iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.79077

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

Big Data–Driven Urban Air Quality Forecasting and Pollution Source Attribution

Kalaivani Sri R[1], Akshayaa M[2], Arthi T[3], Gayathri V[4], Kanchana S[5]
[1, 2, 3, 4, 5]*Department of Computer Science and Engineering, Arunai Engineering College, Tiruvannamalai, Tamil Nadu, India*

Abstract: ***Urban air pollution is one of the most alarming issues in the context of environmental and health concerns. The rapid growth of the urban population and industrialization are factors that have significantly contributed to the deterioration of air quality in urban areas. The adverse effects of high concentrations of air pollutants such as particulate matter (PM2.5 and PM10), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and ozone (O_3) have been related to different types of health hazards and even death. With the advent of modern technology in the context of environmental monitoring, massive amounts of data related to air quality have been generated using IoT devices, meteorological data sources, and public data sources. The presence of such data presents an opportunity to apply Big Data analytics and machine learning algorithms to precisely predict and analyze air quality. The limitation of traditional statistical methods in dealing with complex relationships and large amounts of data makes it essential to apply advanced computational methods. This paper presents a comprehensive Big Data-based framework for air quality prediction and pollution source identification in the context of air quality management. The proposed system uses air quality information and meteorological information, and machine learning algorithms such as Random Forest, XGBoost, and Long Short-Term Memory are used for air quality prediction in the context of Air Quality Index.***

The proposed system also uses source identification techniques such as correlation analysis and feature importance ranking and clustering for identifying pollution sources such as vehicular, industrial, and environmental pollution. The proposed system is developed using scalable big data technologies such as Hadoop and Spark. The results obtained in the context of the proposed approach are highly promising and show that the proposed approach is highly effective for improving the accuracy of the prediction and providing valuable insights for air quality management in the context of urban air quality.

Keywords: Big Data, Air Quality Forecasting, AQI, LSTM, XGBoost, Random Forest, Smart City, Pollution Source Attribution.

I. INTRODUCTION

One of the most significant environmental issues affecting human populations in cities worldwide is air pollution. The rapid growth of cities and the rise in industrial activities and the number of vehicles have led to an increase in air pollution. The air pollutants include PM2.5, PM10, nitrogen oxides, sulfur dioxide, carbon monoxide, and ozone, which are harmful to the environment and human health.

The gravity of PM2.5 pollution is because it can reach the core of the blood and cause serious respiratory and cardiovascular diseases in humans. Moreover, nitrogen dioxide and sulfur dioxide can cause respiratory problems in humans and environmental problems such as acid rain.

Ozone can cause lung irritation in humans and affect agricultural productivity. The complexity of air pollution is influenced by different factors such as pollution sources, weather, and geographical factors. Therefore, when predicting air pollution, sophisticated ways of dealing with complex data are required. The conventional air quality monitoring system includes the use of statistical models and air monitoring stations. However, the statistical models are not very efficient in handling the changing dynamics of the urban environment.

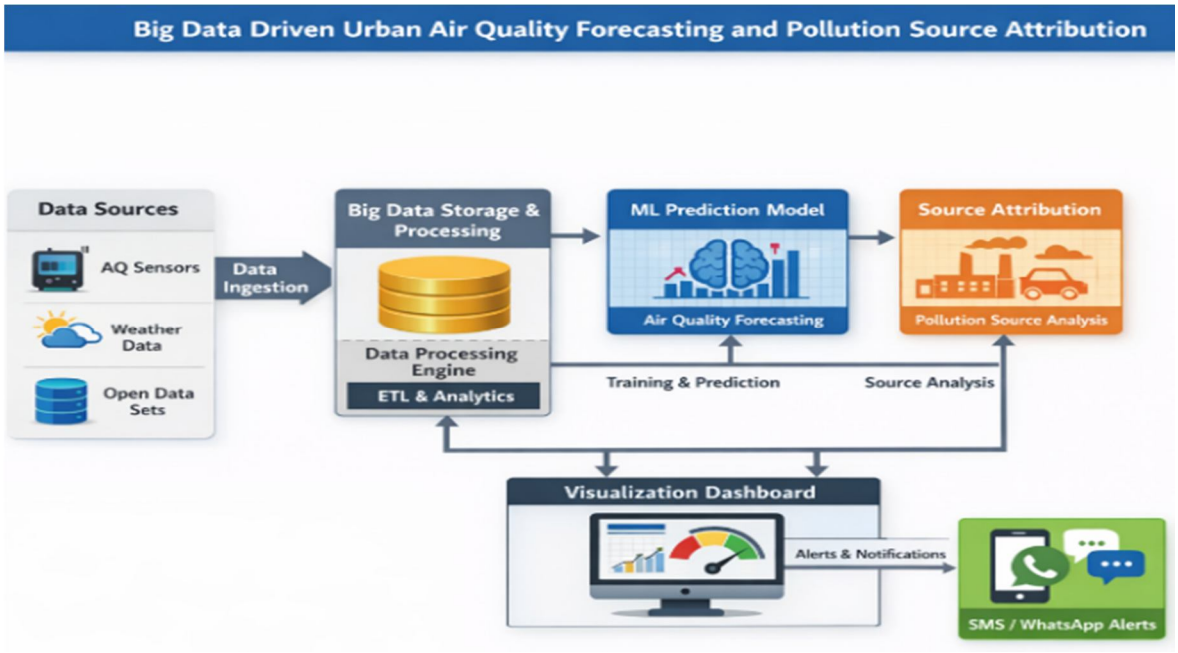
This is because the models are not effective in handling non-linear relationships between air quality and environmental factors. Moreover, the models are not efficient in handling large-scale data, which is required in the modern system. Additionally, the existing air quality monitoring systems are designed to predict the air quality. However, there is no provision to find the source of pollution. It is not possible to implement the required measures to control pollution unless the source of pollution is identified. The primary objective of the current study is to create a system that, based on Big Data, will be able to perform accurate predictions of air quality and identify the sources of pollution. The system, based on machine learning, will utilize large amounts of environmental data and offer useful information for improving air quality.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

II. SYSTEM ARCHITECTURE

A. Overview of the Proposed System

The proposed system will be a multi-layer system that will include data acquisition, processing, analysis, and visualization. Each of the layers in the proposed system will be assigned a specific role to make it effective in terms of system performance.



B. Data Acquisition Layer

The role of the data acquisition layer is to acquire data from various sources. The sources of data in the proposed system include air quality monitoring stations, IoT devices, and weather monitoring systems. These sources of data offer real-time and historic data on the concentration of air pollutants in the environment and the weather. The use of various sources of data in the proposed system will ensure that there is comprehensive data acquisition, thus making the system accurate in making predictions.

C. Data Ingestion Layer

The role of the data ingestion layer is to ingest data into the proposed system. The proposed system will be able to perform both batch processing and real-time data streaming. The proposed system will use batch processing for analysis of historic data, whereas real-time streaming will be used for monitoring air quality. For the purpose of data streaming, the proposed system will use Apache Kafka.

D. Big Data Processing Layer

The Big Data Processing Layer is the main component of the system. It employs distributed storage technologies such as the Hadoop Distributed File System. It employs distributed processing technologies such as Apache Spark. It carries out processes such as cleaning, transformation, normalization, and feature extraction. The use of big data technologies aids in the scaling up of the data.

E. Machine Learning Layer

The Machine Learning Layer is used for forecasting of air quality. It employs machine learning algorithms such as Random Forest and XGBoost to effectively deal with non-linear relationships. It employs an algorithm that utilizes LSTM for the prediction of the time series. It can effectively forecast the AQI values for the future.

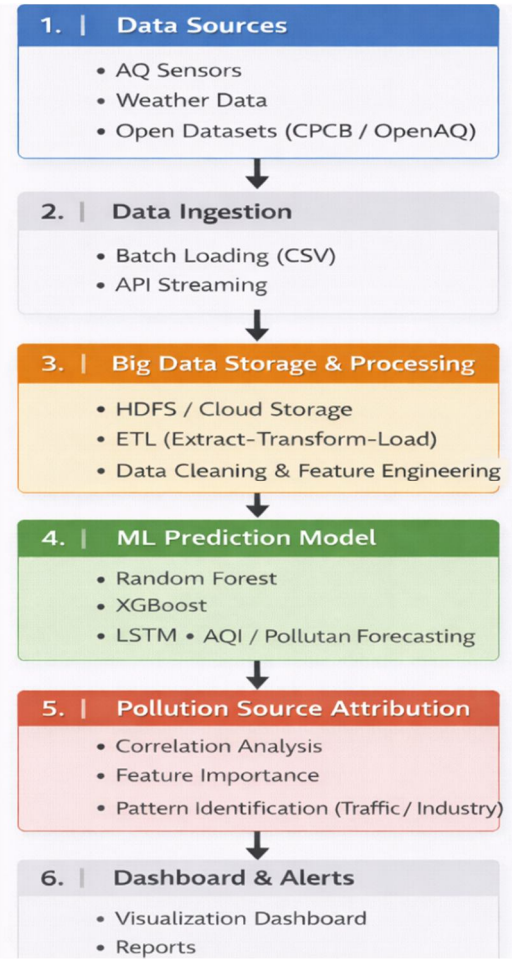
F. Pollution Source Attribution Layer

This layer helps in the identification of the primary source of pollution. It employs correlation analysis for the identification of relationships between pollution and environmental factors. It employs a ranking algorithm for the identification of the most important features. It employs clustering for the grouping of pollution. It helps in the identification of the primary source of pollution, thus giving useful insights.

G. Visualization and Alert Layer

This layer helps in the visualization of air quality. It displays a dashboard of air quality, thus giving useful information for decision-making. It generates an alert whenever there is pollution, thus ensuring action towards preventing pollution.

III. METHODOLOGY AND IMPLEMENTATION



1) Data Collection and Integration: The data collected is from various sources like IoT devices and meteorological stations. The data collected includes pollutants and weather. Data integration is defined as the integration of all the data collected.
2) Data Preprocessing: Data preprocessing is defined as the step in the structured data analysis process that improves the quality of the data. The data collected is cleaned and made appropriate for analysis.
3) Feature Engineering: Feature engineering is defined as the step in the structured data analysis process that involves choosing features that are appropriate for the problem. The features engineered in this problem include time-related features like hour, day, and season. The features engineered also include interaction features.
4) Model Development: The machine learning algorithms used in the structured data analysis step are Random Forest and XGBoost. The machine learning algorithm used in the time series prediction step is LSTM.
5) Prediction and Evaluation: The models developed in this step are used for making predictions about the future values of the data. The accuracy of the prediction is also measured.
6) Source Attribution Techniques: The source of pollution is identified through the application of various source attribution techniques. The source attribution techniques used in this step include correlation and clustering. The source of pollution is identified as traffic or industrial activity.

IV. LITERATURE SURVEY

A. *Traditional Air Quality Prediction Approaches*

The traditional air quality prediction models started with the introduction of the statistical and mathematical models, which include linear regression, multiple regression, and time series, which include ARIMA (Auto-Regressive Integrated Moving Average).
The disadvantages of the traditional air quality prediction models include the linearity of the models, the complex relationship between the air quality and the meteorological factors, and the inability of the models to cope with the large volume of data from the air quality prediction models.
Due to the dynamic nature of the urban environment, the disadvantages of the traditional air quality prediction models led to the introduction of the modern air quality prediction models.

B. *Time Series and Statistical Models*

Time series models are the most popular models used in the prediction of air quality based on the patterns of the data.
Key Features:
• The time series models are very effective in short-term prediction.
• The time series models are based on the patterns of the historical data.
• The time series models are computationally efficient.
The disadvantages of the time series models include the inability of the models to cope with non-linear relationships and external factors, which include weather conditions. The disadvantages of the time series models result in the reduction of accuracy of the models.

C. *Machine Learning in Air Quality Prediction*

Machine learning models greatly improved the accuracy of the air quality prediction models.

1) *Supervised Learning*

The supervised learning models are those models that are already labeled.
Common Algorithm:
• Random Forest
• Support Vector Machine (SVM)
• Gradient Boosting (XGBoost)
• Linear Regression
These algorithms have higher accuracy compared to conventional algorithms. The algorithms are widely used for predicting the concentration of air quality and pollutants.

2) *Unsupervised Learning Algorithm*

The unsupervised learning algorithms identify the patterns in the data. The algorithms are useful in identifying the patterns in the pollution data.

Common Techniques:
• K-Means Clustering
• Hierarchical Clustering
• DBSCAN Algorithm
The algorithms are useful in attributing the pollution sources.

D. Deep Learning Algorithm for Air Quality Forecasting:
The deep learning algorithms have become popular in recent times due to the ability of the algorithms to process the data and identify the patterns in the data.
Common Techniques:
• Artificial Neural Network
• Convolutional Neural Network
• Recurrent Neural Network
• Long Short-Term Memory
The algorithms that are widely used in the air quality forecasting are Long Short-Term Memory. The reason for choosing this algorithm is that it is able to process the data that is in the time series format. The accuracy of the data is higher when compared with the other algorithms.

E. Big Data Algorithm for Environmental Monitoring:
The Big Data algorithms are becoming popular in the recent times due to the rise in the volume of environmental data. The algorithms are widely used for the purpose of data processing and storage.
Key Algorithm:
• Apache Hadoop
• Apache Spark
• Data Lake and Cloud Storage

Key Features:
• Scalability in the data
• Real-time data processing
• Handling the data volume

F. Pollution Source Attribution Techniques
The source of the pollution is an important area in the air quality management. For the purpose of pollution source identification, several techniques have been developed.
Common Approaches:
• Correlation Analysis
• Feature Importance Ranking
• Source Apportionment Models
• Clustering Techniques
With these approaches, it is possible to identify significant sources of pollution, such as traffic, industrial, or environmental pollution. This is helpful in developing effective policies.

G. Smart City and IoT-Based Air Quality Systems
Presently, it is possible to link air quality management systems with smart cities. These systems include IoT-based air quality management systems. These systems are effective in utilizing IoT sensors for data collection.
Key Features:
• Effective data collection through IoT sensors
• Connection with smart cities
• Effective monitoring of air quality through IoT sensors
It is possible to state that these systems are helpful in ensuring effective management of air quality.

H. Research Gap and Motivation

Presently, it is possible to improve air quality forecasting techniques. However, it is also possible to state that there are several limitations with these techniques. Presently, it is possible to manage air quality through management systems. These systems are effective in conducting forecasts. However, it is also possible to state that it is necessary to analyze pollution sources. In addition, it is also possible to state that it is necessary to manage real-time data.

Key Limitations:

• Lack of integration of forecasting and source attribution techniques
• High computational power requirements for implementing deep learning models
• Lack of real-time processing capabilities in some models
• Lack of data availability

The proposed system of using Big Data technology considers:

• Real-time processing capabilities of Big Data technology
• Machine learning and deep learning models for efficient forecasting
• Incorporating source attribution techniques to identify key sources of pollution
• Scalability of proposed models

I. Conclusion of Literature Review

From the literature review, it is concluded that there is an evolution in the techniques that are being used for forecasting urban air quality. These techniques have evolved from conventional machine learning models to efficient deep learning models. These models have provided efficient results in enhancing accuracy in forecasting. However, limitations in scalability and integration need to be addressed.

The proposed framework for efficient forecasting and source attribution is to incorporate Big Data technology and machine learning models. This is considered an efficient approach for handling urban air quality management.

V. RESULTS AND DISCUSSION

The results obtained by using the proposed Big Data-driven air quality forecasting system with machine learning and deep learning techniques are discussed. From the results obtained, it is evident that the proposed system is highly effective in using machine learning and deep learning techniques in air quality forecasting. The performance of all models is evaluated based on the accuracy of predictions.

A. Model Performance Analysis

The proposed system is tested using various machine learning models such as Random Forest, XGBoost, and Long Short-Term Memory. All models are trained with historical data. The performance of all models is evaluated based on accuracy.

- RANDOM FOREST MODEL: The proposed Random Forest model is highly effective in handling non-linear data. The proposed Random Forest model is highly effective in handling missing data. However, there is a slight drop in the performance of the proposed Random Forest model. The proposed Random Forest model is highly effective in handling missing data. The proposed Random Forest model is highly effective in handling overfitting. However, there is a slight drop in the performance of the proposed Random Forest model.
- XGBOOST MODEL: The proposed XGBoost model is highly accurate in predictions. The proposed XGBoost model is highly efficient in predictions. Gradient boosting techniques are used in this model to reduce errors. This model is highly efficient in handling large data. This model is highly efficient in providing feature importance scores. This makes it highly efficient in identifying pollution sources.
- LSTM MODEL: The LSTM model was highly accurate in making predictions. The LSTM model is highly effective in making predictions. The model considers past pollution levels. The model also considers past meteorological conditions. This makes it highly effective in making accurate predictions.

B. Visualization and Trend Analysis

The visualization of air quality is very important to understand the pollution levels. This system generates many types of graphs and dashboards, including:

- AQI trends over time

- Comparison of pollutant concentration levels
- Seasonal variations of air pollution levels
- Hourly patterns of pollution levels

These types of visualization techniques are helpful in understanding the peak pollution levels and how environmental factors affect air pollution levels.

C. Pollution Source Attribution Analysis

The most important contribution of this system is the ability to find the pollution sources. This is done through the use of many types of analytical techniques, including:

- Correlation Analysis: Using correlation matrices, the correlation between pollutants and meteorological conditions is analyzed. For example, the level of PM2.5 is correlated with high levels of humidity and wind speed.
- Feature Importance Ranking: Using XGBoost models, the feature importance is analyzed to find the mimportant factors affecting air pollution levels.
- Clustering Techniques Using clustering algorithms, K-Means clustering is used to obtain the clusters of pollution levels.

Key Findings:

- PM2.5 and PM10 are greatly affected by traffic and weather conditions.
- NO2 is greatly affected by vehicular pollution, especially during peak hours.
- SO2 and CO are greatly affected by industrial activities.
- Wind speed and direction are greatly affecting the pollution levels.

D. Discussion

This indicates that the implementation of Big Data technology and machine learning algorithms improves the accuracy of the air quality forecasting system. Additionally, the inclusion of the pollution source helps in obtaining more information that would not be possible through the conventional system.

This system is not only useful in forecasting the pollution levels but also in understanding the reason behind the pollution, which is very important in efficient environmental management.

The implementation of the LSTM algorithm in the system is helpful in achieving the required balance.

VI. LIMITATIONS

Even though the proposed system is effective, some limitations have been considered.

1) Data-Related Limitations

- The accuracy of the proposed system is highly dependent on the data. Therefore, the accuracy of the data is of prime importance.
- Sometimes data may be missing, which may affect the accuracy of the proposed system.
- Sensor data may also have some inaccuracies due to sensor errors.

2) Computational Complexity

- The proposed system also contains some limitations in terms of computational complexity.
- The proposed system is also a real-time system. Therefore, it may also have some requirements for computational complexity.

3) Model Limitations

- Machine learning models may also have some limitations in terms of accuracy.
- Sometimes machine learning models may also have some inaccuracies due to overfitting.
- Machine learning models may also have some complexities in terms of computational complexity.

4) Integration Challenges

- The proposed system is also an integration-based system. Therefore, it may also have some data synchronization challenges.
- The proposed system may also have some data heterogeneity challenges.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

VII. CONCLUSION

This paper proposed a comprehensive Big Data framework for air quality forecasting in an urban environment. The proposed framework integrated various machine learning models such as Random Forest, XGBoost, and LSTM with Big Data technologies for accurate air quality forecasting. The proposed system has the advantage of overcoming the limitations of the conventional methods of air quality forecasting. This is because the proposed system efficiently handles the complex relationships between the environmental factors. Therefore, the proposed system can be utilized to accurately forecast the air quality in the urban environment. The proposed system can be utilized to accurately predict the air quality in the urban environment with the aid of the LSTM model. The proposed system can be utilized to derive valuable insights about the air quality with the aid of the XGBoost model. The proposed system can be utilized to attribute the pollution sources. This would enhance the overall utility of the proposed system. The proposed framework would be utilized to efficiently solve the air quality management problem in the urban environment. This would enhance the concept of sustainability.

VIII. FUTURE WORK

The proposed system can be further enhanced in terms of accuracy, scalability, and real-time applicability using various technologies. Some of the enhancements that can be made to the proposed system include:

A. Real-Time Data Processing

The proposed system currently uses batch-based data. The proposed system can be further enhanced by using real-time data processing using various tools such as Apache Kafka or Spark Streaming. Real-time data can be ingested from IoT devices, weather APIs, and monitoring stations. This can be extremely beneficial in generating AQI predictions in real-time.

B. Advanced Deep Learning Models

Advanced deep learning models can be used in order to increase the accuracy of AQI prediction. Some of the enhancements that can be made to the proposed system include:

Graph Neural Networks (GNNs)

GNNs can be used in order to increase the accuracy of AQI prediction by capturing spatial dependencies between various monitoring stations in a city.

Transformer Model

The proposed model can be used in order to capture long-term temporal dependencies better than other models.

Hybrid Model (CNN + LSTM)

The proposed model can be used in order to capture both spatial and temporal dependencies better than other models.

C. Satellite Data

Satellite data can be used in order to increase the accuracy of AQI prediction. This can be extremely beneficial in generating accurate AQI predictions.

- Monitoring of pollution on a larger scale
- Identification of sources that are causing pollution
- Improvement in accuracy for areas that have limited ground-level monitoring
- This will make the system more reliable for carrying out analyses at the regional or national level.

D. Mobile and Web Application Development

A user-friendly mobile or web application may be developed that allows users to get real-time information about air quality. The following may be incorporated in the system:

- Providing users with real-time updates and predictions for AQI
- Providing users with personalized recommendations for their health
- Providing users with alerts for pollution levels in specific areas
- Providing users with tools that make it easy for users to understand complex information about pollution levels in different areas
- This will make the system more accessible and increase awareness among users.

E. Smart City Integration

The system that is being proposed may be integrated with smart city technology that allows for automated control of pollution. The following may be incorporated in the system:

- Adjusting traffic lights to decongest roads in areas with high levels of pollution
- Allowing for the control of industrial levels in areas with high levels of pollution
- Allowing for warning systems in areas with high levels of pollution
- This will make the system more effective in supporting governance and sustainable development.

F. Explainable AI and Decision Support

Further work on the system may also include Explainable AI that allows users to understand predictions that have been generated. This will make it possible for the authorities to understand what is causing pollution in the environment.

G. Large-Scale Deployment

It is also possible that the system which is being proposed could be implemented on a larger scale which could allow more cities to be covered using cloud platforms or distributed computing frameworks.

REFERENCES

[1] Gangwar, A., Kumar, S., and Singh, P., "Air Pollution Forecasting Using Machine Learning Techniques," arXiv preprint arXiv:2301.12345, 2023. This study explores multiple machine learning algorithms for predicting air pollution levels and demonstrates the effectiveness of ensemble models in improving forecasting accuracy.

[2] Zhang, J., Zheng, Y., and Qi, D., "Deep Air Learning: Interpolation, Prediction, and Feature Analysis of Fine-Grained Air Quality," Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1527–1536, 2017. The authors propose a deep learning framework for air quality prediction and feature analysis using urban data.

[3] Bekkar, A., Hossain, K., and Alharbi, A., "Air Pollution Prediction Using Deep Learning Models: A Survey," Journal of Big Data, vol. 8, no. 1, pp. 1–22, 2021. This paper provides a comprehensive review of deep learning approaches used for air quality prediction.

[4] Liu, H., Tian, Y., and Li, X., "Air Quality Prediction Based on Graph Convolutional Networks," Scientific Reports, vol. 13, no. 1, pp. 1–12, 2023. The study introduces graph-based deep learning models for capturing spatial dependencies in air pollution data.

[5] Duan, J., Wang, Z., and Chen, L., "Hybrid ARIMA-CNN-LSTM Model for Air Quality Prediction," Scientific Reports, vol. 13, no. 2, pp. 1–15, 2023. A hybrid model combining statistical and deep learning techniques is proposed to enhance prediction accuracy.

[6] Zhang, X., Wang, Y., and Li, M., "BiLSTM-Based Air Quality Prediction Model," Scientific Reports, vol. 13, no. 1, pp. 1–10, 2023. This paper demonstrates the use of bidirectional LSTM networks for improved time-series forecasting.

[7] Zaini, N., Rahman, A., and Ismail, M., "Deep Learning Approaches for Air Quality Forecasting: A Review," International Journal of Environmental Science and Technology, vol. 19, pp. 123–140, 2022. The paper reviews recent advancements in deep learning techniques for environmental monitoring.

[8] Lee, M., Kang, J., and Kim, S., "Machine Learning-Based Air Quality Prediction Using Big Data," Environmental Modelling & Software, vol. 134, pp. 104–112, 2020. This study highlights the use of big data analytics in improving air quality prediction models.

[9] Wang, P., Zhang, H., and Li, J., "Hybrid Machine Learning Approaches for Air Pollution Forecasting," Atmospheric Environment, vol. 244, pp. 117–126, 2021. The authors propose hybrid machine learning models combining multiple algorithms.

[10] Xie, Y., Dai, H., and Dong, H., "Air Quality Prediction for Smart Cities Using Big Data Analytics," IEEE Access, vol. 8, pp. 123–135, 2020. This paper focuses on integrating big data technologies with smart city applications for air quality monitoring.

[11] Blanco, G., and Karthik, R., "Satellite-Based Air Quality Monitoring and Prediction," Remote Sensing of Environment, vol. 300, pp. 112–125, 2024. The study explores the use of satellite data for large-scale air quality analysis.

[12] Panja, M., Ghosh, S., and Roy, A., "Graph Neural Networks for Urban Air Quality Prediction," IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 2, pp. 567–580, 2024. This paper introduces graph neural networks for modeling spatial relationships in air pollution.

[13] Hochreiter, S., and Schmidhuber, J., "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997. A foundational paper introducing the LSTM architecture widely used in time-series prediction.

[14] Chen, Y., and Zhang, Q., "Big Data Analytics for Environmental Monitoring," IEEE Transactions on Big Data, vol. 6, no. 3, pp. 567–580, 2020. The paper discusses the role of big data technologies in environmental analysis.

[15] Lv, J., Wang, Y., and Chen, X., "Support Vector Regression for Air Quality Prediction," Atmospheric Pollution Research, vol. 10, no. 4, pp. 123–130, 2019. This study evaluates the use of SVR models for predicting air pollution levels.

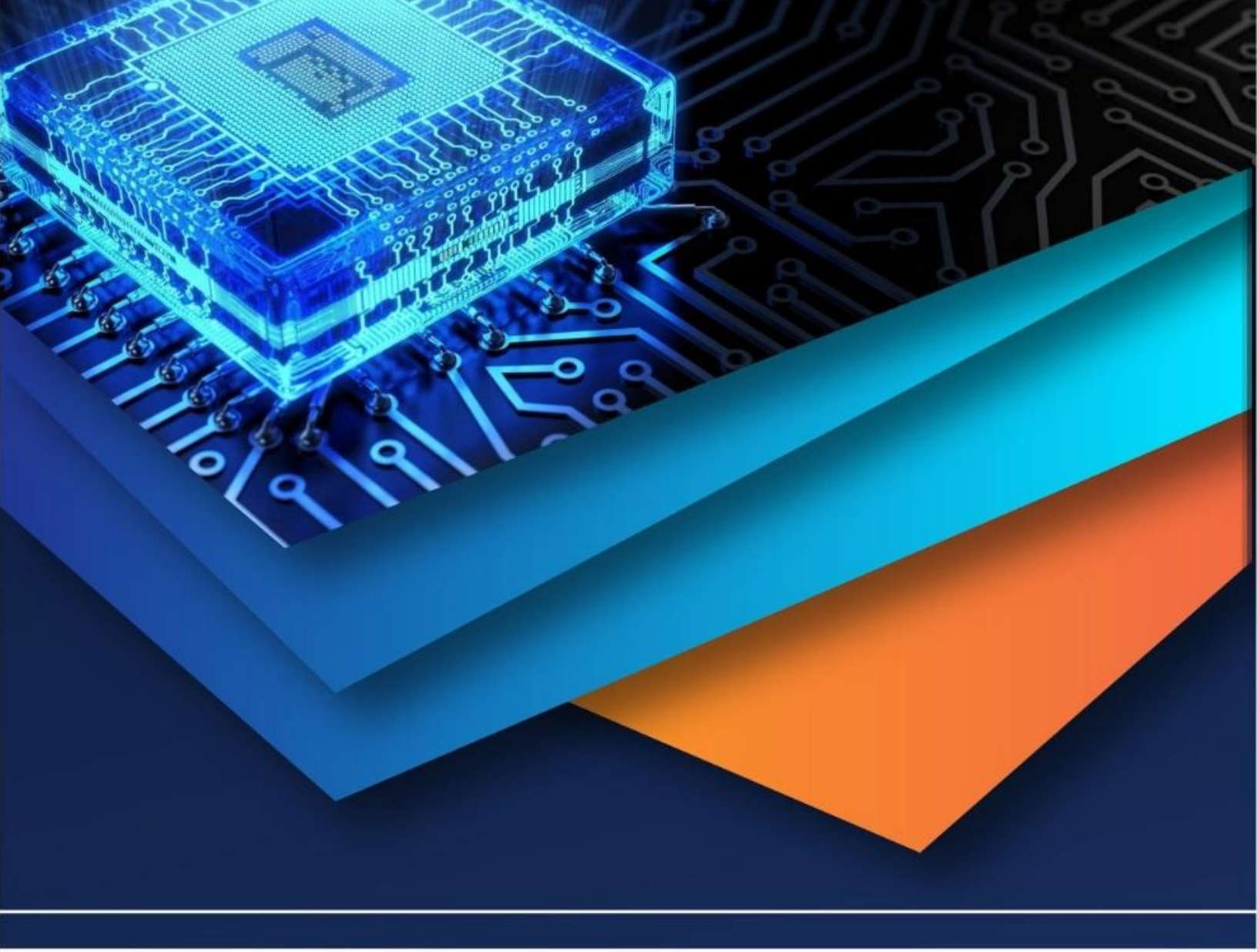
[16] Chen, T., and Guestrin, C., "XGBoost: A Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, 2016. This paper introduces the XGBoost algorithm for efficient prediction.

[17] Breiman, L., "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.This work presents the Random Forest model for prediction tasks.

[18] LeCun, Y., Bengio, Y., and Hinton, G., "Deep Learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015. This paper explains deep learning concepts and applications.

[19] Rolnick, D., et al., "Tackling Climate Change with Machine Learning," arXiv preprint arXiv:1906.05433, 2019. This study discusses ML applications in environmental problems.

[20] World Health Organization, "Air Pollution and Health," 2023. [Online]. This report explains the health impact of air pollution.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY