# A Comprehensive Review of Machine Learning and Deep Learning Approaches for Air Quality Index Prediction

**[1]Abhishek Prasad, [2]Ayushi Mehra**

[1,2] Department of Computer science and Engineering, Rabindranath Tagore University, Raisen, (MP) India

Email id- prasadabhishek048@gmail.com ayushi.mehra@rntu.ac.in

**Abstract:** Air pollutants are one of the major environmental and public health challenges worldwide stemming from rapid industrialization, urbanization, and increasing vehicles use. The Air Quality Index (AQI) is a widely used measure for evaluating the levels of pollution in the environmental and its impact on human and the environment. Over the last decades, real-time air-quality forecast methods have been the most rated essence for pollution monitoring, safeguarding public health, and generating policies useful for environmental planning. This review article gives an in-depth analysis of the current statistical, machine-learning, and deep-learning techniques used in the prediction of AQI. The review reveals that relatively older forecasting methods, such as Linear Regression, Multiple Linear Regression, ARIMA, and Exponential Smoothing, coupled with their advantages and limitations when faced with time-series pollution data. The focus shifts to some advanced machine-learning techniques such as Random Forest, Support Vector Machines, Gradient Boosting, XGBoost, etc., which are known to explore nonlinear relationships between pollutants and meteorological parameters. Furthermore, traction has been seen in the technological growth of deep learning architectures, such as Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and hybrid CNN-LSTM to focus on improved forecast of AQI. Comparison of earlier studies indicated that ensemble and deep learning approaches generally outperformed traditional statistical methods in terms of prediction accuracy and robustness. However, several impediments including data quality issues, computational complexity, and lack of realtime deployment still remain. The study concludes by outlining open research areas and future data for developing efficient, interpretable and real-time AQI forecasting systems in sustainable environmental management.

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

## I. Introduction

One of the environmental pollutants that impacts human health, causes instability in climate, and, at the same time, imposes threat to the ecological sustainability worldwide is the high levels of air pollution. Rapid industrialization, urbanization, population increases, and increased transportation activity are major causes of the increasing concentrations of pollutants released into the atmosphere. Pollutants such as particulate matter (PM2.5 and PM10), sulfur dioxide ($SO_2$), nitrogen dioxide ($NO_2$), carbon monoxide (CO), ozone ($O_3$), and volatile organic compounds (VOCs) are continuously increasing in both developed and developing countries; these pollutants lower air quality and lead to health conditions such as asthma, respiratory infections, cardiovascular diseases, lung cancer, and pre-mature deaths. Training documents suggest that millions of people are regularly exposed to air pollution levels that exceed safe levels. Air pollution also contributes to global warming, climate change, acid rain, decreased agricultural productivity, and environmental degradation. Urban areas, industrial areas, and population-heavy areas are particularly vulnerable to deteriorating air quality conditions due to high emission rates from vehicles, factories, construction activities, and burning of fossil fuels [1]. The Air Quality Index (AQI) has been developed as the universal yardstick for evaluating air pollution. It expresses overall air quality in a simple and easy-to-interpret manner by essentially converting complex pollutant concentration data into a numerical scale that ranges from the fulfilling the specified requirements of the air quality condition, up to that at which immediate action is required. Various AQI systems from different countries and environmental agencies seek to generate public awareness regarding the air quality being enjoyed at the time and the potential health repercussions associated. The categories of each AQI level are general: good, satisfactory, moderate, poor, very poor, severe, and worse. AQI thus runs between less severe and higher health-risk levels, with somewhat lower AQI representing clean air and lower health risk compared to the highest health risk, where pollution may seriously impair human health. The instruments that continuously monitor AQI, namely AQI monitoring systems, provide governments, environmental organizations, and public health departments with automatic aids in decision making for the execution of air pollution control strategies and timely issuing of health advisories. In current environmental management systems, viewing AQI as a fundamental indicator has come as a precondition of its success in understanding variations in air pollution and laying the very foundation to support elements of sustainable urban planning [2].

The growing seriousness of air pollution implies that accurate AQI-forecasting systems for pollution assessment have a strong need for more advanced application modeling capabilities to predict also the future pollution levels. The traditional air quality monitoring systems are only used mainly for polluted data collection, and they do fall quite short of making much futuristic pollution-causing situations go successfully. The accurate forecasting of AQI, which peaks at preventive interventions before hazardous pollution events can occur, enables planning and discretion by both government offices and the public [3]. Regulation of industry emissions, maintenance of safe corridors for transportation, and environmental campaigns aimed at enhancing some policies on environmental concerns are plausible recommendations based on forecasts of better air quality. Forecasting studies in low-air quality studied by public health institutions are a mechanism of reducing health hazards arising among the most vulnerable populations, including children, seniors, and those with respiratory conditions. AQI projections also lend considerable assistance to the development of smart cities, the promotion of sustainable environmental commitment, as well as environmentally voluminous in initiatives that require interaction with disasters [4]. Air quality prediction systems are of heightened significance, adding value to urban locations, where air pollution levels find continuous change with time by virtue of unstable weather conditions, industrial activities, moving vehicles, and other sources. Linear regression, Multiple Linear Regression (MLR), Autoregressive Integrated Moving Average (ARIMA), and Exponential Smoothing are commonly employed traditional methods in forecasting AQI. All these are based on historical records of different air pollutants, resulting in the prediction of future AQI values through relationships among several environmental variables, based on criteria that are mostly statistical or mathematical. Statistical methods are easier to comprehend; making them suitable for many short-term forecasting applications. Viewed from other perspectives, air pollution datasets would be seriously nonlinear, challenging and exacerbated by a range of environments such as meteorological conditions, seasonal variations, dense traffic, industrial pollution, etc. Thus normal statistical models do not really do a good job in providing a dynamic and nonlinear relationship to data related to air quality. This served as an eye-opener to researchers who would eventually look into avenue to see how to forecast AQI using more innovative computational techniques [5].

Machine Learning (ML) has become popular in many AQI prediction contexts for the complex patterns and relationships it can learn from the large environmental datasets. Machine learning models are another technology that has the potentials for working efficiently with multidimensional pollutant and meteorological variables in air quality predictions. Popularly used machine learning algorithms in predicting AQI include the Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Gradient Boosting, XGBoost, and ensemble learning techniques, transcending traditional statistical methods. The ensemble learning forms, such as Random Forest and XGBoost, put together diverse weak learners to raise the accuracy of forecasts, reduce overfitting, and beef up resilience. Even machine learning can deal effectively with missing data, noisy sensor data, and large environmental databases. Deep Learning (DL) approaches have furthered the transformation of atmospheric quality (AQ) predictors through the ability to perform automatic feature extractions and advanced temporal forecasting [6]. Such DL models comprise of Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks and have shown commendable performance in air quality forecasting tasks. Long Short-Term Memory (LSTM) networks are particularly suitable for AQ forecasting because they can capture long-term temporal dependencies and sequential pollution patterns in time-series datasets. Hybrid models that incorporate the CNN and LSTM architectures provide better forecasting accuracy by bringing spatial feature extraction together with temporal sequence learning. Attention mechanisms, autoencoders, and transformer-based architectures are being explored for the more advanced AQI forecasting applications. These intelligent forecasting systems help in the real-time monitoring of air pollution, smart environmental-management systems, and public-health protection [7].

Despite the remarkable advancements in AQI forecasting, several challenges persist in the creation of sustainable and robust forecasting systems. The environmental datasets invariably have either missing values, varied sensor readings, seasonality, or alarming spikes of pollution incidents, all of which disable the prediction output. The advanced models in deep learning are always heavy on resources and demand high-quality datasets, as well as optimization of highly sensitive parameters; these models, however, are difficult to decipher and hence more intriguing to policymakers and various environmental agencies [8]. The real-time extension and integration adopting the advancements in the field of the Internet of Things (IoT) also remain areas where substantial research is happening. To this end, continuous intelligence in the form of a blend of accurate predictions adaptable to real-time variables, easy to interpret, and computationally efficient ways for the calculation of AQIs is inevitable for the sustainable management of an environment. There is growing necessity for identifying with machine learning, deep learning, IoT, and real-time monitoring technologies that makes the powerfully designed air quality forecasting systems operational more efficiently and effectively. These smart prediction models should allow users not only to keep track of environmental conditions but also assist in decision-making processes of pollution control, health care, and sustainable urban development. Moreover, as air pollution-associated environmental sustainability assimilates on a national level, the smart AQI prediction system is due to make greater contributions to human health and environmental quality [9].

## II. Fundamentals of AQI Prediction

The prediction of the Air Quality Index (AQI) is establishing itself as a crucial research subject within environmental science, public health, and intelligent computing, primarily due to the immense impacts of air pollution on society and ecological sustainability. AQI prediction is generally defined as a set of processes for predicting the air quality to be in existence given previous pollution data, atmospheric conditions, and computation-based models or tools. The primary aim of AQI prediction systems is to give forecasts about the degree of pollution in the atmosphere so that required actions are executed in advance to prevent potential perils related to health and other damages to the environment. The contemporary AQI modeling frameworks integrate environmental surveillance technologies with statistical analysis [10], machine-learning algorithms, and deep-learning methods to offer detailed and real-time estimates of atmospheric pollution conditions. Immensely aiding in determining the degree of air quality in sections of the urban or global environment, AQI is a standard measure derived from the resident pollutants in the air. Again, this typically combines PM10 and PM2.5, sulfur dioxide minerals, nitrogen dioxide, carbon monoxide, meth-ane, ammonia, and lead. The various environmental institutes convert the particulate concentration into a value of AQI. It is worth deciding where different relatives of critical study and clas-sification of AQI represent different thresholds [11]: poor, moderate, satisfactory, good, severe, and, as such, very severe. Where the attending AQI value goes down lower, the air might be even cleaner and healthier, while the rising AQI values appear under air pollution, which will affect the health of humans across the range. The AQI system is a great way to convey to the public and decisionmakers the severity of pollution in a simple and understandable format.

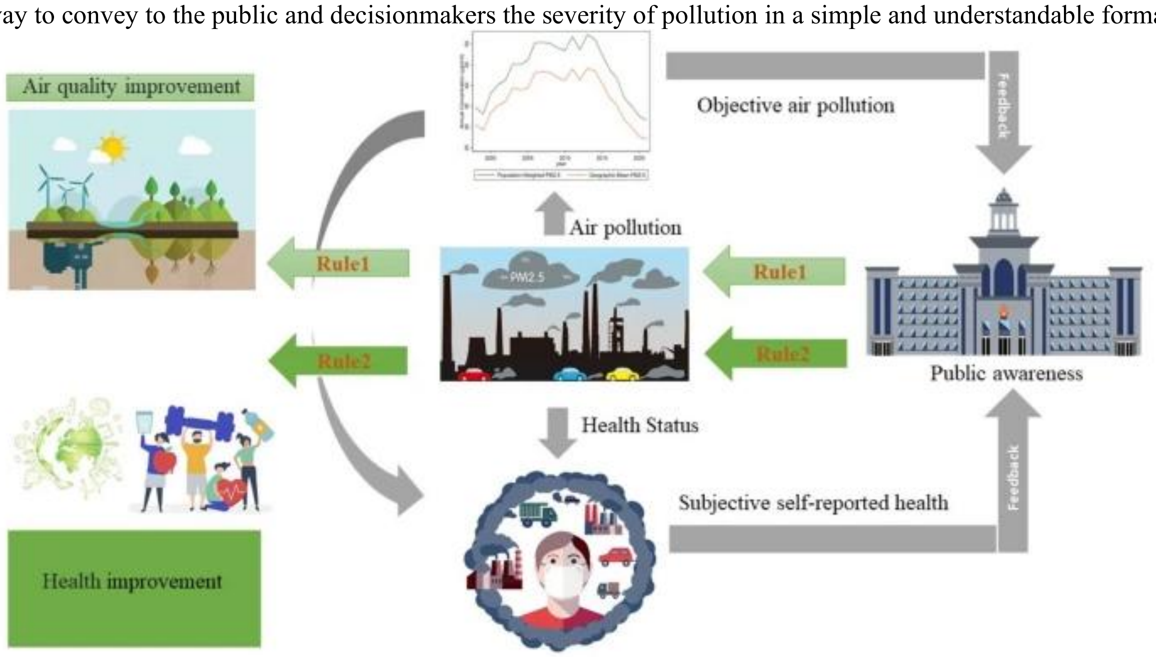


**Figure 1: Relationship Between Air Pollution, Public Awareness, Health Status, and Environmental Improvement [9]**

The figure 1 illustrates how public awareness and pollution control measures influence air quality improvement and human health conditions through feedback-based environmental management.

The AQI prediction system primarily depends on environmental datasets retrieved from monitoring stations, satellites, IoT sensors, and meteorological departments. The datasets provide pollutant concentrations and weather-related variables such as temperature, humidity, wind speed, precipitation, solar radiation, and atmospheric pressure. Meteorological conditions significantly influence the dispersal and accumulation of environmental pollutants in the atmosphere. High humidity and low winds, for instance, may lead to the accumulation of pollutants through stagnation, whereas rain reduces the accumulation of airborne particulate matter. Consequently, the amalgamation of meteorological-analyzed AQI predictive model performance is, therefore, a necessary one [12]. Data-preprocessing techniques such as missing data handling, normalization, outlier removal, and feature engineering have been adapted to further improve prediction accuracy for the AQI forecasting models. Former AQI forecasting strategies largely resort to statistical and time-series estimations. Linear regression and different multiple linear models are the oldest statistical tools used to project the AQI based on pollutant and environmental data. They are appealing for their simplicity and ease of computation and interpretation; however, they suppose linear relationships between variables, which in, by extension, does not permit learning complex structures of pollution. Time-series forecasting methods such as ARIMA, Seasonal ARIMA, and Exponential Smoothing are frequently employed for forecasting AQI because they examine historical AQI trends and behavior left from seasonal variations so as to predict future levels of pollution. While, indeed, statistical methods can ably project pollution levels over short time spans, the methods never depict any important criteria of cross-variables and/or sudden changes in the environment.

Over the years, to circumvent the drawbacks of the traditional statistical models, ML has been brought in by more researchers in the AQI prediction game. Those algorithms have been able to learn and form the complex nonlinear relationships between concentrations of different pollutants or/and meteorological factors from large datasets. AQI forecasting with ML has been found useful using Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Gradient Boosting, and Extreme Gradient Boosting (XGBoost) as the modeling approaches [13]. RF and XGBoost are among the popular methods when it comes to AQI forecasting with respect to prediction accuracy, generalization capability, and ability to fight overfitting. Not only can these algorithms work very well with high-dimensional environmental data but they are able also to interpret and pinacle the important features within the data influencing the air pollution. Ensemble learning methods have gotten the model improvements by combining multiple weak learners into a strong predictive model. Approaches based on Deep Learning (DL) are robustly considered because they naturally discover the hidden features given enough large-scale datasets and learn the time dependencies of complex nature. Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks are agents in the context of modeling air quality predictions. In readying the ANN models, they try to model and function after the biological neurons to gauge non-linear relationships among environmental variables. CNN processes use the spatial architecture in manifold pollution datasets, thus they are good for spatial feature learning, while the LSTM models are relevant when dealt towards limited-time forecasting sequence appurtenances [14]. Given this, LSTM networks have again been proven worthy to be useful in predicting AQI if they can retain wide-ranging time-wise data generation and steppollutant patterning guaranteed to be colored solidly. The hybrid models, on the other hand, would become a successful exemplar from CNN-LSTM configurations for their capability to learn spatial along with temporal features for air pollution prediction.

Finally, it is naive to appreciate that real-time air quality index forecasting systems are tied up with IoT and Edge Computing technologies which cater to the ultimate evaluation of the environment. The IoT-based air quality monitoring system incorporates low-cost sensors to produce real-time data representing various pollutants. These sensors upload environmental information to cloud platforms or edge devices for further computation and prediction. Edge-AI makes the whole prediction happen in real time locally and thus shrinks latency and overhead on communication. Smart cities utilize AQI forecasting systems for their traffic management, industrial emission control, and public health advisories. Mobile applications and web dashboards render real-time live updates on AQIs, facilitating alerts, which enables to spread awareness and environmental safety to the public [15]. It is imperative to know whether our AQI-forecasting system works properly or it is a misnomer! Statistically, several ways exist to judge the forecasting accuracy and modeling efficiency. Some examples of performance measures are Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), coefficient of determination ($R^2$), accuracy, precision, recall, and F1-score. Lower values of MAE and RMSE indicate better prediction, whereas the higher the $R^2$ value, the stronger the association of predicted AQI and actual AQI values. Comparative studies often use these performance metrics to talk different machine learning and deep learning models out of naming the most successful to forecast AQI under different environmental conditions.

Despite the significant advances accomplished regarding predictions of air quality, challenges remain. The data centering environmental concerns often includes anomalies, sensor readings, and varied results that drastically undermine model reliability. Pollution levels are always in flux. The dynamic nature of pollution patterns is governed by changes in the weather, traffic levels, and industrial events [16]. The computational power required for deep learning models, which in turn increases overall implementation costs, along with the challenges of handling large datasets and tuning complex hyperparameters obstructs efficient moving forward. Again, many advanced methods in air quality prediction do not allow insight visualization, meaning that the result could be known too well from the previous decisions that made this incertitude difficult for policymakers. Material hindrances to smart air quality monitoring systems include real-time deployment, scalability, and cross-regional adaptability.

## III. Machine Learning Models for AQI Prediction

A comparative assessment of Decision Tree, Random Forest, Support Vector Machine, and XGBoost was provided in [1]. This comparison illustrated which machine learning technique provides a superior forecast for air pollutant prediction. The results stated that ensemble learning methods performed better than the traditional techniques because they offered maximum efficiency in handling the nonlinear pollutant relationships. The study underscored the importance of appropriate feature selection and preprocessing techniques when undergoing the procedure of forecasting in improving its practice. Furthermore, it was found that the study was not widely applicable. The lack of real-time implementation and geographically diverse datasets, which are a portion of major weaknesses that limit the generalization ability of the suggested framework, must be addressed. A unique forecasting model for AQI combining ARIMA, Artificial Neural Networks, and regression techniques along with the HM4AQI web-based application was presented in [2]. The hybrid model, integrating both quantitative statistical-based time series analysis and neural network based-based learning capabilities, seemed to improve the forecasting capabilities of the proposed mechanism. One of the novel features of the web-based version consisted of allowing public access and visualization of pollution in real-time. In spite of the achievement of superior prediction tasks, the system demands heavy computational power and the parameters need a thorough standardization, thereby limiting its utilization in low-resource settings. The study designed and tested on a composite learning system to forecast the AQI with generalized concept of the boosted Gradient, XGBoost, and stack,kuja, making the model utilize SHAP-based intelligibility [3]. It attained good prediction accuracy while offering explainable results displaying noteworthy pollution and climatic influencers. SHAP analysis has enhanced the transparency and interpretability of our predictive models. The ensemble structure, however, built a high level of computational hurdle necessitating huge training datasets for its optimal performance.

Focus was laid on interpreting forecasting to improve the reliability of air pollution predictions with the target of capturing trends in air pollution and seasonal variations. The proposed framework performed well in predicting day-to-day pollution trends and seasonality. A weakness of the model may be its incapability to serve as a tool against extreme air pollution anomalies with respect to abrupt environmental volatility due to various climatic and industrial activities. Some machine learning algorithms, including Support Vector Machine, Random Forest, Artificial Neural Network, Decision Tree models, were evaluated for AQI forecasting. The conclusion was that more than one ensemble and neural-networks-based method scored over statistical alternatives in the areas of predicting performance and robustness. The study highlighted the power of machine learning in acquiring performance data for multidimensional pollution datasets; yet, a drawback was that it did not do real-time validation or discuss the efficiency pertaining to large machine-learning.

## IV. DEEP LEARNING MODELS FOR AQI PREDICTION

In [6], a multiscale quantile neural network for AQI prediction was suggested with mixed out-of-selling pollution data sets and meteorological variables. The present model captured essentially and approximately uncertainties and deterministic nonlinearity in the environmental data, thus warranting improved forecasting stability and forecast accuracy for developing a quantile prediction model. Also, reliability could be assured even in uncertain pollution conditions. But the major drawback is that the model requires very high computational power, massive training datasets were required, and its application is not feasible in resource-starved systems. In [7], an approach drove by feature engineering through ensemble learning for accurate AQI forecasting was introduced. Substantial extraction of advanced environmentally set of features driven the capability of ensemble models far beyond expectations for an earlier assumed predictive accuracy. Having shown an increase in forecasting accuracy and systems robustness as compared to standalone machine learning methods, this feature engineering process came with its cons and prepossessed great preprocessing complexly besides the necessity of domain expertise to undertake a sound feature selection. A labeling graph system for improved accuracy by fuzzy inference was proposed [8]. A fuzzy inference system improved the accuracy of the aggregate knowledge and dealt effectively with uncertain and imprecise environmental data. The model was highly capable in the nonlinear pollution pattern and complex AQI categories. In contrast, the model had virtually unacceptable computational efficiency and scalability limitations in real-time applications. An announcement on how to back prop-predict A-QI from time-series data looked in [9]. The developed system effectively memorized pollution trends over time and long-term dependencies in an environmental dataset with improved prediction accuracy over statistical methods. The existing model being computationally heavy and data-hungry must render it much less applicable for real-time applications. Another work on AQI quantization was carried out in [10] by adoption of the Edge-AI framework with non-tensor AI computations. The goal was to introduce deep learning within lightweight edge-device models, which could further be offered for real-time air quality estimates. With an energy reduction in consumer electronics (due to quantization techniques), edge-devices-computers did run with acceptable results and the transparency in compromise between performance of forecasts and solids (harsh) model simplifications and limited edge-processing capabilities was achieved. A hybrid deep learning model for air quality prediction and healthcare impact was suggested in [11]. Multiple branches of deep learning were introduced with the aim of boosting AQI forecasting accuracy and analyzing health-related pollution effects. One good gain from the model was its performance in determining pollution trends and linked healthcare risks. The hybrid system necessitated considerable data to demonstrate excellence in performance, and large numbers of data were used in optimization processes. [12] analyzed the seasonal dynamics and time-series forecasting of urban air quality within Andhra Pradesh using environmental datasets collected from urban cities. It was successful in extracting seasonal pollution patterns and long-term air quality variations by statistical and forecasting techniques. Results provide improved understanding of regional pollution behaviors and forecasting reliability. On the other side, the study was geographically restricted and failed to consider advanced machine learning strategies to improve forecasting. A multi-attention enhanced CNN-LSTM model for AQI forecasting is presented [13]. Integration of the attention mechanisms with the CNN-LSTM architecture considerably improved spatiotemporal feature learning from pollution data. The architecture performed well in AQI forecasting tasks as well as in analyzing the dependency of pollution components. Nonetheless, a hefty training time and computational overhead were created because of the complex architecture, making smooth implementation a bit difficult. A modular multi-layered LSTM framework for real-time AQI forecasting, designed for solving problems is introduced [14]. The framework was extremely good in catching long-term temporal dependencies in urban pollution datasets and provided accurate real-time forecasting in urban corridors in Rajasthan. The modular framework has better scalability and prediction efficiency. However, fast changing environmental conditions require continuous retraining for the model to adjust. In [15], Authors have presented the Genetic Algorithm-based improved Extreme Learning Machine model. The approach with the Genetic Algorithm optimized the model parameters and improved the accuracy of forecasting according to GT 2010. The approach does give us better time-to-convergence properties and stronger generalization capabilities compared to those attained from the conventional Extreme Learning Machine algorithms. A disadvantage is that it led to immensely increased processing times and the model became unnecessarily bulky. A two-stage XGBoost pipeline for environmental parameter and consequent AQI forecasts for smart indoor air quality monitoring systems were proposed in [16]. The model in question used advanced boosting techniques to predict indoor environmental conditions and AQI levels. This algorithm has high prediction accuracy, and it is compatible with smart monitoring applications. However, as indicated by the thrust of research being on the indoor environment, outdoor counterparts were tested on a considerably smaller scale. A multi-scale decomposition and learning model for forecasting complex cyclic data were presented in [17]. The model based on decomposition techniques aimed to fetch the influence of the multiscale temporal pattern and enhance predictive stability. The model also improved predictive performance compared to its benchmark conventional time-series methods. Yet it brought about higher implementation complexities and computational requirements. A notable-integration-framework was proposed in [18] for machine learning and time-series analysis for urban temperature dynamics across climate zones in Sri Lanka. As it comes to show, machine-learning techniques clearly outstrip traditional forecasting methods in capturing climate variations. The framework increases environmental prediction confidence across different climate regions. The model lacked, however, the imperative need to be altered and tuned for each specific climatic situation to make it work smoothly. Furthermore, a multimodel approach is explored towards downscaling and forecasting future trends of temperature and precipitation for Northern Nigeria [19]. This framework explained the contribution of various statistical regression and predictive postulation between suggested predictors in order to have a more refined approach towards climate forecasting in a world full of changing environmental conditions; that study improved further upon the existing model in a better viewpoint—long-range-environmental planning. The models exhibit inadequacy in decrypting extreme uncertainty and sudden fluctuations in the environment. An efficient Grey Wolf Optimized-Markov Grey model is proposed for forecasting carbon emission at an expressway during various maintenance periods pointedly made in [20]. The integration of Grey Wolf Optimisation and Markov models increases the anticlockwise of accuracy and efficiency of a forecast; recognition was made for improving performance with forecast in emission trends observed from varying maintenance compositions. The process is developed so advanced in the application of making real-time large-scale environmental systems less practicable, and it is adorned with many a complex procedure for optimization activities. In line with the consumption data and disease patterns seen in hospitals, a forecasting method for medicine requirements was introduced in [21]. Through predictive analytics and massive health data from past events, resource-based planning and inventory management capabilities were enhanced. The interlacing of systemic elements like seasonal fluctuations and historic dependencies resulted in improved forecasting performance coupled with a decrease in medicine shortages. This approach still relied on historical data consistency and had little in the way of adaptability for dealing with sudden healthcare emergencies. The prediction framework based on ARIMA featured in [22] addresses matters pertaining to the analysis of the trends involved in the mpox outbreak using epidemiological datasets. It provides the trend of the outbreak and dynamics across time for forecasting purposes, which is the essence of the framework. Building healthcare master plans and addressing capacities are among the potential areas where the proposed framework could be applied in. However, the ARIMA model lacks the ability to capture epidemic patterns that are more often non-linear and abrupt case fluctuations.

One multimodal predictive model for service workload prediction under ever fluctuating factors such as holidays and promotions was put forth by some earlier research works [23]. The inclusion of multiple prediction variables in the model seriously improved service load prediction accuracy. The ensuing results were well-reasoned in favor of planning improvement and better operational implementation. However, the shorthand for this strength ought to be the necessity for real-time data updates, aside from handling unexpected demand spikes. Prediction model was utilized for input, and statistical forecasting techniques were employed to analyze the scenarios of pollution trends and variations on AQI in periods. From the outcomes, it emerged that the act of implementing temporal models for environmental monitoring has now paid dividends. At the same time, the capability for inculcation of Ai and online predictions for pollution as to having accrued a small time-lag on their reliable prognostication effort. Version one presents the elaboration of a standardized PM2.5 concentration index for monitoring and forecasting air pollution characteristics [25]. It smoothened the air quality monitoring under comparable monitoring settings. The investigations led to a broader conceptual framework of understanding of the behavior of PM2.5 with regard to household or individual exposure and urgent need for environmental risk assessments. However, with a focus only on particulate pollution, other key atmospheric pollutants were overlooked.

The idea proposed in [26] of a Fourier Series-based forecasting approach was primarily designed to deal with AQI data for Kolkata in seasonal cycles among various seasonal cycles. The model did well in capturing the seasonal AQI periodic aspects of pollution concentration frequently as well as capturing AQI datasets repetitively, right through the season. The model reached improved performance for periodic forecasting tasks. However, an analysis of the model showed that the model performed less effectively when it dealt with irregular pollution conditions and abrupt environmental disturbances.

It was put forward in [27] that an advanced air quality forecasting framework based on smoothing method and white noise analysis be introduced for Indore city. The research sought accuracy improvements in seasonal AQI through the incorporation of smoothing with noise reduction approaches. Therefore, the framework had successfully discovered pollution trend on the long run and seasonal variation. But the model faced challenges due to non-linearity of pollutant interactions and dynamic environmental uncertainties. In [28], the use of a fully automated machine learning technique was highlighted for predicting hourly AQI in Azamgarh, India. The study showed that the hourly pollution variations can be handled quite effectively using machine learning models with high accuracy in forecasting. The suggested framework gives a very good answer to meeting real-time AQI monitoring capacities. Nevertheless, the system demands gigantic data preprocessing for the maintenance of its performance. A comparative analysis of machine learning algorithms for the prediction of AQI, the authors used various models including Random Forest as well as Support Vector Machine and Neural Network to predict air quality efficiently. Ensemble and deep learning models were found to outperform traditional approaches according to their research result. However, computational complexity and interpretability remained key challenges. Presents a study of the evaluation of machine learning technologies for the prediction of AQI. The construction of a variety of pollution-related features for understanding air pollution and future AQI behavior using video analytics will aid in developing mechanisms for predicting pollution and AQI. Such models provide better accuracy of forecasting and an improved awareness of environment of the system. However, the critical missing elements are interpretability and computational cost on a larger scale. Table 1 presents a comparative summary of previous AQI prediction studies, highlighting the techniques used, datasets considered, major findings, and limitations of each approach.

**Table 1: Comparative Study of AQI Prediction Techniques**

| Ref. | Technique Used | Dataset Used | Key Findings | Limitations |
|---|---|---|---|---|
| [1] | Decision Tree, Random Forest, SVM, XGBoost | AQI and pollutant datasets | Ensemble models achieved higher AQI prediction accuracy | Limited real-time implementation and regional diversity |
| [2] | ARIMA, ANN, Regression Techniques | Historical AQI and meteorological data | Hybrid model improved forecasting accuracy and web-based monitoring | High computational complexity and tuning requirements |
| [3] | Gradient Boosting, XGBoost, Stacking, SHAP | Environmental and AQI datasets | Ensemble learning with SHAP improved prediction and interpretability | Large training data and high computational cost required |
| [4] | Time-aware Machine Learning Pipeline | Daily AQI data from Indian cities | Improved daily AQI forecasting and temporal analysis | Difficulty handling sudden pollution spikes |
| [5] | ANN, Random Forest, SVM, Decision Tree | AQI and environmental datasets | Machine learning models outperformed traditional methods | Lack of real-time validation |
| [6] | Multi-scale Quantile Neural Network | Multiscale air pollution datasets | Better uncertainty handling and forecasting stability | High computational resource requirement |
| [7] | Feature Engineering and Ensemble Learning | AQI and meteorological datasets | Feature engineering improved forecasting accuracy | Complex preprocessing and feature selection |
| [8] | Graph-based Fuzzy Inference System | Environmental and AQI datasets | Improved handling of uncertain and nonlinear pollution data | Scalability issues for large datasets |
| [9] | Temporal Deep Learning Framework | Time-series AQI datasets | Captured temporal dependencies effectively | Requires large training datasets |
| [10] | Edge-AI and Quantized Deep Learning | Real-time sensor-based AQI data | Reduced computational cost for edge devices | Slight reduction in prediction accuracy |
| [11] | Hybrid Deep Learning Model | Air quality and healthcare datasets | Improved AQI prediction and health impact analysis | Complex architecture and optimization |
| [12] | Time-Series Forecasting Techniques | Urban AQI datasets from Andhra Pradesh | Identified seasonal air pollution patterns | Limited geographical scope |
| [13] | Multi-Attention CNN-LSTM | AQI time-series datasets | Enhanced spatial and temporal feature extraction | Increased training complexity |
| [14] | Multi-Layered LSTM Framework | Real-time AQI datasets from Rajasthan | Accurate real-time AQI forecasting | Requires continuous retraining |
| [15] | Genetic Algorithm-based Extreme Learning Machine | AQI and environmental datasets | Improved optimization and prediction accuracy | Higher computational time due to optimization |

## V. Emerging Trends and Future Directions

Considering the advancements in AI, IoT, cloud computing, and environmental sensing technologies, the field of AI-enabled systems for predicting pollution and quality of air is rapidly growing. Instead of traditional statistical methods, smart machine learning and deep learning models are being considered due to their ability to cope with environmental pollution datasets that are nonlinear and widely-scattered in nature. In AQI, the great demand is for the latest architectures abided by deep learning, for instance, LSTM, CNN, Transformer models, and CNN-LSTM hybrid frameworks. These frameworks are designed to enhance the performance of forecasting models by identifying temporal and spatial dependencies between air pollution data. Also, systems like attention mechanisms would facilitate better explainable AI and model trust [7].

The adoption of IoT-enabled smart air quality systems is another major trend. This comprises the use of low-cost sensors connected through IoT that carry around-the-clock real-time environmental observation data through urban and industrial sectors. Live and real-time monitoring of pollutants-such as PM2.5, PM10, $NO_2$, $SO_2$, CO, and $O_3$-help an accelerated and more precise AQI forecasting. The immersed power of Edge Computing and Edge-AI has both enhanced the real-time modelling by significantly holding back the real-time data processing on local hardware boxes as opposed to complete dependency on cloud infrastructures. This reduces latency, cuts communication overhead, decreases power consumption and minimizes communication time for pollution alerts [15]. The use of satellite image, remote sensing data, and Geographic Information Systems (GIS) is becoming more popular in AQI forecast applications. Satellite-based pollution monitoring helps to analyze regional and global pollution distribution fields, especially where very few ground-level air pollutant monitoring stations exist. Machine learning technology assists in improving the large-scale evaluation and prediction accuracy when compared to the vast conventional regression. Further, hybrid and ensemble learning methods are very popular in the stream as they tend to combine strengths of various forecasting models to better ensure robustness as well as prediction performance [16].

## VI. Conclusion

This work reviewed the atmospheric air AI methods of conventional statistical, machine learning algorithm, deep learning approach in AQI prediction. Furthermore, linear regression, ARIMA are some of the simple and easily implemented forecasts that can work with AQI prediction models, but they fail miserably when faced with nonlinear pollution tendencies. Machine learning methods of forecasting: Random Forest, Support Vector Machine, XGBoost, ensemble models provide a very good estimate for the pollutant's strength. New deep-learning approaches like CNN, LSTM-whatever one makes sense-tend to capitalize on temporal and spatial relationships in environmental data corpus for enhanced forecasting capability and long-term performance. The review also identified several challenges including data quality issues, computational complexity, lack of interpretability, and real-time deployment. However, the aforementioned challenges can be suggested to be dealt with using new opportunities from advanced technologies like Internet of Things (IoT), Edge-AI, satellite monitoring, and interpretable AI for intelligent systems predicting AQIs. In the context of future research development, lightweight, scalable, and interpretable forecasting frameworks are yet to be established to hold real-time environmental monitoring; therefore, the integration of AI-enabled AQI systems with smart cities, healthcare systems, and sustainable urban planning would serve to improve pollution, environmental protection, and climate change. The foreseen enhancement hybrid models combining both machine learning, deep learning, and IoT technologies would be pivotal in the future efficient air quality predictions.

## REFERENCES

[1] N. Srinivasa Gupta, Y. Mohta, K. Heda, R. Armaan, B. Valarmathi, and G. Arulkumaran, “Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis,” *Journal of Environmental and Public Health*, vol. 2023, Article ID 4916267, pp. 1–26, Jan. 2023, doi: 10.1155/2023/4916267.

[2] Somboonmark, Witchaya, and Jularat Chumnaul. "Enhancing AQI forecasting accuracy: integrating ARIMA, ANN, and regression techniques with the development of HM4AQI web application." *Journal of Big Data* (2026).

[3] Singh, Sukhendra, et al. "Ensemble learning for air quality index prediction: integrating gradient boosting, XGBoost, and stacking with SHAP-based interpretability." *Scientific Reports* (2026).

[4] Goetzinger, Philipp, Sebastian Noy, and Karsten Huffstadt. "Machine learning approaches for predicting air pollution levels: a transparent, time-aware pipeline for daily AQI in Indian cities." *Modeling Earth Systems and Environment* 12.2 (2026): 97.

[5] Mangla, Manya, Tejna Khosla, and Bhanu Prakash Reddy Rella. "Machine Learning Approaches for Air Quality Index (AQI) Prediction: A Comparative Study." *Transforming Healthcare With AI And IoT*. CRC Press, 2026. 38-53.

[6] Zhao, Shangrui, et al. "Multi-scale quantile neural network for air quality index prediction." *Theoretical and Applied Climatology* 157.4 (2026): 229.

[7] Charjan, Ojas, et al. "A feature engineering–driven ensemble approach for accurate AQI forecasting." *Discover Applied Sciences* (2026).

[8] Kulandhainadar Mariavalavan, Ordenshiya, and Revathi Govindasamy Krishnamoorthy. "Enhanced air quality index prediction using fuzzy center merge labeling graph based fuzzy inference system model." *Cybernetics and Systems* 57.3 (2026): 534-576.

[9] Sakunthala, Kovelakonda, et al. "A Temporal Deep Learning Framework for Air Quality Index Prediction using Time-Series Data." *2026 6th International Conference on Expert Clouds and Applications (ICOECA)*. IEEE, 2026.

[10] 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.

[11] Madan, Tanisha, et al. "Hybrid deep learning model for air quality prediction and its impact on healthcare." *Scientific Reports* (2026).

[12] Shreyas, D., et al. "Deciphering seasonal dynamics and time series forecasting of urban air quality: a case study in Urban Cities of Andhra Pradesh." *Scientific Reports* (2026).

[13] Xie, Tianbao, and Sijia Zhong. "Multi-Attention-Enhanced CNN-LSTM Model for Air Quality Index Forecasting." *Proceedings of the 2026 2nd International Conference on Big Data, Communication Technology and Computer Applications*. 2026.

[14] Gaur, Bhanu. "A Modular Multi-Layered LSTM Framework for Real-Time Air Quality Index Forecasting: A Case Study of Urban Corridors in Rajasthan (2023–2025)." *Int. Journal of Engineering, Science and Environment (IJESE)* 1.1 (2026).

[15] Lakshmi, Chaganti BN, et al. "Air Quality Index Forecasting via Genetic Algorithm-Based Improved Extreme Learning Machine." *International Journal of Data Science and IoT Management System* 5.2 (2026): 2098-2107.
[16] Jadhav, Deepali S., et al. "A Two-Stage XGBoost Pipeline for Environmental Parameter and AQI Forecasting in a Smart Indoor Air Quality Monitoring System." *Proceedings of the International Conference on Sustainable Innovation with Artificial Intelligence and Machine Learning 2025 (ICSIAIML 2025)*. Springer Nature, 2026.
[17] Wang, Jianping, et al. "Forecasting LNG Prices Using a Multi-Scale Decomposition Learning Model." *International Review of Economics & Finance* (2026): 105258.
[18] Rupasinghe, R. A. K. P., and Lakshika S. Nawarathna. "Urban Temperature Dynamics in Sri Lanka: A Comparative Machine Learning and Time-Series Analysis Across Wet, Intermediate, and Dry Climate Zones." (2026).
[19] Jibrin, Salim, and Jazuli Abdullahi. "Evaluating Future Temperature and Precipitation Trends in Northern Nigeria: A Multi-Model Approach to Downscaling and Forecasting." *Techno-computing Journal* 2.1 (2026): 50-75.
[20] Ren, Yuan, et al. "Research on carbon emission forecasting during expressway maintenance period based on GWO-Markov optimized grey model." *Scientific Reports* (2026).
[21] Kambey, Ronald Reidi, Vinnie Widayanti, and Kurnia Popy Rakhmawati. "METHOD OF FORECASTING MEDICINE REQUIREMENTS BASED ON CONSUMPTION DATA AND DISEASE PATTERNS IN HOSPITALS." *Integrative Perspectives of Social and Science Journal* 3.04 April (2026): 1805-1812.
[22] Daodu, Lanre Peter, et al. "Temporal dynamics and predictive modelling of mpox outbreak using autoregressive integrated moving average: insights from the Nigeria Centre for Disease Control and Prevention, 2017–2025 data." *IJID One Health* (2026): 100125.
[23] Zhang, Yajing, and Minhao Li. "Accuracy Evaluation of Multi-Factor Forecasting Methods for Customer Service Workload Prediction Under Holiday and Promotional Fluctuations: Evidence from US Service Industry Data." *Journal of Advanced Computing Systems* 6.5 (2026): 12-20.
[24] Howard, C. C. "Temporal analysis and prediction of air quality index for atmospheric pollutants in Port Harcourt, Rivers State." *World Scientific News* 204 (2025): 311-327.
[25] Batool Awan, Wajiha, and Zulfiqar Ali. "Development of standardized PM2. 5 concentration index (SPM2. 5I) for monitoring and forecasting air pollution characteristics." *Theoretical and Applied Climatology* 156.10 (2025): 504.
[26] Ghosh, Anirban, and Raju Maiti. "Forecasting of Multiple Seasonal Categorical Time Series Using Fourier Series with Application to AQI Data of Kolkata." *arXiv preprint arXiv:2603.03954* (2026).
[27] Jayant, Jitendra, Aseem Chandra Tiwari, and Angur Bala Jayant. "Integrating winter's method and white noise for advanced air quality forecasting: a case study of Indore, India." *Environmental Research Communications* 7.3 (2025): 035031.
[28] Ansari, Asif, and Abdur Rahman Quaff. "Advanced machine learning techniques for precise hourly air quality index (AQI) prediction in Azamgarh, India." *International Journal of Environmental Research* 19.1 (2025): 15.
[29] Desai, Tanay, et al. "Comparative analysis of machine learning algorithms for air quality index prediction." *Machine Learning for Computational Science and Engineering* 1.1 (2025): 14.
[30] Abdelmalek, Mostafa M., Hatem Mahmoud, and Hassan Shokry. "Prognosis of air quality index and air pollution using machine learning techniques." *Scientific Reports* 15.1 (2025): 25890.