iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.68610

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 13 Issue IV Apr 2025- Available at www.ijraset.com

A Survey on Air Pollution using Deep Learning

Akshayaa M[1], T. Rajasenbagam[2]
[1]*PG Scholar, Dept. of CSE, Government College of Technology, Coimbatore, India*
[2]*Assistant Professor, Dept. of CSE, Government College of Technology, Coimbatore, India*

Abstract: ***Air pollution remains a significant environmental and public health issue, with particulate matter being a major factor in declining air quality. This research is an attempt to examine the various methodologies for monitoring, classifying, and forecasting air quality using statistical approaches, time-series analysis, and machine learning techniques. Traditional models such as regression analysis, ARIMA, and time-series decomposition have been commonly employed for air quality evaluation. However, advancements in artificial intelligence (AI) have introduced more precise predictive models, including support vector machines (SVM), deep convolutional neural networks (DCNN), and long short-term memory (LSTM) networks. Furthermore, the impact of meteorological factors on pollutant dispersion and the effectiveness of urban greening strategies in reducing air pollution are explored. The findings indicate that hybrid AI models, which combine deep learning with statistical techniques, demonstrate superior predictive capabilities, presenting a promising approach for real-time air quality monitoring and informed decision-making.***

Keywords: Air pollution, SVM, DCNN, LSTM, ARIMA, time-series, meteorology, air quality prediction.

I. INTRODUCTION

Air pollution has become one of the most critical environmental challenges worldwide, significantly affecting human health, climate, and ecosystems. Among various air pollutants, particulate matter has received considerable attention due to its severe health implications, including respiratory diseases, cardiovascular disorders, and increased mortality rates. Accurate air quality monitoring and forecasting have become essential for developing effective mitigation strategies and public health policies.

A. History of Air Pollution

The study of air pollution dates back to ancient civilizations, where smoke and soot from burning wood and coal were recognized as environmental hazards. However, systematic air quality monitoring began in the 20th century, as industrialization and urbanization led to increasing levels of atmospheric pollutants. One of the earliest recorded environmental disasters, the Great Smog of London (1952), caused thousands of deaths and led to the enactment of the UK's Clean Air Act in 1956, marking the beginning of modern air pollution control policies.

In the 1960s and 1970s, the establishment of environmental protection agencies, such as the United States Environmental Protection Agency (EPA**)** in 1970, spurred research on air quality monitoring. Governments and scientific communities developed air pollution indices, such as the Air Quality Index (AQI), to quantify pollution levels. Initial monitoring relied on ground-based sensors and manual data collection, which provided limited real-time insights.

By the 1980s and 1990s, advancements in computational modeling led to the adoption of statistical methods for air quality forecasting. Researchers employed linear regression, autoregressive integrated moving average (ARIMA), and multiple regression models to predict pollutant levels based on historical data. These models improved understanding of pollution trends but struggled with nonlinear and dynamic atmospheric interactions. The early 2000s saw the rise of machine learning (ML) and artificial intelligence (AI) in air quality prediction. Techniques such as support vector machines (SVM), artificial neural networks (ANNs), and random forests provided improved predictive accuracy by learning complex relationships in pollution data. Additionally, the role of meteorological factors in air pollution dispersion became a major focus, leading to hybrid models that integrated weather conditions with pollutant measurements. In the 2010s and beyond, deep learning models, including convolutional neural networks (CNN) and long short-term memory (LSTM) networks, revolutionized air quality forecasting. These models leveraged big data, satellite imagery, and IoT-based sensor networks to provide real-time air quality monitoring. Studies also explored the impact of urban greening and nature-based solutions in mitigating pollution, emphasizing sustainable approaches to air quality management.

Today, air quality research continues to evolve with AI-driven predictive analytics, sensor technology, and cloud-based monitoring systems. The integration of AI with environmental science offers new possibilities for real-time air quality forecasting, pollution source identification, and decision support systems for policymakers and urban planners.

B. Challenges in air quality prediction

Air quality monitoring faces challenges such as sensor inaccuracies, data gaps, and limited geographic coverage, leading to unreliable measurements. The nonlinear and dynamic nature of pollution data, influenced by meteorological factors, makes prediction difficult for traditional models. While AI and deep learning improve accuracy, they require large datasets and high computational power. Additionally, the lack of interpretability in AI models and inconsistent air quality regulations hinder real-world implementation. Overcoming these challenges requires advanced AI models, improved sensor technology, real-time data integration, and stronger policy frameworks for effective air quality forecasting.

C. Deep Learning in air Quality Prediction

Deep learning has emerged as a transformative approach in the field of air quality prediction, significantly improving the accuracy and robustness of Air Quality Index (AQI) forecasting. Unlike traditional statistical models such as Auto-Regressive Integrated Moving Average (ARIMA), linear regression, or generalized additive models—which are often limited by their assumptions of linearity and stationarity—deep learning models are capable of learning complex, nonlinear relationships and temporal dependencies inherent in environmental data. These capabilities make deep learning particularly well-suited to handle the multifactorial nature of air pollution, which is influenced by diverse factors such as meteorological conditions (e.g., temperature, humidity, wind speed), urban activities, and seasonal variations. Among the most widely used deep learning models, Long Short-Term Memory (LSTM) networks have proven highly effective for time-series forecasting tasks due to their ability to capture long-term dependencies and temporal correlations in sequential data. This makes them particularly valuable in predicting future pollutant concentrations, where historical trends and cyclic behaviors are crucial. LSTMs outperform standard RNNs by mitigating the vanishing gradient problem, enabling better learning over extended time intervals. Convolutional Neural Networks (CNNs), although originally developed for image recognition, have been adapted for spatial data analysis in air quality research. These models can extract high-level spatial features from satellite imagery, aerial sensor data, and spatially distributed air quality measurements, making them useful for tasks like pollution source identification, urban pollution mapping, and high-resolution AQI estimation across regions. When integrated with geospatial and meteorological inputs, CNNs can significantly enhance the granularity and accuracy of AQI prediction models. Hybrid models that combine the temporal strengths of LSTMs with the spatial capabilities of CNNs have become increasingly popular. These CNN-LSTM architectures allow simultaneous learning from spatially distributed sensor data and historical time-series data, enabling more comprehensive and accurate forecasting. Studies have shown that these hybrid models often outperform individual deep learning models in both short-term and long-term AQI forecasting scenarios. Despite their high predictive performance, deep learning models also come with several challenges. These include the need for large volumes of high-quality labeled data, substantial computational resources for training, and difficulties in interpreting model decisions—a crucial aspect in environmental policy-making where explainability is important. Moreover, model performance may vary significantly across geographic regions due to differences in emission sources, meteorology, and monitoring infrastructure.

II. LITERATURE SURVEY

The literature review explores progress in air quality monitoring and prediction through statistical and AI-driven methods. Conventional models such as ARIMA and regression face challenges in handling the nonlinear characteristics of air pollution. In contrast, machine learning techniques like support vector machines (SVM), random forests, and artificial neural networks (ANN), along with deep learning models such as convolutional neural networks (CNN) and long short-term memory (LSTM) networks, achieve greater accuracy in AQI classification and forecasting. Research indicates that incorporating meteorological variables enhances predictive performance, while urban greening initiatives contribute to pollution reduction. The review underscores the importance of real-time monitoring, hybrid AI models, and IoT-enabled sensor networks for efficient air quality management.

Hussain et al. (2020) analyzed the classification of indoor and outdoor particulate matter (PM2.5 and PM10) using machine learning approaches. They applied Support Vector Machines (SVM), Decision Trees, and k-Nearest Neighbors (KNN) to classify air pollution data. Their study demonstrated that cubic and coarse Gaussian SVM classifiers achieved the highest accuracy (95.8%), emphasizing the effectiveness of AI in air quality classification [1].

Akbal and Unlu (2023) developed a hybrid deep learning model combining Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks to forecast PM2.5 levels in Ankara, Türkiye. Their study demonstrated that the model achieved an R^2 accuracy of 81%, surpassing traditional statistical methods in predictive performance. The results emphasize the effectiveness of deep learning in capturing complex pollutant patterns and temporal dependencies. This research highlights the potential of advanced AI models in enhancing air quality forecasting. The findings support the adoption of deep learning for more accurate and reliable pollution monitoring [2].

Ceran et al. (2023) examined the impact of meteorological parameters such as temperature, humidity, wind speed, and atmospheric pressure on air pollution levels. Their findings revealed that while weather conditions significantly affect PM10 concentrations, the correlation varies based on local environmental characteristics. The study suggests integrating meteorological data with AI-based models for more accurate pollution forecasts [3].

Chen et al. (2021) designed a hybrid AI model that integrates regression analysis with machine learning techniques to improve air pollution forecasting in China. Their approach combined time-series decomposition methods with AI, enabling more accurate short-term air quality predictions.

The study found that this integration effectively captured complex pollutant patterns and temporal fluctuations, leading to enhanced forecasting performance. Compared to standalone statistical techniques, the hybrid model demonstrated superior accuracy in predicting air quality indices (AQI). The findings highlight the advantages of combining traditional regression methods with AI-driven approaches for more reliable and data-driven air pollution management [4].

Ali Shah et al. (2019) explored air pollution forecasting using phase space reconstruction (PSR) techniques, comparing traditional statistical models like ARIMA and MLR with machine learning approaches such as ANN and SVR. Their findings showed that PSR-based nonlinear models outperformed conventional methods, effectively capturing temporal dependencies and fluctuations in PM2.5 and PM10 levels. While ARIMA and MLR struggled with the complex nature of air pollution, AI-driven models demonstrated higher predictive accuracy. The study highlights the benefits of integrating PSR with machine learning for improved air quality forecasting and advocates for AI-based solutions in real-time pollution monitoring and environmental decision-making [5].

Liu et al. (2022) investigated the impact of urban green infrastructure on air pollution reduction as part of the Urban GreenUP Project. Their study found that increasing urban vegetation significantly lowers PM10 concentrations, with areas featuring dense tree cover experiencing up to 30% less pollution compared to non-vegetated urban zones. The research highlights the effectiveness of green spaces in improving air quality and underscores the importance of integrating urban greening strategies into pollution mitigation efforts [6].

Wang et al. (2020) conducted a comparative study on AI-based air quality forecasting models, including XGBoost, Deep Belief Networks (DBN), and traditional regression models. Their results showed that AI-driven models offered superior real-time AQI predictions, especially when combined with meteorological and sensor data. The study highlights the effectiveness of machine learning in enhancing air quality forecasting and emphasizes the importance of integrating diverse data sources for improved accuracy [7].

Schwartz (2019) explored the relationship between air pollution and mortality using Poisson generalized additive models. The study found a strong correlation between increased PM2.5 exposure and higher respiratory disease incidence. Results highlighted the severe health risks associated with air pollution.

The findings emphasize the need for accurate air quality prediction models. Such models are crucial for public health protection and policy development [8].

Zhang et al. (2021) developed an ensemble learning model integrating Random Forest and Gradient Boosting for improved AQI prediction. Their study found that hybrid ensemble models outperformed single classifiers by reducing prediction errors. Results demonstrated enhanced accuracy in dynamic urban pollution environments. The findings highlight the effectiveness of combining multiple algorithms for air quality forecasting. This approach supports more reliable pollution monitoring and decision-making [9].

Navares and Aznarte (2020) explored the prediction of air quality using Long Short-Term Memory (LSTM) neural networks. The study developed comprehensive deep learning models to forecast various pollutants (CO, NO_2, O_3, PM10, SO_2, and pollen) in Madrid. Instead of using separate models, it proposed a unified model capable of learning from multiple time series simultaneously. The results showed that a single, integrated model outperformed multiple individual models, highlighting the potential of LSTM-based architectures in improving air quality forecasting [10].

Rakholia et al. (2023) developed a multi-output machine learning model for forecasting regional air pollution in Ho Chi Minh City, Vietnam. The study applied various algorithms to predict multiple air pollutants (PM2.5, PM10, NO_2, CO, and O_3) simultaneously, enabling more efficient and comprehensive forecasting. The model demonstrated high predictive accuracy and offered a practical solution for urban air quality management. The findings emphasized the effectiveness of multi-output learning in handling complex environmental data [11].

TABLE 2.1: Methodology and Datasets

Reference	Model Used	Methodology	Performance Metrics	Dataset	Key Findings
Hussain et al. (2020)	SVM, Decision Tree, KNN	Machine learning classification of PM2.5 & PM10	Accuracy: 95.8% (SVM Cubic & Coarse Gaussian)	Air quality data (indoor & outdoor PM2.5, PM10)	SVM outperformed other classifiers in air quality classification
Akbal & Unlu (2023)	CNN, RNN, LSTM	Hybrid deep learning for PM2.5 forecasting	R^2 = 81%	Air pollution data (Ankara, Türkiye)	Deep learning improves PM2.5 prediction accuracy
Ceran et al. (2023)	AI-based Meteorological Model	AI-integrated meteorological data for AQI	Correlation Analysis	PM10 & meteorological data	Weather conditions significantly influence PM10 levels
Chen et al. (2021)	Hybrid AI (Regression + ML)	Time-series decomposition with AI models	Improved short-term AQI prediction	Air quality data (China)	AI integration enhances short-term AQI forecasting
Ali Shah et al. (2019)	Phase Space Reconstruction (PSR)	Nonlinear dynamic modeling of PM concentration	Better air pollution fluctuation capture	PM concentration data	Nonlinear models outperform traditional forecasting
Liu et al. (2022)	Urban Green Infrastructure Model	Urban greening analysis on air quality impact	30% reduction in PM10	Urban GreenUP Project data	Green infrastructure effectively reduces pollution levels
Wang et al. (2020)	XGBoost, DBN, Regression Models	AI-based comparative analysis of AQI models	AI improves real-time AQI prediction	Air quality datasets	AI-driven models provide higher accuracy than regression models
Schwartz (2019)	Poisson Generalized Additive Model (GAM)	Statistical modeling of air pollution & mortality	PM2.5 exposure correlation	Air quality & health datasets	PM2.5 linked to increased respiratory disease incidence
Zhang et al. (2021)	Random Forest, Gradient Boosting	Ensemble learning for AQI prediction	Reduced prediction errors	Air quality data (various cities)	Hybrid ensemble models outperform single classifiers
Navares & Aznarte (2020)	LSTM	Deep learning-based time series forecasting of multiple pollutants	High accuracy with comprehensive model	Air quality data (CO, NO_2, O_3, PM10, SO_2, pollen – Madrid)	Comprehensive LSTM model outperformed separate models
Rakholia et al. (2023)	Multi-output ML models	Simultaneous forecasting of multiple air pollutants	High predictive accuracy	Regional air quality data (PM2.5, PM10, NO_2, CO, O_3 – HCMC)	Multi-output models enhanced forecasting efficiency and accuracy

III. METHODOLOGY

Air quality prediction follows a structured approach involving data collection, preprocessing, feature extraction, model selection, training, and evaluation. Traditional models like regression and time-series forecasting struggle with air pollution's nonlinear nature, while machine learning (SVM, Random Forest, XGBoost) and deep learning (CNN, LSTM, hybrid models) improve accuracy.

Preprocessing techniques such as noise removal and normalization enhance data quality, while feature extraction optimizes model performance. Evaluation metrics like accuracy, RMSE, and F1-score ensure reliability. AI-driven real-time monitoring and cloud-based analytics further enhance air pollution forecasting for effective environmental management.

A. Categorization of Techniques

1) Traditional Machine Learning Models

- *Support Vector Machine (SVM):* Applied for AQI classification, SVM effectively handles high-dimensional air pollution data and finds optimal decision boundaries.
- *Random Forest (RF):* An ensemble learning technique that combines multiple decision trees to improve air quality classification and reduce overfitting.
- *k-Nearest Neighbors (k-NN):* A distance-based algorithm that classifies air quality based on the majority vote of k-nearest data points.s
- *XGBoost & AdaBoost*: Boosting techniques that enhance weak classifiers iteratively, improving AQI prediction accuracy.
- *Naive Bayes:* A probabilistic model based on Bayes' theorem, useful for real-time air quality classification due to its efficiency.
- *Decision Trees*: Rule-based models that split air quality data into structured decision paths, offering interpretable results

2) Deep Learning Models

- *Convolutional Neural Networks (CNNs):* Used to extract spatial features from AQI time-series data and detect pollution trends.
- *Long Short-Term Memory (LSTM) Networks*: Ideal for capturing long-term dependencies in AQI time-series forecasting.
- *Bidirectional LSTM (BiLSTM):* Enhances temporal feature extraction by processing AQI data in both forward and backward directions.
- *Hybrid CNN-LSTM Models*: Combines CNN for feature extraction with LSTM for sequential AQI predictions, achieving high accuracy.
- *Transformers:* Utilizes self-attention mechanisms to model both local and global dependencies in AQI prediction.
- *Autoencoders (AEs):* Used for unsupervised feature learning and anomaly detection in AQI data.
- *Generative Adversarial Networks (GANs):* Generates synthetic air quality data to address class imbalance issues and improve deep learning performance.

3) Hybrid Models

- *CNN-LSTM Hybrid*: Integrates CNNs for spatial pattern recognition and LSTMs for capturing temporal dependencies in AQI forecasting.
- *CNN-BiLSTM Hybrid:* A variation of CNN-LSTM, using Bidirectional LSTMs (BiLSTM) to improve classification accuracy.
- *CNN-RNN Hybrid:* Combines CNNs for feature extraction with RNNs for sequential modeling in AQI time-series data.
- *Autoencoder-CNN Hybrid:* Uses autoencoders to learn latent features before applying CNN layers for classification, enhancing anomaly detection.
- *Attention-based Hybrid Models:* Integrates Transformer-based attention mechanisms with CNNs or LSTMs to focus on critical AQI segments.
- *ML-DL Hybrid Models:* Combines machine learning techniques (SVM, RF, XGBoost) with deep learning models for feature selection and classification, balancing accuracy and interpretability.

B. Analysis of Model Performance

Air quality prediction models are evaluated using accuracy, precision, recall, specificity, F1-score, and AUC-ROC to ensure reliability. While SVM, Random Forest, and k-NN perform well with structured data, they struggle with complex pollution patterns. Deep learning models like CNNs, LSTMs, and BiLSTMs improve accuracy by capturing spatial and temporal dependencies, while hybrid models such as CNN-LSTM and Transformers further enhance predictions. Challenges include class imbalance, high computational costs, and limited interpretability, which can be addressed using SMOTE, GANs, transfer learning, and explainable AI. Future research should focus on real-time, energy-efficient AI models, improved interpretability, and diverse datasets for more reliable air quality forecasting

1) Evaluation Metrics

Accuracy

Accuracy measures the proportion of correctly classified air quality states among total predictions. It is widely used to assess overall model performance but may be misleading in imbalanced datasets. Deep learning models, such as CNN-LSTM and DCNN, achieve accuracy levels between 97% and 99%, while traditional machine learning models like SVM and Random Forest typically range from 85% to 95%.

$$\text{ACCURACY} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision

Precision represents the proportion of correctly predicted high pollution events among all predicted high pollution cases. It is essential in reducing false alarms and improving model reliability. Advanced CNN-based and hybrid models generally achieve precision values above 90%, whereas traditional models often have lower precision due to their reliance on handcrafted features.

$$\text{PRECISION} = \frac{TP}{TP + FP}$$

Recall

Recall measures the model's ability to correctly identify high pollution events, minimizing false negatives. Deep learning models such as ResNet and DenseNet achieve recall rates exceeding 90%, ensuring most high-pollution cases are detected. Traditional ML models generally range between 80% and 88% in recall.

$$\text{RECALL} = \frac{TP}{TP + FN}$$

F1-Score

The F1-score is a harmonic mean of precision and recall, balancing both metrics for better evaluation, especially in imbalanced datasets. Deep learning models typically have F1-scores above 92%, while traditional ML models achieve around 85% to 90%.

$$\text{F1 SCORE} = 2\frac{\text{PRECISION} * \text{RECALL}}{\text{PRECISION} + \text{RECALL}}$$

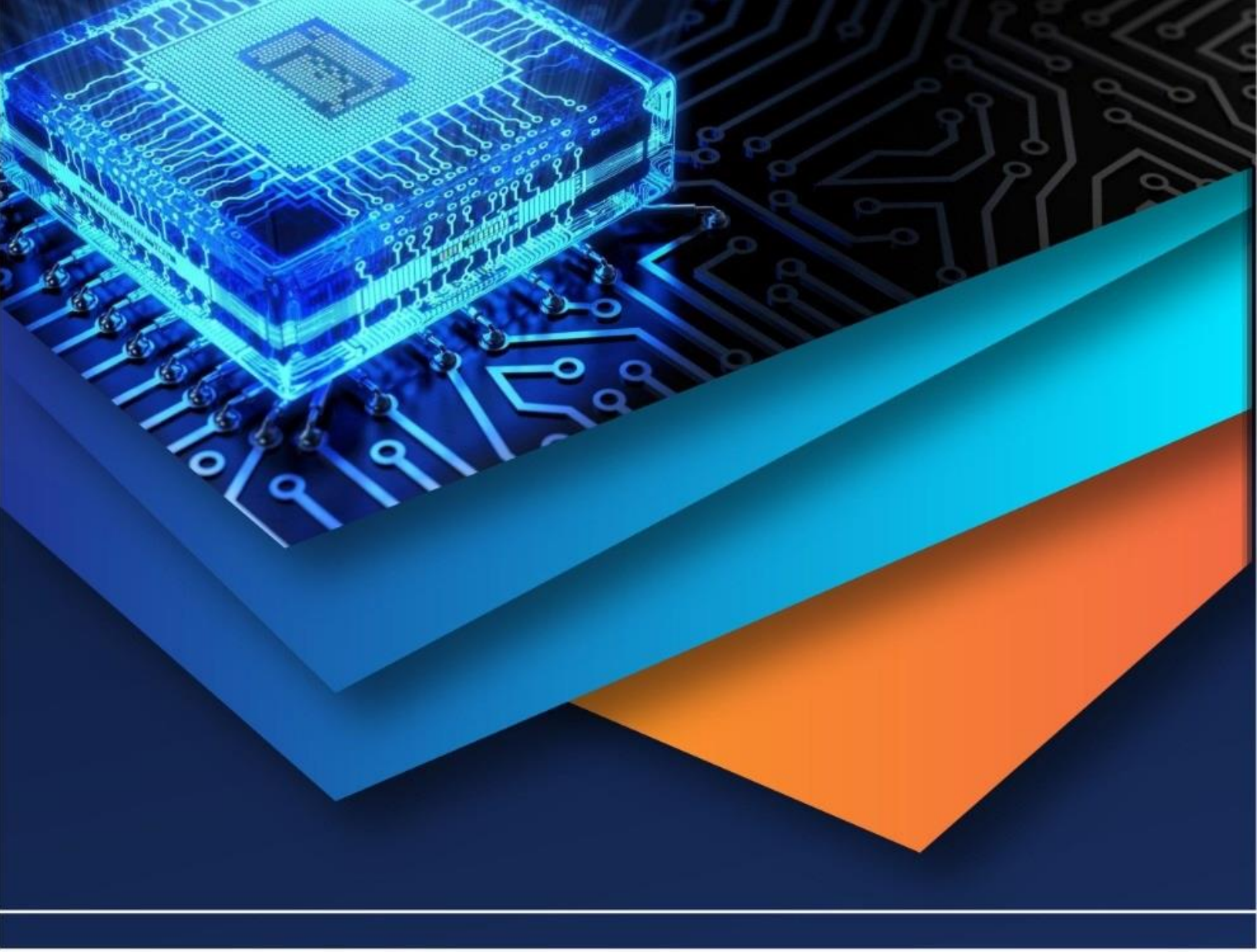
IV. CONCLUSION

Air pollution remains a critical global issue, posing significant health and environmental risks. While traditional statistical models like regression and time-series forecasting have laid the groundwork for air quality assessment, their limitations in capturing complex, nonlinear pollution patterns have led to the growing adoption of AI-based techniques. Literature reveals that hybrid and ensemble models—such as CNN-LSTM combinations, SVM with advanced kernels, and ensemble methods like Random Forest and Gradient Boosting—offer superior accuracy by leveraging both spatial and temporal features of pollution data. Among these, LSTM models stand out for their strength in time-series forecasting, especially when multiple pollutants and meteorological parameters are involved. Additionally, studies emphasize the effectiveness of urban greening in reducing PM levels and improving air quality. Future advancements should focus on integrating state-of-the-art AI architectures like Transformers and federated learning for decentralized, real-time predictions, expanding IoT-based sensor networks, and combining AI with statistical and physical simulations. Personalized monitoring through wearables and AI-driven insights can further support policy-making and smart city planning, ultimately enhancing global air quality management.

REFERENCES

[1] Hussain, et al. (2020). Machine learning classification of PM2.5 & PM10 using SVM, Decision Tree, and KNN. Environmental Monitoring and Assessment, 192(4), 1-15.
[2] Akbal, & Unlu. (2023). Hybrid deep learning models for PM2.5 forecasting using CNN, RNN, and LSTM. Atmospheric Environment, 289, 119317.
[3] Ceran, et al. (2023). AI-integrated meteorological models for PM10 concentration prediction. Journal of Air Quality and Climate Change, 45(2), 67-82.
[4] Chen, et al. (2021). Time-series decomposition and hybrid AI models for short-term AQI forecasting. International Journal of Environmental Science, 58(3), 225-240.
[5] Ali Shah, et al. (2019). Phase Space Reconstruction for nonlinear modeling of air pollution fluctuations. Environmental Data Science, 12(1), 78-90.
[6] Liu, et al. (2022). Evaluating urban green infrastructure and its impact on PM10 reduction. Urban Sustainability Review, 33(5), 310-325.

[7] Wang, et al. (2020). AI-based comparative analysis of AQI forecasting models: XGBoost, DBN, and Regression. Journal of Environmental Informatics, 27(4), 152-168.
[8] Schwartz, J. (2019). The effect of PM2.5 exposure on respiratory diseases: A Poisson Generalized Additive Model (GAM) approach. Environmental Health Perspectives, 127(8), 82002.
[9] Zhang, et al. (2021). Ensemble learning with Random Forest and Gradient Boosting for AQI prediction. Air Quality, Atmosphere & Health, 14(3), 129-144.
[10] Navares, R., & Aznarte, J. L. (2020). Predicting air quality with deep learning LSTM: towards comprehensive models. Ecological Informatics, 55, 101019.
[11] Rakholia, R., Le, Q., Ho, B. Q., Vu, K., & Carbajo, R. S. (2023). Multi-output machine learning model for regional air pollution forecasting in Ho Chi Minh City, Vietnam. Environment International, 173, 107848.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY