Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

Air Quality Index Prediction: A Review of Machine Learning, Deep Learning, and Hybrid Approaches

[1]Abhishek Tiwari, [2]Atesh Kumar
[1] MTech Scholar, Department of Computer Science & Engineering, Sanjeev Agrawal Global Education University, Bhopal Madhya Pradesh
Assistant professor, Department of Computer Science & Engineering, Sanjeev Agrawal Global Education University, Bhopal Madhya Pradesh

Abstract - The increasing evolution of air pollution as a global environmental and public health issue worldwide warrants the timely provision of accurate predictions for the Air Quality Index (AQI). Therefore, this review paper will explore in-depth all the traditional, machine learning, and deep learning models put into effect for AQI forecast modeling. Initially, some conventional statistical models like ARIMA and regression-based models are analyzed. It is seen that they are less complex and easy to interpret, but due to their incapability to handle nonlinear and complex data patterns, the results do not speak highly of these models. This raises issues around machine learning algorithms, such as those of decision tree, support vector machine, and ensemble methods, which actually increased the accuracy of data predictions thanks to better treatment of features and generalization. Further embraces on computational neural networks (CNN), long short term memory (LSTM), and hybrid models have been smartly improved to catch temporal and spatial linkages in air quality data. The paper also covers discussed up-coming trends such as federated learning, edge-AI, remote sensing integration, and explainable AI for much-needed larger deployment and more privacy and interpretable AI prediction options for the AQI. The key challenges identified in the work were data scarcity, model complexities, computational needs, and lack of evaluation frameworks complying with a set of standards. This review brings out these research gaps and clearly accentuates the need for these hybrid, scalable, and real-time AQI forecasting solutions. It, therefore, provides a comprehensive understanding of all available methodologies and helps foresee the construction of durable and efficient air quality predictions for the betterment of environmental management.

Keywords: Air Quality Index (AQI), Machine Learning, Deep Learning, Time Series Forecasting, Environmental Monitoring, Air Pollution Prediction

I. INTRODUCTION

Air pollution is one of the most notable challenges facing the environment into the 21st century, with a significant impact on human health, ecosystems, and climate systems. Rapid industrialization, urbanization, population growth, and increased vehicular emissions have given rise to large concentrations of atmospheric pollutants that are air-borne, including particulate matter (PM 2.5 and PM 10), nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (CO), and ozone (O3). These pollutants spoil the air quality and pose the serious risk of health disease, including respiratory diseases, cardiovascular system disorders, and premature mortality. Hence, air quality monitoring and prediction have become very essential for environmental sustainability and to save and protect the health of the public. The Air Quality Index (AQI) comprehensively simplifies the complexity of air pollution data into an obvious numerical value, thus making it easier for the general public and policymakers alike to understand air quality levels, thereby fortifying the linkage between air pollution risks and health risks [1]. In determining the classification of air quality as denoted in a range from very good to very hazardous, the AQI quantitatively deals with the levels other than the ones assigned to the standards by the WHO in terms of concentration of a pollutant. Accurate predictions of AQI have the potential to put in place early warning systems, support the decision-makers in controlling pollution, and enable immediate steps by individuals to prevent its occurrence. Thus, the development of superior forecasting algorithms for AQI, which could be very robust and effective, has become paramount in recent years. There are several standard statistical techniques that have been employed for predicting the AQI. These are like ARIMA, linear regression, and any other time series approach [2]. These models are naive and have fewer features of complexity. Large-scale or huge-dimensional RAQI prediction databases due to modern sensing technology quite challenge traditional models in many respects. Faster and more intelligent modeling techniques need to be enforced to enhance the accuracy of AQI prediction as the structure of air pollution is influenced by numerous factors like meteorological conditions, traffic pattern, emissions by industry, and geographical location. Artificial intelligence's development ushers machine-learning applications for the forecast of AQI, the usage of which has been rising; this includes decision trees, SVMs, KNNs, random forests, and more, all of which have given better results than traditionally accepted statistical approaches. The naming of these models implies that the techniques are skillful enough to learn the complicated intricacies of data and nonlinear relationships, effectively prediction models. Combining different models ensures increased accuracies, with ensemble learning reducing bias and variance [3]. These models, however, normally require extensive feature engineering and suffer from the problems of data overfitting and quality. In recent

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

years, deep learning (DL) techniques have emerged as powerful tools for AQI forecasting. The power of automatic feature extraction and modeling of complex temporal and spatial dependencies are the key driving mechanisms behind these growths. Architecture of neural networks, such as convolutional neural networks (CNN), recurrent neural networks (RNN), long short-term memory (LSTM), and gated recurrent units (GRU), have been extensively used for AQI prediction. Time-series data is particularly assayed as a measure of success in these models and the much-imported capture for long-term of gets of pollutant concentrations. Successful hybrid models which are comprised of distinct deep learning designs or which integrate more traditional machine learning and statistical techniques have shown superior forecasting performance [4]. In counterpoint, deep learning models, for all their success, generally require a high computational cost and a large amount of data for training, which is not always available. It is remarkable to observe that AI-aided and traditional approaches are being pursued for research in improving the air quality prediction paradigm for indoor as well as outdoor environments. To counterbalance the governed and outrageously fine-tuned electronic beast learning-based models, new technologies need to be used to develop engineering dogmas. Federated learning tries to bring about really state-of-the-art models for the purpose of training using availabilities or sandboxes of models across fused datasets instead of interacting. AI-like frameworks of technology, which operate only at the boundary of network history for Edge Computing and Edge-AI application like instance, dramatically trim latency from the raw data to the processed data from point to point and keep the results close by to the processing source [5]. Ground-based sensing limitations for point-reference in wide-open regions are apparent, much needed for measuring air quality during vast developmental process targets. Another If or a concise term is "Explainable AI." These concepts of effective deliverable very high variability can flow lucidly from developing model interpretability and understandable complex models, thereby providing increased understanding of air quality causation. Significant deficits in AQI forecast have been observed since many challenges still persist. Concerns over data quality and availability remain unaddressed with missing values, noise, or discrepancies in any of these data sets. Also, data heterogeneity---involving the integration of datasets from various sources like sensor information, meteorological data, and satellite imaging---renders more complexity to developing the models. To compare with other models, the difficulty is also caused by the absence of standard metrics for benchmarking models. Costs involving computation and scalability inside deep learning and hybrid models are both important [6]. A majority of the earlier studies restrict their scope to particular regions, which makes their generalizability tough for any different geography. The purpose of this review is to present an exhaustive analysis of the predictive tools for air quality index prediction, which covers standard statistical methods, machine learning models, deep learning models, and emerging artificial intelligence-based framework. Strengths and limitations thereof are discussed, overview of recent developments is given, and the paper attempts to identify key research gaps. By compiling results of various studies, this paper provides meaningful insights into more developed and robust forecasting methods for the AQI. Ultimately, AQI forecasting systems may be well-designed to support air quality monitoring systems that are intelligent and help sustain environmental management and public health.
Figure 1.1 represents the Air Quality Index (AQI) as a standardized scale used to measure and communicate the level of air pollution and its potential health impact on the public.

Figure 1. 1: Air Quality Index (AQI)

II. BACKGROUND AND SIGNIFICANCE OF AQI PREDICTION

Air pollution is now an important global issue, owing to the grave impact it inflicts on human health, environmental sustainability, and economic development. The rapidly expanding processes of industrialization, urbanization, transport, and energy production are pumping, in considerably increased proportions, the harmful pollutants in the air. The increased concentration of particulate matter (PM_{10} and $PM_{2.5}$), nitrogen oxides (NO_x), sulfur dioxide (SO_2), carbon monoxide (CO), ground-level ozone (O_3), etc., eventually lead to air quality deterioration. Pollutions like these can be traced back to numerous sources: vehicular emissions, industrial operations, fossil fuel burning, constructions, and natural occurrences - some of which include wildfires and dust storms []. The increased levels in concentration of these pollutants ultimately bring the issues of environmental pollution and public health to the fore, thus opening for new areas of research into monitoring and forecasting of air quality. The Air Quality Index (AQI) is a standardized indicator used to signify the overall air quality in a specific region. It converts the data concerning complex pollutant concentration into a unique numerical value, thereby making it understandable for policymakers, researchers, and the general

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

population to gauge air quality. The AQI-related levels are mostly divided into categories like good, moderate, satisfactory, poor, very poor, and hazardous, each corresponding to a specific health implication. If the AQI value is high, poor air quality has resulted and there may be health risks, particularly for vulnerable groups such as infants, senior citizens, and individuals who are already suffering from respiratory or cardiac illnesses. Therefore, it is said that AQI plays a significant role in simplifying air quality translation and raising public awareness to develop preventive measures. The understanding of the background of the AQI prediction lies in the domain of the evolution of air quality monitoring systems. Land-based monitoring stations have traditionally been used for air quality assessment, equipped with sensors that measure pollutants' concentration. Even though it gives precise findings, the requirement of resources is a significant obstacle in terms of cross-station and maintenance that becomes too expensive, causing severe impairment of spatial consistence [8]. Due to these constraints, there are some geomorphological regions and developing countries that still lack significant monitoring infrastructure. Researchers now often seek alternatives to good data sourcing, particularly remote sensing from satellites, sensor networks with low-cost sensor, open/crowd-sourced data, to address these constraints. This has led to an influx many times in 'big data' such that there is now considerable volume of real-time data converge at various scales of resource, necessary to envisage good AQI models. The significance of predicting AQI early is the ability to address potential air pollution events. Accurate predictions allow the authorities to timely respond, like closing cities off to traffic, regulating industrial emissions and advising the public [9]. Indeed, when air pollution is anticipated to be highest, the government shall opt to take preventive steps to reduce human exposure and thwart health risks. Policy planning can incorporate decisions regarding long-term environmental policy, such as how successful pollution control strategies are in achieving the targets, or in forwarding air quality as a key variable for the formulation of sustainable urban development plans. Understanding where AQI insights into climate change might come, one could point toward the tight relationship between atmospheric processes and pollutant-related greenhouse gas emissions.

From a public health perspective, the AQI prediction is highly indispensable to reduce the detrimental effects relative to pollution. Various health adversities, including asthma, chronic obstructive pulmonary disease (COPD), lung cancer, and heart-related issues have been linked with exposure to high pollution episodes. Short-term exposure further leads to actions like irritation in the eyes, nose, and throat-and over the long-term, it causes severe problems, like reducing life expectancy. To know about real-time and forecasted AQI values will prepare the people to choose their outdoor activities more reasonably, decide on all the protective measures, and make the healthcare planning grassroots-effective. This mainly targets extremely crowded urban areas since the pollution level there gets fluctuated in any moment [10] Advanced technologies are being employed in smog forecasting systems, resulting in markedly superior forecasting capacities. Machine learning/deep learning techniques now stood out as the critical factors behind the successful building of models that establish complex relationships between air pollutants and related meteorological or environmental fields like certain weather scenarios, traffic, and rules. These models can work through vast volumes of data, exposing complex relationships that just would not fall into view under normal conditions. Additionally, the time and location collection of data has become a possibility due to the development of IoT boxes thus making model predictions more reliable and hence more significant. All these technologies, combined on the same platform, led ASTute to form the AI in air quality management that's flexible with respect to environmental leadership. Broad awareness, still facing challenges, is needed in predicting AQI [11]. The quintessential concern happens on the availability and quality of data. Environmental datasets often carry missing values, noise, or inconsistencies, which ultimately wreak havoc on model performance. Moreover, data source heterogeneity (sensor data, meteorological information, satellite imagery) contributes to data integration and analysis complexities. Other reasons are data typology and the varying sensitivities of air pollution, which may change quickly under a variety of changing factors. In view of these, some solid-state adaptable and real-time solutions need to be pursued with simultaneous leading accuracy model designs. AQI forecasting significance also extends to economic and policy domains [12]. The severity of poor air quality has economic consequences; severe health care costs, reduced labor productivity, and diminished tourism engender economic benefits. Accurate AQI forecasting helps increase the probability of mitigating all these compensatory forces through more timely distribution of resources and quality of policy-making. Governments and environmental agencies can use predictive models to control it through some of the most suitable regulations, monitor compliance, and review the success of pollution control measures. Industries, on the other hand, require a mechanism for their operations that strategically lowers emissions lessens ahead of further abuse of the phenomenon by sustainable development [13].
AQI prediction holds an indispensable place in modern environmental management. It has come to bear any value in assessing air quality trends, enhancing public health, as well as training and informing decision-makers, from family level or company level to government level. The context of AQI forecasting is indicative of the progress from conventional compliance with somewhat uncertain air monitoring methods to AI-flavored cutting-edge methods through massive data and robust computational technologies. Despite strong headways achieved in this regard, there are challenges and opportunities to make further improvements in the accuracy, scalability, and interpretability of the AQI forecast models. By sharing techniques in AQI forecasting, future researchers and government bodies may join forces and tackle all the air pollution problems ensuring a decent environment around us.

III. DATA SOURCES AND PREPROCESSING TECHNIQUES

Data Sources

1. Ground-Based Monitoring Stations

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

Ground-based air quality monitoring stations are the primary source of accurate pollutant data, measuring parameters such as $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, and O_3. These stations are typically maintained by government agencies and environmental organizations. They provide high-resolution temporal data essential for AQI prediction. However, their spatial coverage is limited due to high installation and maintenance costs. Despite this limitation, they serve as benchmark datasets for validating predictive models. Their reliability and precision make them indispensable for training and testing machine learning and deep learning models in AQI forecasting.

2. Satellite-Based Remote Sensing Data

Satellite data provides large-scale atmospheric information, enabling air quality monitoring over wide geographic regions. Remote sensing techniques capture parameters such as aerosol optical depth (AOD), which correlates with particulate matter concentrations. This approach is especially useful in regions lacking ground-based monitoring infrastructure. Satellite data enhances spatial coverage and supports global AQI prediction. However, it may have lower temporal resolution and can be affected by weather conditions such as cloud cover. Integrating satellite data with ground-based observations improves prediction accuracy and enables more comprehensive environmental analysis.

3. Meteorological Data Sources

Meteorological parameters such as temperature, humidity, wind speed, wind direction, and atmospheric pressure significantly influence air pollutant dispersion and concentration. These data are collected from weather stations and meteorological departments. Incorporating meteorological data into AQI prediction models improves their ability to capture environmental dynamics. These parameters help explain seasonal and temporal variations in air quality. However, inconsistencies and missing values in weather data can affect model performance. Proper integration of meteorological data with pollutant data enhances forecasting accuracy and provides a more holistic understanding of air quality patterns.

4. IoT-Based Sensor Networks

Internet of Things (IoT) devices and low-cost sensors are increasingly used for real-time air quality monitoring. These sensors are deployed in urban areas to collect continuous environmental data. They provide high spatial and temporal resolution, making them suitable for smart city applications. IoT-based systems are cost-effective compared to traditional monitoring stations. However, they may suffer from calibration issues and data noise. Despite these challenges, IoT sensors enable scalable and distributed data collection. Their integration with cloud and edge computing platforms supports real-time AQI prediction and environmental monitoring [14]-[15].

5. Public and Open Data Repositories

Public datasets from platforms such as government portals, research institutions, and environmental agencies provide historical air quality data. These datasets often include pollutant concentrations, meteorological parameters, and geographic information. They are widely used for research and model development. Open data repositories facilitate reproducibility and benchmarking of AQI prediction models. However, data quality and consistency may vary across sources. Proper validation and preprocessing are required before use. These datasets play a crucial role in training machine learning and deep learning models for AQI forecasting.

Data Preprocessing Techniques

1. Data Cleaning and Missing Value Handling

Data cleaning is a crucial preprocessing step that involves removing noise, errors, and inconsistencies from datasets. Missing values are common in environmental data due to sensor failures or transmission issues. Techniques such as mean imputation, interpolation, and regression-based methods are used to fill missing values. Proper handling of missing data ensures the reliability of prediction models. If not addressed, missing values can lead to biased results and reduced accuracy. Data cleaning improves dataset quality and prepares it for further analysis and modeling in AQI prediction systems.

2. Feature Selection and Extraction

Feature selection involves identifying the most relevant variables that influence AQI prediction. Techniques such as correlation analysis, mutual information, and principal component analysis (PCA) are commonly used. Feature extraction helps reduce dimensionality and improve model efficiency. Selecting appropriate features enhances model performance and reduces overfitting. It also simplifies model interpretation. By focusing on important variables, feature selection ensures that the model captures key patterns in the data. This step is essential for building accurate and efficient AQI prediction models.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

3. Data Normalization and Scaling

Normalization and scaling are used to transform data into a consistent range, ensuring that all features contribute equally to the model. Techniques such as min-max scaling and z-score normalization are commonly applied. These methods are particularly important for machine learning and deep learning models, which are sensitive to feature magnitudes. Proper scaling improves convergence during model training and enhances performance. It also prevents bias towards features with larger values. Data normalization ensures stability and efficiency in AQI prediction models.

4. Time-Series Transformation

AQI data is inherently time-dependent, making time-series transformation essential for prediction models. Techniques such as lag features, rolling averages, and seasonal decomposition are used to capture temporal patterns. These transformations help models understand trends, seasonality, and periodic variations in air quality data. Time-series preprocessing improves the ability of models to forecast future AQI values. It is particularly important for recurrent neural networks and other temporal models. Proper handling of time-series data enhances prediction accuracy and model reliability [16]-[17].

5. Outlier Detection and Removal

Outliers are abnormal data points that can significantly affect model performance. In AQI datasets, outliers may occur due to sensor errors or extreme environmental conditions. Techniques such as z-score, interquartile range (IQR), and clustering-based methods are used to detect outliers. Removing or correcting these values improves model accuracy and robustness. However, care must be taken to avoid removing genuine extreme pollution events. Proper outlier handling ensures that models learn meaningful patterns from the data and produce reliable AQI predictions.

IV. TRADITIONAL STATISTICAL METHODS FOR AQI PREDICTION

Modern deep learning methods have had a huge influence on the performance of the AQI models. These extremely powerful methods and algorithms have accelerated the speed of research in this area. Traditional statistical methods have been the go-to solutions for air quality development though, and they all played him. A greater part of the literature on the analysis of air pollution data and the forecasting of future trends is predominantly reliant on time-series ARIMA, exp ARIMA, GARCH, and other mathematical formulations, under assumptions about data distribution. These methods are simpler, interpretable, and computational [18]. However, they are limited in some ways, such as assumption adherence, error autocorrelation, model stability, and high predictive heterogeneity. Despite the limitations, they are still anchor models, and, thanks to their inherent interpretability, they perform sophisticated duties as essential ingredients of hybrid solutions. ARIMA model is one of the most prominent techniques to forecast AQI. It is a time series forecasting method that understands linear relationships in the sequential data, its components being autoregression (AR), difference (I) and moving average (MA). The ARIMA model is rather efficient in modeling stationary time-series data and has been applied broadly for predicting the concentration of pollutants, namely $PM_{2.5}$, PM_{10}, among others. ARIMA, however, adheres to linearity in that it fails to provide for the representation of nonlinear relationships, which always exist in the setting of multiple interacting factors in an environmental dataset. Another method that is often employed is linear regression, which builds a model representing the relationship between air quality index and meteorological parameters from which it draws influence. The extension or advancement of this is marker multiple regressions, which incorporates several independent variables to attain a better level of prediction. The good thing with regression models is that they are easily interpreted and give insight into the effects of different factors on air quality [19]. Conversely, on balance, they can be limited to cope with complex interactions and nonlinear dependencies among variables. Which would bring down their predictive efficiency in extremely dynamic and nonlinear situations. Seasonal models like SARIMA have been dreamt up to deal with seasonal variability in air quality data. Air pollutants are frequently subjected to seasonal variation because of weather changes, human activities, and natural phenomena. SARIMA is splashed with seasonal features, extending ARIMA corresponding to cyclic trend modeling. These models help in the long-term forecasts and seasonal-oriented trend analysis. Nevertheless, they are still based on a linear belief and may not fit well with sudden or atypical patterns in data. Exponential smoothing methods, including Holt linear and Holt-Winters seasonal methods, are commonly used to predict AQI. Such techniques, which attribute the exponentiated weights to past information from an exponentially decreasing range of chronological records, have been developed to lend heavy consideration to the most recent information obtained with the help of empirical data. The essence of an exponential smoothing model is simply implementation with nice short-period forecasting properties. The method always shows superior performance when the time series shows an upward or downward trend and when seasonality is incorporated. Ultimately, however, one cannot expect almost any traditional method to work well within more complex non-linear circumstances, interaction effects, or multivariate relationship sampling. Time series decomposition methods are important statistical methods used for predicting AQI [20]. These methods break down a time series into components, such as the trend, seasonality, and residuals. By analyzing all of the components individually, researchers can gain a much better understanding of the underlying pattern in air quality data. Decomposition techniques are often used alongside other models to get higher accuracy of a forecast, although they may not prove much effective entirely when considered for prediction. It is natural for

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

one to find certain deficiencies in the traditional statistical methodologies when one attempts to predict AQI values. One of the challenging factors is that nonlinear and complex relationships are difficult for most of the existing models. Air pollution is controlled by a series of mechanisms like meteorological conditions, traffic density, industrial emissions, and geographical features. This nonlinearity and dynamics further lend complexity to the models for prediction under lean spatial scenarios. In conclusion, simplest statistical models often do not aptly approximate these relationships. In addition, assuming stationarity and normality are often devoid of the truth for much real environmental data [21].

Another significant limitation found in the big data contexts is the problem of t interventions involving large and high-dimensional datasets. The technological advances of the modern age, including sensory sensors and monitoring systems based on IoT, have fueled a large amount of environmental data to be continuously born. So that adapting traditional statistical models for effective treatment of big datasets is a nontrivial task. Not only because they lack automatic data registration in relevant features from the raw data but also, they require manual feature engineering, which of course could be tardy or error-laden. At the same time, traditional statistics are still very useful when predicting AQIs. Their simplicity and interpretability make them the right choice for determining some basic trends and relationships in data. Often they are used as a reference model to be compared from more sophisticated models. In addition, they can work with the machine learning/deep learning methodologies to create hybrid models that leverage the advantages of both methodologies. An instance is the combination of ARIMA, which can collect linear components and neural nets, relating to nonlinear patterns.

V. MACHINE LEARNING APPROACHES FOR AQI PREDICTION

The complexities within the datasets make them tough to predict since different parameters define varied dynamics. Dynamic and unorganized data may be rotely adjusted, and traditional prediction methods generate unsatisfactory performance. Given the poor performance of conventional statistical methods, machine learning (ML) approaches have gained prominence due to the greater flexibility in identifying the effects of variables that are significantly related to air pollution. Some of these factors include seasonality, connectivity, air patterns, and industrial influence. An MLP model tries to clearly and effectively model the structure of the AQI. Among the several machine learning techniques used for AQI forecasting, the Decision Tree algorithm is highly effective. It creates a tree-like structure to split the data based on feature values, allowing for easy interpretation and implementation. However, decision trees can be inclined to over-fitting in situations like with large datasets. Ensemble methods like Random Forest and Gradient Boosting have been put forth as solutions for overcoming this barrier. Random Forest integrates many decision trees to produce a more accurate prediction and reduce variance, whereas Gradient Boosting upsurges the minimizing of errors by sequentially improving weak learners. Renewable ensemble approaches show good performance in AQI forecasting. Another important category of ML techniques is Support Vector Machine (SWM), which is effective in dealing with high-dimensional and nonlinear relationships using strategically defined kernel functions. SVM has shown success in classifying AQI levels and predicting the concentrations of pollutants.[22] However, this method becomes cumbersome as the dataset grows, especially on parameter tuning issues and efficient means of scaling. Equally, K-Nearest Neighbors (KNN) is simple yet effective. In this algorithm, AQI is predicted on the basis of the similarity of data points. KNN is easy to implement, but it has higher computational overhead for large datasets and is prone to noise. Feature engineering is critical to the accurate AQI prediction using ML. Techniques to use for identifying the most relevant variables that affect air quality include feature selection, dimensionality reduction, and transformation. Common features include pollutant concentration levels, temperature, humidity, wind speed, and seasonal indicators. Proper feature selection has been known to improve model performance and reduce computational complexity [23]. It is important to verify the generalization of the model and ensure that no overfitting occurs using the cross-validation techniques. Even when they have diverse advantages, there exist some challenges while applying machine learning models in AQI forecasting. It is well known that data quality problems, namely missing values, noise, and inconsistent data, can degrade the abilities of a predictive model. In particular, machine learning models require extensive preprocessing and feature engineering. Moreover, there is a certain extent to which interpretability is a barrier when it comes to institutions of machine learning techniques compared with traditional statistical methods. This becomes more difficult, particularly when dealing with a complex ensemble model. Nonetheless, machine learning models maintain a powerful balance between accuracy and computational efficiency. In conclusion, machine learning approaches have become, in large part, the most popular method in AQI predictions due to the ability to model the intricate nonlinear relationships, considering that several high-quality datasets are provided [24]. Apart from SVM, ensemble methods, and a host of others that are somehow superior compared to traditional means, a considerable prediction power has successfully been obtained in regard to various mathematical problems pertaining to AQI. If the data is properly processed and appropriately pre-engineered, ML would be a perfect solution in delivering reliable and accurate AQI predictions for effective air-quality management and decision-making.

VI. DEEP LEARNING AND HYBRID MODELS FOR AQI FORECASTING

The deep learning (DL) ones thus explicitly feature in the best predictors for AQI. In the presence of learning and capturing intricacies in the myriad domains and temporal and spatial relationships, they involve units that inflict automatically upon themselves the imagination of hierarchical features. In contrast to conventional MA, these exceptional characteristics mean that a minimal level of manual feature engineering is now needed; irrespective of scaling up the dataset size etc., holding wide-eyed in a wild manner which generally depicts deep learning models' supreme brutal honesty with large-scale, high-dimensional

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

environmental datasets. One of the most widely used DL architectures for AQI prediction is the LSTM, which is a current neural network (RNN) with advanced variants like the Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) [25]. These models have been created to cope with sequential data and bring out long-term dependencies in time-series data. LSTM achieved fantastic forecasting of pollutant concentrations mainly due to the allowance to retain the historical information for long periods. GRU found applications in similar opportunities, but with reduced complexity over computation. AQI prediction utilizes Convolutional Neural Networks (CNNs), focusing on spatial features extracts from environmental data. CNNs can work on grid-based data like satellite images and the spatial distributions of pollutants. When configured into RNN models, CNNs enhance the ability to capture both spatial and temporal patterns, resulting in better forecast accuracy. Such hybrid models have specifically become the trend in recent literature. CNN-LSTM hybrid models are abundantly used in recent literature. Hybrid models that combine multiple techniques have gained significant attention in AQI forecasting. For example, models like ARIMA-LSTM combine the best features of both statistical and deep learning approaches. Similarly, combining machine learning and deep learning algorithms like Random Forest with LSTM provide better generalization and robustness [26]. Hybrid learning addresses the weakness of individual models and overall prediction performance improves. Recent innovations in deep learning consist of configuring data-driven precision with which models can attend to features and time steps. Technology examples from federated learning and Edge-AI models are almost real-time over AQI forecasting; here, data privacy remains uncompromised with a minimal latency in the edge, yet are mostly ahead for environmental monitoring system solutions to large-scale smart city applications.

Moreover, deep learning models indeed have many limitations. They require detailed data sets, and their implementation depends upon high computational cost. Training a deep neural network can be very time-consuming and can require specialized hardware like GPUs. To the complete lack of interpretability in DL, the use of deep learning models for prediction purposes instilled an additional layer obfuscating reasoning. To sum up, deep learning models and hybrid models provide a great advancement in predicting AQI. They are known to handle complicated patterns and are useful in handling large data sets, both of which guarantee accurate forecasting. Combining the strength of different methods, hybrid approaches enhance performance of the models [2]. Future research should focus much on the improvement in model efficiency, interpretability, and scalability, putting together a fast, trustable AQI prediction production system for sustainable environmental management.

VII. EMERGING TRENDS (FEDERATED LEARNING, EDGE-AI, XAI, REMOTE SENSING)

The emerging trends within AQI forecasting are driven by the most recent developments that have unraveled in artificial intelligence and data-driven technologies. These trends are found to be highly valuable because they focus on positively addressing major constraints associated with traditional and deep learning models, such as the issue regarding data privacy, complexity of calculations, the lack of interpretability, and restricted spatial coverage. This trend is seen very much in real time hence the strong likelihood that AQI forecasting on a much larger scale will change this kind of situation. What has drawn much attention among all of these is federated learning, edge AI (edge-cloud computing), XAI Solutions, Remote Sensing, and so on, which only boost scalability, cost-effectiveness, and data quality assurance for AQI forecasting. Federated Learning (FL) has become a recent notion for collective model training without data sharing. In classical machine learning systems, from several sources converged data has been used for model training, which raised issues on data personal information and security. These concerns are minimized by federated learning, which allows distributed learning components, meaning that models are trained on an individual device or institution, where only model updates are sent to a centralized server. This shines a major spotlight on applications like the prediction of AQI, concerning where information could be collated from many cities, organizations, or sensor networks with privacy restrictions. FL is used to keep confidential the data while using diverse datasets across the trainings to improve model generalization. But to ensure success in future implementations, it needs to address challenges like communication overhead, model synchronization, and heterogeneity of local data distributions. The existence of Edge Artificial Intelligence as a major trend that led toward real-time prediction about AQI is one based considerably more than simply centralized organizations: it leads AI channels for this information directly at its source. With the immense evolution of IoT-based air quality sensors composed of large datasets comes in an intermittent manner. Now consider the transmission of such sensory data to the cloud, which would therefore entail latencies and immense limitations regarding bandwidth. This problem is solved with Edge AI in deploying lightweight AI models on edge devices such as sensors, smartphones, or embedded systems. This allows quicker decision-making, reducing delay times, and the ability of the entire system to be more reactive. Edge-AI becomes particularly beneficial for smart city applications where real-time monitoring and issuing of alerts are imperative. Nevertheless, scarce computational resources and energy constraints make it difficult to involve large models in devices at the edge.

CHALLENGES, LIMITATIONS, AND RESEARCH GAPS

This table 1 summarizes key issues such as data quality, model complexity, scalability, interpretability, and real-time processing, while highlighting existing limitations and identifying future research directions for improving AQI prediction systems.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

Table 1: Challenges, Limitations, and Research Gaps in AQI Prediction

Category	Challenge / Issue	Description	Impact on AQI Prediction	Existing Methods Used	Limitations of Existing Methods	Research Gap	Future Direction
Data Quality	Missing and Noisy Data	Environmental datasets often contain missing values, noise, and inconsistencies due to sensor faults	Reduces model accuracy and reliability	Imputation, interpolation, filtering	May introduce bias and reduce data authenticity	Robust data cleaning frameworks lacking	Develop adaptive and intelligent data imputation methods
Data Availability	Sparse Monitoring Stations	Limited number of ground stations in many regions	Poor spatial coverage and inaccurate predictions	Data interpolation, satellite integration	Low accuracy in rural/remote areas	Need for scalable data collection methods	Integration of IoT and satellite data
Data Heterogeneity	Multi-source Data Integration	Combining sensor, meteorological, and satellite data	Increased model complexity	Data fusion techniques	Difficult alignment and synchronization	Lack of standardized fusion frameworks	Develop unified multi-modal data models
Model Complexity	High Computational Cost	Deep learning models require significant resources	Limits real-time implementation	GPU-based training, cloud computing	Expensive and energy-intensive	Efficient lightweight models needed	Develop energy-efficient AI models
Model Interpretability	Black-box Nature of Models	DL and ensemble models lack transparency	Reduces trust and usability	XAI techniques (SHAP, LIME)	Partial explanations only	Need for fully interpretable models	Develop transparent AI frameworks
Generalization	Region-Specific Models	Models trained on one region may not work elsewhere	Poor scalability	Transfer learning	Limited cross-region adaptability	Need for generalized models	Develop domain adaptation techniques
Temporal Dynamics	Changing Environmental Conditions	Pollution patterns vary over time	Reduced prediction stability	Time-series models (LSTM, ARIMA)	Struggle with abrupt changes	Adaptive real-time models lacking	Develop dynamic and self-learning models
Feature Engineering	Manual Feature Selection	Requires domain expertise and effort	Time-consuming and error-prone	PCA, correlation analysis	May miss important features	Automated feature extraction needed	Use deep learning-based feature learning
Real-Time Processing	Latency Issues	Delay in processing large-scale data	Ineffective early warning systems	Edge computing	Limited device capability	Need for real-time efficient models	Develop Edge-AI optimized models
Scalability	Handling Big Data	Large datasets from IoT and sensors	Slower model performance	Distributed computing	Complex implementation	Need scalable architectures	Use cloud-edge hybrid systems

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

Evaluation Metrics	Lack of Standardization	Different studies use different metrics	Difficult to compare models	RMSE, MAE, R^2	No unified benchmark	Standard evaluation framework missing	Develop benchmark datasets and metrics
Data Privacy	Sensitive Data Sharing	Data sharing across organizations is restricted	Limits model training	Federated learning	Communication overhead	Efficient privacy-preserving methods needed	Improve FL efficiency and security
Overfitting	Model Overfitting	Models perform well on training but poorly on unseen data	Reduced generalization	Regularization, cross-validation	Not always effective	Robust generalization techniques needed	Develop hybrid and ensemble approaches
Extreme Events	Handling Pollution Spikes	Sudden pollution events are hard to predict	Poor emergency response	Time-series models	Cannot capture rare events well	Lack of extreme event modeling	Use anomaly detection and hybrid models
Resource Constraints	Limited Hardware	Edge devices have limited memory and power	Restricts model deployment	Model compression	Accuracy trade-offs	Need efficient lightweight models	Develop optimized deep learning models
Spatial Variability	Geographic Differences	Pollution varies across locations	Reduced prediction accuracy	Spatial models (CNN)	Limited spatial generalization	Need better spatial modeling	Use geospatial AI models
Integration Issues	Lack of Unified Systems	Different technologies are not well integrated	Inefficient AQI systems	Hybrid models	Complex implementation	Need integrated frameworks	Develop end-to-end AQI systems

VIII. CONCLUSION AND FUTURE DIRECTIONS

This review highlights that while traditional models such as ARIMA and regression provide simplicity and interpretability, they are limited in capturing complex nonlinear relationships. Machine learning methods improve prediction accuracy by learning patterns from data, whereas deep learning and hybrid models further enhance performance by modeling temporal and spatial dependencies. Additionally, emerging technologies such as federated learning, Edge-AI, explainable AI, and remote sensing are transforming AQI forecasting by addressing challenges related to data privacy, real-time processing, and large-scale monitoring. Despite these advancements, several challenges persist, including data quality issues, limited spatial coverage, high computational requirements, and lack of model interpretability. These limitations indicate the need for more robust, scalable, and efficient AQI prediction systems. Furthermore, the integration of heterogeneous data sources and the development of standardized evaluation frameworks remain open research problems. Future research should focus on developing lightweight and energy-efficient models suitable for real-time applications, especially in smart city environments.

REFERENCES

[1] Sarkar, Nairita, et al. "Deep Learning and Federated Learning in Air Quality Forecasting: Trends, Insights, Challenges, and Future Perspectives." *Archives of Computational Methods in Engineering* (2026): 1-40.
[2] LAKSHMI, UPPADA, and A. Naga Raju. "Intelligent Air Quality Index Prediction System Using Machine Learning and Deep Learning Techniques." *International Journal of Data Science and IoT Management System* 5.2 (2026): 1618-1630.
[3] Dutta, Anandi, Kazi Sifatul Islam, and Damian Valles. "Edge-AI Framework for Air Quality Index (AQI) Forecasting with Quantized Deep Learning Models." *2026 IEEE 16th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2026.
[4] Singh, Pooja, and Sapna Nagpal. "Hybrid RF-LSTM: An Advance Machine Learning Technique for Prediction of Air Pollution in Indian Cities." *CLEI Electronic Journal* 29.2 (2026): 1-1.
[5] Somboonmark, Witchaya, and Julara Chumnaul. "Enhancing AQI forecasting accuracy: integrating ARIMA, ANN, and regression techniques with the development of HM4AQI web application." *Journal of Big Data* (2026).
[6] Kumari, Sheetal, et al. "Next-generation air quality management: unveiling advanced techniques for monitoring and controlling pollution." *Aerosol Science and Engineering* 10.1 (2026): 5-26.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

[7] Kumar, Vijendra, et al. "Advancing air quality prediction with hyperparameter optimization and innovative feature analysis using deep learning models in Phoenix, Arizona, USA." *Theoretical and Applied Climatology* 157.1 (2026): 60.
[8] Çelik, Mehmet Ali, Melahat Batu Ağırkaya, and Dessalegn Obsi Gemeda. "Bridging data scarcity with AI methodology: Global research trends in machine and deep learning for air quality prediction." *Journal of Atmospheric and Solar-Terrestrial Physics* (2026): 106777.
[9] Alam, Nayema Sanzida, et al. "Modeling and Predicting Air Quality Index of Bangladesh and India Using Machine Learning Approach." *The Nexus Between Geography and Sustainability: Exploring Emerging Tools and Techniques*: 455.
[10] Bhatt, Divyanshu, and Shikha Goswami. "Air Quality Prediction and Forecasting Using Machine Learning Algorithms: A Review." *DMPedia Lecture Notes in Multidisciplinary Research* (2026): 1095-1108.
[11] Mohandas, Prajul, P. Subramanian, and R. Surendran. "AI-Driven Hybrid Deep Learning Approach for Predicting Air Quality in India Using AlexNet-TabNet Architecture." 2026 International Conference on Computing Technologies & Data Communication (ICCTDC). IEEE, 2025.
[12] Susilawati, Tuti, and Bustomi Bustomi. "Air Quality Index Prediction Using Machine Learning Algorithms on the Beijing PM2. 5 Dataset." *Technema: Journal of Intelligent Engineering and Computing* 1.1 (2026): 20-29.
[13] Madan, Tanisha, et al. "Hybrid deep learning model for air quality prediction and its impact on healthcare." *Scientific Reports* (2026).
[14] El Mghouchi, Youness, and Mihaela Tinca Udristioiu. "Integrated AI Framework for Sustainable Environmental Management: Multivariate Air Pollution Interpretation and Prediction Using Ensemble and Deep Learning Models." *Sustainability* 18.3 (2026): 1457.
[15] Muralikrishnan, R., et al. "Machine learning-driven prediction and interpretation of air quality index in industrial environment." *Asian Journal of Civil Engineering* 27.3 (2026): 1473-1491.
[16] Singh, Sukhendra, et al. "Ensemble learning for air quality index prediction: integrating gradient boosting, XGBoost, and stacking with SHAP-based interpretability." *Scientific Reports* (2026).
[17] Goswami, Mausumi. "Air quality index improvement through machine learning and quantum computing: a framework for advancing air quality prediction using quantum-inspired metaheuristics on climate change to achieve positive health." *Multilevel Quantum Metaheuristics*. Academic Press, 2026. 341-384.
[18] Ullah, Abaid, et al. "Real time air quality monitoring and forecasting using AI and nature-based recommender system for climate-resilient air quality management." *Fusion Journal of Engineering and Sciences* (2026).
[19] Charjan, Ojas, et al. "A feature engineering–driven ensemble approach for accurate AQI forecasting." *Discover Applied Sciences* (2026).
[20] Khan, Darakhshan, Archana B. Patankar, and Jyotika Kakar. "Optimizing hourly air quality index forecasting: a particle swarm optimization-enhanced hybrid approach combining convolutional and recurrent neural networks." *International Journal of Electrical & Computer Engineering (2088-8708)* 16.1 (2026).
[21] Arifuzzaman, Md, et al. "Cost-effective and scalable urban air quality monitoring using image-based deep learning." Scientific Reports (2026).
[22] Goutham, M. R., et al. "Time-series based predictive modeling and explainable artificial intelligence for air quality index forecasting towards sustainable environment in rajahmundry, AP, India." (2026).
[23] Tawabini, Bassam. "Using machine learning algorithms to study the relationship between meteorological conditions and air quality parameters." Scientific Reports (2026).
[24] Edeh, Chijioke George, et al. "A Machine Learning-Based Framework for Real-Time Environmental Health Risk Prediction and Spatial Air Quality Intelligence in Urban-Industrial Ecosystems." Asian Journal of Advanced Research and Reports 20.4 (2026): 1-20.
[25] Gupta, Deepak, et al. "Remote Sensing-Based Air Quality and Atmospheric Pollution Modeling Using AI." AI and Remote Sensing for Earth Sciences. IGI Global Scientific Publishing, 2026. 121-156.
[26] Kohila, C., K. Meena Alias Jeyanthi, and P. Kasthuri Rengan. "Predicting Air Quality using a Hybrid Deep Learning Model to achieve Environmental Sustainability." Water, Air, & Soil Pollution 237.5 (2026): 267.
[27] Kumar, Sunny, et al. "Ensemble Machine Learning for Reliable Air Pollution Prediction and Sustainable Environmental Management." International Journal of Scientific Research in Science and Technology 13.1 (2026): 53-67.
[28] Haruna, Syaku Uba, and Wiwied Virgiyanti. "Deep Learning–Based Prediction of Urban Air Quality Using Multisource Environmental Data." Journal of Computer Applications and Information Technology (2026).
[29] Iqbal, Asif, and Nandini Mukherjee. "A Systematic Review and Comparative Study of Machine Learning Techniques for Air Quality Prediction: A. Iqbal, N. Mukherjee." Water, Air, & Soil Pollution 236.12 (2025): 789.
[30] Das, Aritra, and Himanshu Gupta. "Air quality prediction model for monitoring aqi." EPJ Web of Conferences. Vol. 328. EDP Sciences, 2025.