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

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Air Quality Monitoring and AQI Forecasting Using Time-Series ML

G. Ashmitha, S. Snigdha Chandana, G. Asritha, G. Veera Venkata Aditya Sai, M. Harshini
Department of Computer Science and Engineering
Kakatiya Institute of Technology and Science, Warangal, India

***Abstract*—Air pollution has recently been identified as one of the most pressing issues in the environmental domain, which has severe implications for the health and well-being of citizens across the world. The increased rate of urbanization, industrialization, and vehicular emissions has contributed substantially to the deterioration of air quality, especially in major cities. The Air Quality Index (AQI) has been widely adopted as a standardized measure to reflect the level of air pollution and its associated health effects. The accurate estimation and forecasting of AQI has become increasingly important for pollution control and decision-making.**

A thorough machine learning-based framework for time-series model-based air quality monitoring and AQI forecasting is presented in this paper. For real-time AQI prediction based on pollutant concentrations like PM2.5, NO_2, NOx, SO_2, and CO, the suggested system incorporates a K-Nearest Neighbors (KNN) regression model. Additionally, by identifying seasonal patterns and temporal dependencies in past air quality data, a Seasonal AutoRegressive Integrated Moving Average with eXogenous variables (SARIMAX) model is used for short-term AQI forecasting. The Central Pollution Control Board (CPCB) provided the dataset used in this study, which has undergone extensive preprocessing, including time-series transformation, normalization, and handling of missing values.

The Flask framework is used to implement the suggested framework as a web-based application that allows users to interactively enter pollutant values and obtain multi-day forecasts and AQI predictions.

***Keywords*—Air Quality Index, Machine Learning, Time-Series Forecasting, KNN Regression, SARIMAX, Environmental Monitoring, CPCB Dataset.**

I. Introduction

In recent decades, air pollution has emerged as one of the most pressing environmental issues, especially in areas that are rapidly urbanizing and developing. The air quality in metropolitan areas has drastically declined due to the expansion of industrial activities, rising vehicle emissions, population density, and energy consumption. Numerous health problems, such as respiratory disorders, cardiovascular diseases, decreased lung function, and early mortality, have been directly linked to poor air quality. Global health reports state that prolonged exposure to polluted air causes millions of deaths each year, underscoring the critical need for effective air quality monitoring and forecasting systems.

As a standardized indicator of the general state of air pollution and its possible effects on human health, the Air Quality Index (AQI) is widely used. The concentrations of major pollutants, including particulate matter (PM2.5), nitrogen dioxide (NO_2), nitrogen oxides (NOx), sulfur dioxide (SO_2), and carbon monoxide (CO), are combined to calculate AQI values. AQI offers a simplified numerical representation in addition to qualitative categories like Good, Moderate, Poor, and Severe, each of which has a unique impact on health. Despite the fact that AQI is a useful tool for communicating air quality conditions, its values are highly variable and impacted by a variety of factors, such as emission sources, weather, seasonal variations, and geographic features.

The majority of conventional air quality monitoring systems rely on threshold-based approaches, statistical methods, and manual analysis. Although these methods offer fundamental insights into pollution levels, they frequently fall short of capturing intricate nonlinear relationships between pollutants. Furthermore, a lot of current systems are limited to reporting AQI values in real time and are unable to forecast future trends in air quality. This restriction hinders the prompt implementation of preventive measures to reduce health risks and diminishes the efficacy of early warning systems.

However, the challenge of forecasting air quality is addressed by time series forecasting models, which model trends, seasonality, and correlations that exist in air quality data. Seasonal factors like winter smog, dispersion during monsoons, and air pollution during festivals are important contributors to AQI variations. Hence, a more comprehensive and reliable solution for air quality analysis can be obtained by combining machine learning prediction models with time series forecasting models.

In the proposed research work, an integrated model is developed by incorporating K-Nearest Neighbors Regression for real-time AQI prediction and SARIMAX for short-term AQI forecasting. The proposed model is developed using real-world air quality data collected from the Central Pollution Control Board (CPCB) and is implemented as a web-based application using the Flask framework. The proposed model has the potential to enhance the accuracy of AQI prediction, identify seasonal trends of air pollution, and offer a useful and user-friendly tool for air quality monitoring.

A. Our Contributions

The major contributions of this research work are summarized as follows:

- A holistic air quality monitoring system is developed using real-world CPCB air quality data.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

- A K-Nearest Neighbors (KNN) regression algorithm is proposed to predict AQI values using pollutant concentration measures such as PM2.5, NO_2, NOx, SO_2, and CO.
- A Seasonal AutoRegressive Integrated Moving Average with eXogenous variables (SARIMAX) model is developed to predict short-term AQI forecasts using time-series and seasonal patterns.
- The prediction and forecasting models are combined into a single framework, allowing for both real-time AQI predictions and multi-day forecasts.
- The proposed framework is developed as a web application using the Flask framework, allowing for user interaction and result visualization.
- Experimental analysis is conducted to evaluate the efficacy and robustness of the proposed approach.

II. Related Work

Several studies have explored air quality monitoring and forecasting using machine learning and time-series approaches. Traditional air quality prediction methods relied on chemical transport models, which are computationally expensive and sensitive to input assumptions. Recent research has shifted toward data-driven techniques due to their adaptability and accuracy.

Özüpak et al. [1] performed a comprehensive comparison of machine learning regression models such as XGBoost, SVR, Random Forest, and KNN for air quality prediction. Their study demonstrated that Bayesian optimization combined with ensemble strategies significantly improves prediction accuracy. Gupta et al. [2] applied Random Forest and Support Vector Regression models for AQI prediction across Indian cities and reported improved performance when data balancing techniques were used.

Deep learning-based approaches have also gained popularity. Janarthanan et al. [3] employed deep neural networks to forecast AQI in metropolitan regions, highlighting the effectiveness of temporal modeling. Mao et al. [4] utilized LSTM-based architectures for air quality prediction and achieved superior performance compared to traditional regression models. Similarly, Rahman et al. [5] proposed a web-enabled machine learning system for real-time AQI forecasting.

Hybrid and optimization-based approaches have shown promising results. Lakshmipathy et al. [6] integrated feature selection with ensemble learning to enhance AQI prediction accuracy. Wang and Zhang [7] demonstrated the effectiveness of CNN-based models for high-resolution air quality estimation. These studies collectively indicate that combining machine learning with time-series modeling and optimization techniques yields reliable and accurate air quality predictions.

III. Problem Statement and Motivation

Rapid urbanization and industrialization have been some of the major factors that have contributed to the degradation of air quality in large cities. Prolonged exposure to high concentrations of air pollutants has serious implications for human health, including the development of respiratory and cardiovascular ailments. Although air quality monitoring stations are in place, many of these stations are designed to provide information only about the prevailing levels of air pollution and lack the ability to predict or forecast air quality.

Current methods for air quality prediction are either based on traditional statistical models or machine learning models. Statistical models have limitations when dealing with nonlinear relationships, whereas machine learning models often disregard the temporal and seasonal patterns that exist in air quality data. In addition, many of the existing methods are limited to offline analysis and are not designed as interactive real-time systems.

Another significant issue is the provision of actionable insights to the users. Most of the systems are not able to provide real-time AQI prediction and short-term forecasting together in one system. This makes it impossible for the users and the authorities to predict the future trends of air pollution and take necessary steps to prevent it.

With the above issues in mind, the objective of this project is to design an integrated air quality monitoring system that incorporates machine learning-based AQI prediction and time series forecasting. The proposed system will utilize KNN for real-time AQI prediction and SARIMAX for short-term forecasting.

IV. System Overview

The proposed Air Quality Monitoring and AQI Forecasting system is intended to be developed as an integrated approach that encompasses data preprocessing tasks, machine learning-based prediction, time series forecasting, and web-based implementation. The main aim of the proposed system is to facilitate real-time AQI prediction and short-term AQI forecasting. The overall architecture of the system is illustrated in Fig. 1.

The architecture of the system is divided into five main parts: data collection, data preprocessing, AQI prediction, AQI forecasting, and result visualization. All these parts are essential for ensuring the accuracy of the prediction, reliability of the forecast, and proper interaction with the end users.

The AQI prediction module consists of a K-Nearest Neighbors (KNN) regression model. This module uses the concentrations of the various pollutants as input variables to predict the AQI value based on past patterns of pollution. The choice of the KNN regression model is informed by its simplicity, interpretability, and ability to handle nonlinear relationships between pollutants and AQI.

The AQI forecasting module employs a Seasonal AutoRegressive Integrated Moving Average with eXogenous variables (SARIMAX) model. The AQI forecasting module examines past AQI data to identify patterns and seasonal changes. The SARIMAX model produces short-term AQI forecasts, which help in providing early warnings for possible pollution peaks.

Lastly, the result visualization and user interaction module is developed using a web application built with Flask. The

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

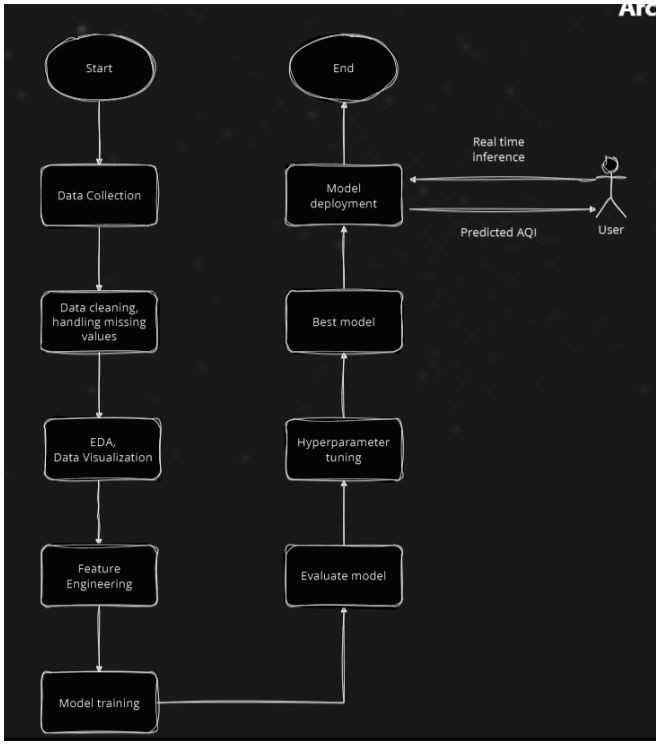

Fig. 1. Overall Architecture of the Proposed AQI Monitoring and Forecasting System

graphical user interface enables users to enter pollutant concentrations and display forecasted AQI values and short-term forecasts. The system design with multiple modules promotes scalability and will allow for easy expansion in the future, such as the integration of real-time sensors and mobile app development.

V. Dataset Description

The data used in this study is collected from the Central Pollution Control Board (CPCB), which is the government agency designated for monitoring the air quality in India. The CPCB supplies the continuous ambient air quality monitoring data collected from various air quality monitoring stations established in major cities. For this study, air quality data for Gurugram city is chosen because of its high pollution level and the availability of historical data.

The data is collected on a daily basis for the major air pollutants that contribute significantly to the calculation of the AQI. The attributes chosen for the study are particulate matter with diameters less than 2.5 micrometers (PM2.5), nitrogen dioxide (NO_2), nitrogen oxides (NOx), sulfur dioxide (SO_2), and carbon monoxide (CO). The dataset also includes the AQI value corresponding to the pollutant concentration, which is calculated as per the CPCB guidelines.

The concentration levels of the air pollutants in the dataset are measured in standard units like micrograms per cubic meter ($\mu g/m^3$) for PM2.5, NO_2, NOx, and SO_2, and milligrams per cubic meter (mg/m^3) for CO. AQI is measured in terms of numerical values that correspond to pre-defined air quality classes.

VI. Data Preprocessing

Air quality data sets obtained from actual monitoring stations may have noisy data, missing values, and inconsistencies due to sensor failure, environmental interference, or communication breakdowns. Hence, data preprocessing is an essential step to guarantee the accuracy and robustness of machine learning models and time series forecasting.

The first step of data preprocessing is dealing with the missing values in the data set. Missing data values in air quality measurements are quite common in environmental data sets and may have a substantial impact on model accuracy if not treated carefully. Missing values are treated using interpolation and forward filling methods in this research.

The second preprocessing task is outlier removal. Outliers in pollutant data due to sensor malfunctions or unusual occurrences can result in biased model training. Statistical methods like z-score tests can be employed to detect outliers that are far from the usual pollutant patterns.

The third preprocessing task is the normalization of pollutant data. Since the units of different pollutants are not the same, min-max normalization is performed to ensure that all pollutant data is on the same scale. This is especially necessary for distance calculations in algorithms like KNN.

VII. Algorithmic Workflow and Pseudocode

The proposed system has a modular workflow that combines data preprocessing, AQI prediction via machine learning, and AQI forecasting via time series analysis. The workflow starts with data retrieval from the CPCB dataset, followed by data preprocessing for enhanced data quality.

Next, two separate models are trained. The KNN regression model makes AQI predictions based on pollutant concentration, while the SARIMAX model makes future AQI forecasts based on past AQI time series data.

User inputs are handled by a Flask-based backend, where real-time AQI predictions and short-term AQI forecasts are carried out. The output is presented via an interactive web interface, facilitating easy analysis of air quality status.

VIII. Comparative Analysis with Existing Methods

The conventional air quality prediction models, like linear regression and ARIMA, have limitations in dealing with the nonlinear relationships and interactions between air quality variables. The machine learning models, like SVR and Random Forest, have improved accuracy but with high computational complexity.

The KNN-SARIMAX framework is a balanced solution that leverages the simplicity and interpretability of KNN and the capability of SARIMAX to model time series data. Unlike other models, the combined framework enables real-time AQI prediction and short-term forecasting in a single system.

Compared to other models, the proposed system has improved deployability with a web-based interface and easier usability for air quality monitoring.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

IX. Mathematical Formulation

Air Quality Index (AQI) is a quantitative value that reflects the overall air quality based on the concentration of multiple air pollutants. Mathematically, AQI is calculated by finding the sub-index for each air pollutant and taking the maximum sub-index as the AQI value. Let C_p be the concentration of air pollutant p. The sub-index I_p for air pollutant p is given by the following piecewise linear function:

$$I_p = \frac{I_{hi} - I_{lo}}{C_{hi} - C_{lo}}(C_p - C_{lo}) + I_{lo}$$

where C_{hi} and C_{lo} represent breakpoint concentrations surrounding C_p, and I_{hi} and I_{lo} are the corresponding AQI breakpoints. The overall AQI is computed as:

$$AQI = \max\{I_{PM2.5}, I_{NO_2}, I_{NOx}, I_{SO_2}, I_{CO}\}$$

This definition emphasizes the nonlinear and multi-dimensional nature of AQI calculation, which necessitates the application of machine learning models for prediction.

In KNN regression for AQI prediction, each data point is represented as a feature vector:

$$X = [x_1, x_2, x_3, x_4, x_5]$$

where x_1 to x_5 correspond to PM2.5, NO_2, NOx, SO_2, and CO concentrations respectively. Given a query point X_q, the Euclidean distance between X_q and a training sample X_i is computed as:

$$d(X_q, X_i) = \sqrt{\sum_{j=1}^{n}(x_{qj} - x_{ij})^2}$$

The predicted AQI value is obtained by averaging the AQI values of the k nearest neighbors:

$$\hat{y} = \frac{1}{k}\sum_{i=1}^{k} y_i$$

For AQI forecasting using SARIMAX, the AQI time series $\{y_t\}$ is modeled as:

$$\Phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D y_t = \Theta_q(B)\Theta_Q(B^s)\epsilon_t$$

where B is the backshift operator, p, d, q are non-seasonal orders, P, D, Q are seasonal orders, s is the seasonal period, and ϵ_t is white noise. This formulation enables SARIMAX to capture both short-term dependencies and long-term seasonal patterns.

X. Methodology: K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) algorithm is a supervised machine learning algorithm that has been widely used for classification and regression tasks. In this study, the KNN algorithm is used as a regression algorithm to forecast the Air Quality Index (AQI) value using the concentration levels of the pollutants. The choice of using the KNN algorithm is informed by its simplicity, interpretability, and ability to handle non-linear relationships between multiple input variables and a continuous output variable.

The KNN algorithm is a non-parametric learning algorithm that is instance-based. This means that the KNN algorithm does not make any assumptions about the underlying distribution of the data. Instead, the KNN algorithm stores all the training data and uses it to make predictions based on the similarity between new input instances and existing data points. This property of the KNN algorithm makes it ideal for use in air quality studies, where the behavior of the pollutants is complex and non-linear.

In the proposed system, the input features are the concentrations of pollutants such as PM2.5, NO_2, NOx, SO_2, and CO, while the target output is the AQI value. Each data point is a vector in a multidimensional space. When a user enters new values for the pollutants, the KNN algorithm calculates the distance between the input vector and all the vectors in the training data.

The Euclidean distance is the similarity measure used to find the proximity of data points. The distance between two points is given by:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

where x and y are two data points in the feature space, and n is the number of input features. Based on the calculated distances, the algorithm selects the k nearest neighbors that are most similar to the input instance.

The predicted AQI value is calculated by finding the average AQI of the k nearest neighbors. The choice of the value of k is critical in determining the performance of the model. A small value of k results in a noise-sensitive model, while a large value of k results in oversmoothing of predictions, leading to a loss of sensitivity to local patterns. For this research, the value of k is determined experimentally to strike a balance between bias and variance.

The trained KNN model is then incorporated into the system for real-time AQI prediction. When users input values of the pollutants via the web interface, the KNN model analyzes the input values, searches for similar patterns of past pollution, and displays the predicted AQI value immediately. This helps to achieve rapid and easy AQI prediction, making the model ideal for real-time applications.

The KNN approach is thus an efficient tool for AQI prediction using past pollutant data and similarity learning. Its simplicity, interpretability, and flexibility make it a trustworthy tool in the proposed air quality monitoring system.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

XI. Methodology: SARIMAX Model

Although machine learning algorithms like KNN work well for real-time AQI forecasting using pollutant concentration values, they do not capture the temporal dependencies and seasonal patterns that exist in air quality datasets. To overcome this drawback, the proposed system uses a Seasonal AutoRegressive Integrated Moving Average with eXogenous variables (SARIMAX) model for short-term AQI prediction. SARIMAX is an advanced time-series forecasting model that can be considered an extension of the popular ARIMA model, which also accounts for seasonal patterns and external variables.

The SARIMAX model is well suited for air quality forecasting because AQI has strong temporal dependencies, including daily, weekly, and seasonal patterns like the peak winter pollution levels and monsoon-season dispersion. The SARIMAX model can capture these patterns explicitly to make more accurate short-term predictions compared to models that do not account for temporal patterns.

The parameters of a SARIMAX model consist of two sets of parameters: non-seasonal parameters (p, d, q) and seasonal parameters (P, D, Q, s). The non-seasonal parameters are used to define the autoregressive order p, differencing order d, and moving average order q, while the seasonal parameters are used to define the seasonal autoregressive component, seasonal differencing, seasonal moving average, and the seasonal period s. In this research, the value of the seasonal period is determined based on the yearly AQI seasonal patterns.

The SARIMAX model is trained on the past AQI data available in the CPCB dataset. While training the model, it learns how past AQI values are related to future values, considering the seasonal variations. The goal of the training process is to maximize the likelihood function to estimate the parameters of the model.

After the model has been trained, it is utilized for short-term AQI forecasting, which is done for the next five days. The result of the forecasting process includes the forecasted AQI values along with the confidence intervals, which represent the uncertainty of the predictions.

The SARIMAX forecasting module is also incorporated into the proposed system to serve as a complement to the KNN-based AQI prediction module. Although the KNN method is capable of providing real-time AQI predictions based on real-time pollutant inputs, the SARIMAX method provides a temporal outlook on AQI predictions by forecasting future AQI trends.

In summary, the SARIMAX approach improves the reliability of the proposed air quality monitoring system by accounting for the temporal aspects of AQI data. The incorporation of the SARIMAX approach with machine learning approaches for AQI predictions provides a holistic solution for air quality analysis and forecasting in urban areas.

XII. Implementation Details

The proposed Air Quality Monitoring and AQI Forecasting system is developed as an end-to-end application that combines data processing, machine learning algorithms, time series forecasting, and a web-based interface. The development of the application is done using Python because of its rich libraries and resources for data processing, machine learning, and web development.

The development of the application starts with offline model training. The KNN regression algorithm is trained on the pre-processed CPCB dataset, with pollutant concentration levels as input variables and AQI levels as target variables. After training, the KNN algorithm is serialized and saved as a pickle file for efficient reuse during deployment. This is because the model does not need to be trained every time a prediction is made.

Likewise, the SARIMAX model is also trained on the historical AQI time series data. The AQI data is indexed according to dates to maintain a chronological order, and the parameters of the model are determined through exploratory data analysis. After training, the SARIMAX model is also stored for future use. This trained model is used to make short-term AQI forecasts according to the identified patterns.

The backend of the system is implemented using the Flask web application framework. The Flask framework serves as an interface between the trained models and the frontend design. Upon receiving the pollutant data input from the user via the web interface, the Flask server processes the data and sends it to the KNN model for AQI prediction. The predicted AQI result is then sent back to the user in real-time.

For AQI forecasting, the Flask backend system loads the trained SARIMAX model and makes short-term forecasts based on the latest AQI data. The forecasted results are then processed and displayed through graphical plots. These plots enable the user to view the future air quality patterns in a simple and comprehensible manner.

The frontend of the system is developed using HTML and CSS for designing a simple and user-friendly interface. Input fields enable users to input the concentrations of the pollutants, whereas output fields are used to display the predicted AQI values and graphs of the forecasts. The interaction between the frontend and backend of the system enables a smooth flow of data.

In conclusion, the modular design of the system improves its scalability and maintainability. The system's different parts, such as data preprocessing, training, prediction, forecasting, and visualization, are developed separately. This will help in the future development of the system by integrating real-time sensors, mobile apps, and cloud computing.

Fig. 2. Web-Based User Interface for AQI Prediction and AQI Scale Visualization

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

XIII. Experimental Setup

The experimental analysis of the proposed Air Quality Monitoring and AQI Forecasting system is carried out by using real-world air quality data collected from the Central Pollution Control Board (CPCB). The aim of the experimental design is to evaluate the efficiency of the KNN regression model in AQI prediction and the SARIMAX model in short-term AQI forecasting.

All experiments are carried out in the Python programming language. Libraries such as NumPy and Pandas are used for data manipulation and preprocessing, and scikit-learn is used for developing the KNN regression model. The SARIMAX model is developed using the *statsmodels* library, which is equipped with efficient tools for time series analysis and forecasting. Result visualization is done using Matplotlib for creating plots related to AQI trends and forecasting results.

The preprocessed dataset is divided into training and testing subsets to evaluate model performance on unseen data. For the KNN regression model, an 80:20 train–test split is employed, where 80% of the data is used for training and the remaining 20% is reserved for testing. This split ensures that the model learns from sufficient historical data while allowing reliable evaluation of prediction accuracy. The value of k in the KNN model is selected experimentally to balance bias and variance.

For the SARIMAX forecasting model, the AQI time series is split in a chronological manner. The past AQI values are employed for training the model, and the latest values are held for validating the accuracy of the forecasts. The seasonal factors are determined according to the yearly trends in the data. The forecast horizon is defined for five days to enable short-term AQI forecasts, which are useful for early warning and decision-making.

The performance of the models is assessed using conventional error measures. For AQI prediction, the accuracy of the predictions is assessed by comparing the AQI predictions with the actual AQI values in the test data. For AQI forecasting, the Root Mean Square Error (RMSE) measure is employed to assess the difference between the forecasted AQI values and the actual AQI values. The RMSE measure is employed owing to its sensitivity to large errors in predictions, which is useful for assessing the reliability of the forecasts.

All the experiments are performed on a common computing platform, and the trained models are implemented using the Flask-based web application.

XIV. Results and Discussion

This section discusses the experimental results obtained from the proposed Air Quality Monitoring and AQI Forecasting system and provides a detailed discussion of the results. The performance of the KNN regression model for AQI prediction and the SARIMAX model for short-term AQI forecasting is analyzed using real-world CPCB air quality data.

A. AQI Prediction Results Using KNN

The KNN regression model is tested by comparing the predicted AQI values with the actual AQI values from the test dataset. The outcome shows that the KNN model is able to accurately identify the relationship between pollutant concentrations and AQI values. Since KNN is a distance-based learning algorithm, the model makes predictions by searching for past pollution patterns that match the input data. This enables the model to make intuitive and interpretable AQI predictions.

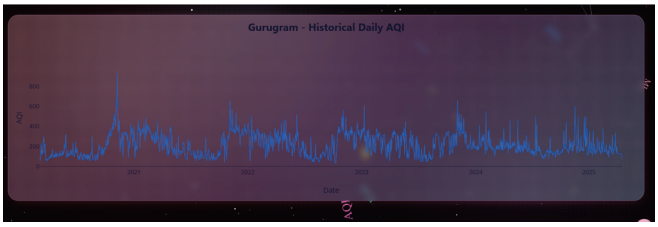

Fig. 3. Historical Daily AQI Trends for Gurugram City

From the experimental results, it can be seen that the KNN model works well for all AQI ranges, whether low, moderate, or high. The accuracy of the predictions is high when pollutant concentrations are following past trends in the dataset. Small discrepancies are noticed when there is a sudden pollution surge, which can be explained by sudden environmental or human-induced changes that are not accounted for in the past trends.

The ease of use of the KNN model is one of the reasons why it has a fast response time and is ideal for real-time AQI prediction in a web-based setting. Users get immediate AQI predictions after they enter the pollutant values, which increases the usability of the system.

B. AQI Forecasting Results Using SARIMAX

The SARIMAX model is tested for short-term AQI forecasting purposes based on the historical AQI time series data. The model produces forecasts for a five-day term, incorporating both trend and seasonal aspects of the AQI time series. The forecast performance is tested using the Root Mean Square Error (RMSE) measure, which estimates the difference between the forecasted and actual AQI concentrations.

The outcome indicates that the SARIMAX model is capable of producing smooth AQI forecasts, which effectively capture the seasonal aspects of AQI, such as periodic pollution peaks during the winter season and reduced AQI concentrations during times of increased atmospheric dispersion. The addition of seasonal parameters to the model has greatly improved the accuracy of AQI forecasts compared to non-seasonal models.

While the SARIMAX model may not accurately forecast extreme pollution peaks, the model is capable of producing accurate short-term trend forecasts, which are essential for early warning systems. The AQI forecast plots produced by the model effectively demonstrate the future AQI conditions, which can be easily interpreted by the users.

C. Comparative Analysis and Discussion

The combination of KNN and SARIMAX models provides complementary benefits. The KNN model has strong capabili-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

ties in real-time AQI prediction based on the current pollutant inputs, and the SARIMAX model has a temporal aspect in AQI forecasting by predicting future AQI trends. The combination of both models overcomes the disadvantages of individual models and improves the overall system reliability.

The proposed framework has better adaptability to the nonlinear behavior of the pollutants and seasonal changes compared to traditional statistical models. The previous research works were focused on either prediction accuracy or time series forecasting independently. However, the combined approach in this work provides both features in a single system.

The web-based implementation further improves the applicability of the system by allowing user interaction and visualization of the results. The users can not only see the AQI prediction values but also examine the short-term trends, which helps in taking proactive health actions.

In conclusion, the experimental results confirm the effectiveness of the proposed air quality monitoring system. The system provides a good balance between accuracy, interpretability, and usability and can be effectively applied to real-world urban air quality monitoring.

XV. Limitations and Future Work

Although the proposed Air Quality Monitoring and AQI Forecasting system performs well, there are some limitations to be considered. The proposed system mainly uses the historical air quality data collected from the CPCB monitoring stations. Therefore, the quality of the data has a direct impact on the accuracy of the forecasting and prediction results. Abrupt pollution events triggered by unforeseen factors such as accidents, fires, construction activities, and natural weather conditions cannot be properly represented by the proposed models.

The KNN regression model, although simple and interpretable, requires the entire training data to be stored in memory and can become computationally intensive for larger datasets. Moreover, the performance of the KNN algorithm is highly sensitive to the choice of the parameter k and the quality of feature scaling. Similarly, the SARIMAX model assumes linear relationships among the time series and cannot properly represent highly nonlinear relationships among AQI factors.

Future work can address these limitations by incorporating additional data sources and advanced modeling techniques. Integration of real-time sensor data and meteorological parameters such as temperature, humidity, wind speed, and rainfall can significantly enhance prediction accuracy. Deep learning models such as Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), and hybrid CNN–LSTM architectures can be explored to capture complex nonlinear and long-term dependencies in air quality data.

Furthermore, the system can be extended to include geospatial visualization using Geographic Information Systems (GIS) for region-wise AQI monitoring. Deployment on cloud platforms and development of mobile applications can improve scalability and accessibility. These enhancements would enable the proposed framework to serve as a comprehensive and intelligent air quality monitoring solution for smart city environments.

XVI. Conclusion

This paper has discussed a holistic machine learning-based approach for air quality monitoring and AQI forecasting through time series models. The proposed system combines K-Nearest Neighbors (KNN) regression for real-time AQI prediction and Seasonal AutoRegressive Integrated Moving Average with eXogenous variables (SARIMAX) for short-term AQI forecasting. Through the use of past air quality data collected from the Central Pollution Control Board (CPCB), the proposed system is able to efficiently model the relationships and dynamics of the air pollutants, as well as the seasonality of AQI patterns.

The experimental results show that the KNN model is capable of providing accurate and interpretable AQI predictions based on air pollutant concentrations, and that the SARIMAX model is able to provide stable and reliable short-term AQI forecasts. The combination of both models into one system improves the overall reliability and usability of the system. The web deployment using Flask also allows for user interaction and visualization of air quality data.

In summary, the proposed framework provides a viable, scalable, and efficient solution for urban air quality monitoring and early warning systems. The work presented in this paper demonstrates the potential of integrating machine learning and time series forecasting methods for solving real-world environmental problems and provides a solid foundation for future research and development in intelligent air quality monitoring systems.

References

[1] Y. Özüpak, F. Alpsalaz, and E. Aslan, "Air Quality Forecasting Using Machine Learning: Comparative Analysis and Ensemble Strategies," *Water Air Soil Pollution*, vol. 236, no. 464, 2025.

[2] N. S. Gupta et al., "Prediction of Air Quality Index Using Machine Learning Techniques," *Journal of Environmental and Public Health*, 2023.

[3] R. Janarthanan et al., "A deep learning approach for prediction of air quality index," *Sustainable Cities and Society*, vol. 67, 2021.

[4] W. Mao et al., "Modeling air quality prediction using deep learning," *Sustainable Cities and Society*, vol. 65, 2021.

[5] M. M. Rahman et al., "AirNet: Predictive machine learning model for air quality forecasting," *Environmental Systems Research*, vol. 13, no. 1, 2024.

[6] M. Lakshmipathy et al., "Advanced ambient air quality prediction through ensemble learning," *Knowledge and Information Systems*, vol. 66, 2024.

[7] S. Wang and Y. Zhang, "An attention-based CNN model for urban air quality estimation," *Atmospheric Environment*, vol. 340, 2025.

[8] K. Siwek and S. Osowski, "Data mining methods for prediction of air pollution," *International Journal of Applied Mathematics and Computer Science*, vol. 26, no. 2, pp. 467–478, 2016.

[9] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th ed., Wiley, 2015.

[10] Y. Cheng et al., "Air quality forecasting using LSTM networks," *IEEE Access*, vol. 8, pp. 123234–123245, 2020.

[11] A. Bekkar and B. Hssina, "Air pollution prediction in smart cities using deep learning," *Journal of Big Data*, vol. 8, no. 1, 2021.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

[12] J. Zhang and W. Ding, "Deep learning-based air quality prediction model," *Neural Computing and Applications*, vol. 31, pp. 819–829, 2019.
[13] X. Li and L. Peng, "Air quality prediction using machine learning models," *Environmental Science and Pollution Research*, vol. 27, pp. 39690–39702, 2020.
[14] P. Rao and A. Kumar, "Urban air quality prediction using ensemble learning," *Environmental Monitoring and Assessment*, vol. 194, 2022.
[15] R. Singh and S. Verma, "Machine learning approaches for pollution forecasting in Indian cities," *Journal of Cleaner Production*, vol. 382, 2023.
[16] Y. Liu and T. Wang, "Seasonal air quality forecasting using SARIMA and machine learning," *Atmospheric Pollution Research*, vol. 13, 2022.
[17] K. Patel and R. Mehta, "Short-term AQI forecasting using hybrid time-series models," *Environmental Challenges*, vol. 15, 2024.
[18] Central Pollution Control Board (CPCB), "National Air Quality Monitoring Programme," Government of India. [Online]. Available: https://airquality.cpcb.gov.in
[19] World Health Organization, "WHO global air quality guidelines," WHO Press, 2021.
[20] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[21] F. Liang, C. Yu, and Z. Zhang, "Air quality prediction based on ensemble learning," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 11, pp. 5054–5063, 2018.
[22] Y. Zheng, F. Liu, and H.-P. Hsieh, "U-Air: When urban air quality inference meets big data," *Proceedings of the 19th ACM SIGKDD Conference*, pp. 1436–1444, 2015.
[23] J. Cao et al., "Air quality prediction using multiple machine learning models," *Atmospheric Environment*, vol. 212, pp. 236–247, 2019.
[24] C. Huang and J. Kuo, "Forecasting air pollution using hybrid time-series models," *Environmental Modelling & Software*, vol. 127, 2020.
[25] M. Kim and S. Lee, "Short-term air quality forecasting using data-driven approaches," *Sustainable Cities and Society*, vol. 78, 2022.