
0.014
0.012
0.010
0.008
0.006
0
5
10
15
20
25
30
Epoch



VI. Conclusion

The air pollution analysis project provides a comprehensive understanding of how various pollutants influence overall air quality and the Air Quality Index (AQI). By performing data preprocessing, feature extraction, and exploratory analysis, the project successfully identifies key pollutants such as PM2.5, PM10, NOx, CO, SO_2, and O_3 as major contributors to environmental degradation. The analysis highlights noticeable variations in pollution levels across different locations and time periods, emphasizing the growing concern of air pollution in urban areas. Through effective data handling and visualization techniques, complex datasets were transformed into meaningful insights that can support informed decision-making. Although the project faced certain limitations, such as missing data and limited feature scope, it still demonstrates the potential of data-driven approaches in environmental monitoring. Furthermore, the project lays a strong foundation for future enhancements, including real-time monitoring and predictive modeling. Overall, this work underscores the importance of continuous air quality assessment and the need for proactive measures to reduce pollution levels, protect public health, and promote sustainable development for a cleaner and healthier environment.

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.