AirCare: Integrating VAR-XGBoost Models for County-Level AQI Prediction and Health Advisory

Dr Swarnalatha P*, Rishabh Sharma 21BCE0943†, Caleb Missier 21BCE3949‡, Aryan Singh 21BCE3026§ * Email: pswarnalatha@vit.ac.in
† Email: rishabh.sharma2021@vitstudent.ac.in
‡ Email: calebxavier.levon2021@vitstudent.ac.in
§ Email: aryan.singh2021b@vitstudent.ac.in
Department of Computer Science and Engineering
Vellore Institute of Technology

Abstract

AirCare builds a state-of-the-art air quality forecast framework by integrating two different machine learning methods to delve further into environmental health. The first one includes the place-centric Vector Autoregression (VAR) model, borrowing from 11 years of historical EPA data for six criteria pollutants-namely CO, NO_2, SO_2, O_3, PM_{10}, and $PM_{2.5}$-for 100 U.S. counties. This model embeds the spatial correlations by analyzing the trends of pollutants in neighboring counties, hence improving the predictive accuracy of traditional time-series models like ARIMA. The other approach is an XGBoost-based weather-driven model that dynamically processes real-time meteorological data-time, humidity, wind speed-pulled from OpenWeatherMap API. Both methods are complementary to each other and enable both long-term and real-time forecasting of AQI.

This novelty of AirCare consists of a personalized health recommendation engine that synthesizes AQI predictions with individual health profiles to provide customized advisories for vulnerable groups, among them respiratory and cardiovascular sensitive ones. Delivered through a web-based interface, the system provides interactive AQI visualization, robust error handling, and secure storage of user data. Empirical evaluation highlights the prediction reliability of the system, with the VAR model outperforming ARIMA and the XGBoost model offering statistical significance with robust real-time adaptability. Translating complex environmental data into actionable health insights, AirCare nudges the intersection of machine learning and public health forward in ways that enable better decision-making by individuals and policymakers alike.

Index Terms

Air Quality Index, Machine Learning, Health Recommendations, Vector Autoregression, XGBoost

I. Introduction

The air quality forecasting and health recommendations are basic in the aim of protecting public health and reducing adverse health impacts due to environmental pollution, which is one of the key problems facing urban areas throughout the world. Poor air quality has well-documented negative effects on both respiratory and cardiovascular health and is an escalating challenge to public health systems. Air pollution, because of rapid urbanization and industrialization in most regions, especially in developing countries, is at alarming rates. Thus, there is an enhanced need for effective air quality prediction systems and personalized health recommendations for proactive health management. While much effort has been invested in developing air quality prediction models and health advisory systems, a number of challenges remain in how historical data is integrated with real-time information, providing personalized health advisories, comprehensive spatial monitoring, and dynamic adaptation to constantly changing environmental conditions.

Among these, the integration of historical data related to air quality with real-time information on current monitoring stands out as one of the most determining factors for further improvement in air quality predictions. While historical data gives good context and long-term trends, this might miss sudden changes in air quality due to transient factors such as abnormal weather, congestion, or industrial activities. On the other hand, real-time data, though critically important for immediate alerts and short-term forecasting, is limited by its insufficient temporal patterns for enabling long-term predictions with higher accuracy. Therefore, recent studies deal with the development of hybrid models that put historical and real-time data together for more reliable and dynamic prediction. The models use machine learning algorithms like XGBoost, Long Short-Term Memory LSTM networks, and other advanced methods that can handle long-term trends and real variations with a view to enhancing robustness in the air quality forecasting system.

Besides the forecast of air quality itself, personalization of health recommendations based on air quality levels also becomes increasingly important. Most air quality prediction systems offer general health advice, such as advising the public against outdoor activities when air quality is poor. However, personalized recommendations are able to give more specific guidance towards the individual health profile. It would take into account pre-existing respiratory conditions, age, and the amount of activity a person does in a day, as well as sensitivity to pollutants. The recent research in this area has taken note and explored the integration of wearable gadgets and health monitoring systems for providing real-time alerts in regard to one's exposure to particular pollutants such as PM2.5, NO2, or ozone. This kind of personalized health recommendation will be very helpful for more sensitive populations, including the elderly, children, and people with pre-existing cardiovascular or respiratory conditions, because a more efficient strategy may be built in order to lower the risk of a given population due to poor air quality.

Another big challenge that has to be overcome is that of lousy spatial monitoring. Traditional air quality monitoring systems have largely relied upon sparse numbers of monitoring stations, which are often distributed very non-uniformly, especially over rural or less-developed areas. The lack of comprehensive spatial coverage in this regard leads to huge gaps in the data, besides probably incorrect predictions of air quality, especially within regions with localized pollution sources. Recent development of the satellite monitoring and low-cost sensing technologies is a promising solution to this problem. These sensors can provide extensive and real-time monitoring that allows better assessment of air quality both in urban and rural settings. At the same time, active citizen-based sensor networks have emerged as one of the parallel solutions for extending data collection by engaging the public in monitoring efforts, thus providing more granular and locally relevant air quality data.

Other factors make the conditions dynamic; hence, this is another challenge for air quality prediction systems. Variables that influence air quality are very many, including meteorological conditions of wind speed, temperature, humidity, and atmospheric pressure, to anthropogenic activities such as transportation and industrial emissions. This is because changes in the weather-sudden reversals of wind patterns, for example-can sometimes dramatically affect the concentrations of pollutants. Events such as wildfires or traffic accidents lead to sudden peaks in pollution levels. Air quality prediction models, usually based on fixed parameters, are bound to fail when these dynamic changes take place, and forecasts may be very poor under atypical conditions. Recent studies have tended to develop more flexible and adaptive models that are able to take real-time meteorological and environmental data for dynamic adjustment of predictions, thus giving more accurate forecasts for variable conditions.

Despite these challenges, there have been significant improvements in air quality prediction and health recommendation systems. Recent development in machine learning and deep learning techniques such as ensemble models, CNNs, and RNNs has much better and more reliable predictions of these parameters. Such models can capture the complex interaction amongst various pollutants, meteorological parameters, and health outcomes, therefore providing more advanced information about trends and risks related to air quality. In this regard, there is a recent upsurge in interest in the integration of health data with air quality predictions. Such systems will enable more actionable health guidance since they will couple real-time air quality levels with health outcomes. For example, the occurrence of respiratory distress or cardiovascular events may be forecast based on pollution levels. In this manner, this data-driven approach will subsequently allow for better-informed and proactive health management, which, in turn, will minimize risks resulting from poor air quality.

Besides that, access and the usability of air quality prediction systems have been well improved by the advances in data visualization techniques, making it easy for users to interpret from interactive dashboards and real-time visualization of complex air quality data on the trends of pollutants and health advisories in an intuitive manner. These tools allow the citizens to take matters concerning their health in hand and, at the same time, provide policymakers and the public health officials with data-driven decisions to help reduce the effects caused by pollution in public health.

While much progress is being achieved in the development of air quality prediction and health recommendation systems, integrating real-time data with historical trends, personalization of health recommendations, comprehensive spatial coverage, and adaptation to dynamic environmental conditions remain some of the difficult challenges that are yet to be fully addressed. Continuous development within the realms of machine learning, sensor technologies, and integration of health data enables a path towards more accurate, personalized, and adaptive systems. These will be ingenious innovations in air quality monitoring and response, thus availing even more sound solutions for the protection of public health and limiting the deleterious effects of air pollution. The air quality forecast system of the future, with further research into technology development, shall be suitably positioned to provide actionable insights in a timely manner that protects vulnerable populations and fosters sustainable urban growth.

II. Literature Survey

A. Fragmented Prediction Approaches

Another important challenge in the field of air quality prediction is the limited integration of historical data with real-time meteorological conditions. Most of the models in existence either rely on location-based historical data or provide weather-based predictions but hardly incorporate the two for an integrated model of forecasting. Many works tried to bridge this gap by proposing hybrid models that merge both kinds of data to improve the accuracy of the prediction.

Wu and Lin (2023) have come up with the hybrid model-SD-SE-LSTM-BA-LSSVM, which incorporates secondary decomposition, AI techniques, and optimization algorithms to perform AQI forecasting. This approach has shown better performance in estimating AQI values through efficient exploitation of both historical air quality data and optimization methods. This again assures the advantage of diverse methodologies being integrated into the model for AQI forecasting.

Sarkar et al. (2022) developed a hybrid of LSTM-GRU-based models for AQI forecasting. The hybrid model ensured better performance compared to the traditional methods with respect to MAE and R^2 scores. It really worked with respect to the amalgamation of various neural network architecture types to enhance the accuracy of predictions. This approach addresses the challenge of integrating multiple machine learning techniques for more accurate and reliable air quality forecasting.

Zhu et al.(2021) proposed two mongrel models, videlicet, EMD- SVR- Hybrid and EMD- IMFs- Hybrid, that make use of empirical mode corruption to enhance the delicacy of AQI soothsaying . These models proved that the incorporation of various forecasting methods would lead to a much higher level of precision in AQI forecasts, especially in regions with complex behaviors of pollutants. These models combine different approaches to provide robust and highly accurate predictions.

B. Lack of Health Recommendation Personalization

Though most of the air quality prediction systems give general information on AQI, the health recommendations are not personalized. Most of the systems do not address the individual health conditions or focus on particular pollutant exposures or individual threshold for the vulnerable sections of the population. A few recent studies have tried to address the gap by focusing on personalized health recommendations based on AQI predictions and individual health profiles.

Zhang et al. (2022) explored the use of machine learning techniques to predict variations in indoor air quality and their impact on health outcomes. They proposed a system that provides personalized health advisories based on predicted indoor air quality levels, specifically targeting individuals with respiratory conditions. Their work emphasizes the importance of tailoring health recommendations to individual health conditions to enhance the effectiveness of air quality prediction systems.

Later on, Gu et al. (2023) developed a hybrid predictive model for $PM_{2.5}$ forecasting. The model indeed showed that it can bring about more precise peak value predictions than individual models. The interpretability coming with such a model was able to bring forth personalized health recommendations for individuals according to exposure levels. This study showed that incorporation of individual health parameters with AQI forecast would result in more efficient and specific health advisories for the vulnerable groups, which include children, aged, and those suffering from past respiratory or cardiovascular disease.

C. Geospatial Monitoring Gaps

Another major lacuna in most air quality models is the failure to take into consideration other neighboring regions and county-level AQI reporting. Most models predict air quality for a given location but do not take into account the greater regional effects from pollution that result in incomplete, if not incorrect, assessments. A number of studies have attempted to develop enhanced geospatial monitoring to overcome this lacuna.

Saez et al. (2021) developed a hierarchical Bayesian spatiotemporal model that integrated sparse monitoring station data into improving the AQI predictions of both short-term and long-term exposures. Their model performed particularly well in areas that have few air quality monitoring stations, thereby overcoming regional AQI forecasting difficulties. This could give a better representation of air quality over a wide area and is very important when a region does not possess far-reaching monitoring infrastructure.

Jurado et al. (2022) applied a convolutional neural network for real-time air pollution prediction with information on wind speed, traffic flow, and building geometry. The proposed approach yielded higher spatial accuracy of air quality forecasts by accounting for factors other than traditional pollutant measurements. In fact, integrating spatial features from the surrounding environment could definitely enable better forecasting, particularly in urban areas characterized by complex dynamics of pollution.

D. Complexity in Data Accessibility and Usability

One of the sustained challenges in air quality prediction is how to present AQI data. AQI data is very often made available in formats difficult to be interpreted by the general public. As a response, researchers focused their attention on the development of more user-friendly systems that could simplify data presentation and improve data accessibility.

Rakholia et al. (2022) identified a model that integrates a wide array of aspects related to air quality: meteorological conditions, traffic flow, and pollution levels. This model further opens access to AQI data while at the same time giving a fuller picture of air quality, making the information more accessible to those people in society who might need or be less conversant with air quality science.

Elsheikh et al. (2021) utilized LSTM networks in order to model air quality and water production to show how machine learning methods can simplify complex environmental data sets. The paper shows a possibility of the use of machine learning models in making complex air quality data more comprehendible and usable to the decision-makers and the general public.

E. Limitation of Real-time Adaptation

Most of the air quality prediction systems are based merely on previous data and do not respond dynamically to the current weather. This is one major limitation that seriously constrains the accuracy and responsiveness of these models at times when unexpected events take place in the environment. Recent studies have developed models that allow the objectification of real- time data to give further adaptive and accurate prognostications.

Zhang et al. (2022) used a CNN-GRU model to predict AQI by combining the CNN for feature extraction and GRU for modeling temporal dependencies. This hybrid approach allows for dynamic responses to real-time meteorological data, hence, enabling the model to adapt quickly to changes in air quality caused either by sudden weather changes or pollutant emission.

Mao et al. (2021) proposed a framework, namely temporal sliding LSTM (TS-LSTME), for AQI prediction based on the historical record of $PM_{2.5}$ data, meteorological data, and real-time temporal data. This model was able to show how this incorporation of real-time data processing can enhance responsiveness and precision of air quality prediction to ensure systems can cope with the existing ambient environment.

F. Multitype Pollution Monitoring

Most air quality models are designed for monitoring one type of pollutant or several types. It is, however, the interaction of many that shows air quality; thus, there is a greater understanding when several types of pollutants are simultaneously monitored. Current research uses multi-pollutant models that give a wide-ranging picture about air quality.

Rakholia et al. (2022) proposed a multi-pollutant model involving CO, NO_2, and $PM_{2.5}$, with meteorological and urban factors. In that way, a wider overview of air quality could be provided by improving the accuracy of forecasts and understanding the interaction among different pollutants.

Gu et al. (2023) then developed a hybrid model for $PM_{2.5}$ prediction, which is more accurate and interpretable compared to previous approaches. The proposed approach fully mines multiple pollutant information in their prediction model and considerably improves the accuracy of air quality forecasting, indicating the importance of multitype pollution in air quality prediction.

G. Limited Public Health Integration

While the air quality models have made significant improvements in AQI forecasting, few have integrated public health considerations or provided specific guidance for healthcare providers. This has created a limitation in how the AQI models can be usefully applied to support public health interventions. Anchoring their work on integrating health impacts with air quality predictions, several studies have tried to fill this gap.

Zhang et al. (2022) analyzed indoor air quality variations associated with human health and constructed personalized health management strategies based on AQI predictions. Their work identified, with urgency, the need for air quality models that provide evidence to inform healthcare decisions, particularly in managing the health of vulnerable populations.

Gu et al. (2023) also showed how inclusion of health aspects with $PM_{2.5}$ prediction models can enhance their interpretability and utilize them efficiently for public health purposes. In this respect, it would be justified to believe that incorporation of health-related information within AQI forecasting models can result in effective health advisories and interventions.

H. Literature Survey - Conclusion

Recent research has considerably improved air quality prediction and health recommendation systems. The hybrid models are improving the forecasting accuracy, using machine learning with statistical methods, while real-time adaptation and personalized health recommendations are becoming increasingly feasible. Geospatial models improved AQI prediction for border regions, while multitype pollutant monitoring offers a more complete picture of air quality. Nevertheless, a lot more gaps need to be filled in order to make these models practical and user-friendly. Future studies should enhance the accessibility of air quality data, application of real-time data to decision-making processes, and personalization of health recommendations according to individual health conditions and geographic factors.

III. Methodology

A. Hardware and Software Used

The hardware and software used in this system are judiciously selected to meet the performance and reliability requirements for air quality prediction and personalized health recommendations.

1) Hardware Specifications:: The system requires an Intel Core i5 processor or higher to handle the computational load of running machine learning models effectively, especially for the algorithms like VAR and XGBoost. For this, a minimum of 8GB RAM is required, though 16GB will be more satisfactory for running larger data sets in real time. The minimum for deployment is 256GB SSD, but for storehouse, 50 GB more can be added for 11 years of the existing historic EPA data. The application will be needing an internet connection, as API interactions are required for real-time access of weather data from OpenWeatherMap and historical data from Google BigQuery. It requires basic integrated graphics for visualizations.

2) Software Requirements:: The application is written in Python 3.8 or later. This provides the available and required libraries and frameworks to process the information and perform machine learning. The IDE with Python inside will be Visual Studio Code; however, any other compatible IDE can substitute it. Core dependencies for the project will include Flask for web development, with Flask-SQLAlchemy for database management. Besides these, a number of libraries are responsible for data processing, namely Pandas and NumPy. For machine learning, Scikit-learn is used, while Statsmodels-when performing VAR modeling. Matplotlib and Plotly libraries are used for static and interactive air quality visualizations, respectively.

The system uses integrations with third-party services that provide access to large volumes of data, such as Google BigQuery and the OpenWeatherMap API for real-time weather. In back-end development, SQLite3 is used, but in production, the system relies on PostgreSQL, which is scalable and securely manages data. Source Control: Git; Repository Management: GitHub. Gunicorn is used as a WSGI web server to deploy the application, with Nginx set up in front as a reverse proxy to handle traffic efficiently. SSL certificates have been implemented to ensure secure HTTPS communication.

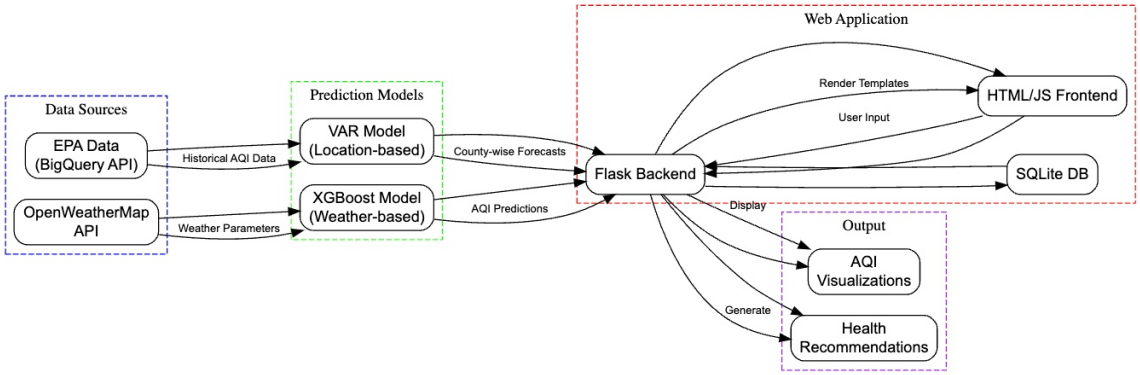

Fig. 1. Block diagram

B. OBJECTIVES OF THE SYSTEM

This system shall develop an integrated holistic air quality prediction platform incorporating machine learning and advanced statistical modeling techniques that enable highly accurate AQI value forecasts. Utilizing historical

air quality data of specific locations and current and real-time weather parameters, the system will develop comprehensive AQI predictions which are geographically specific. It would also implement a VAR model for long-run AQI forecasting in 100 U.S. counties using comprehensive historical datasets to ensure accuracy and reliability. Conversely, the system will also adopt an XGBoost model for dynamic, real-time AQI predictions that integrate meteorological variables such as temperature, humidity, wind speed, and atmospheric pressure to boost the capacities of short-run forecasting.

AQI forecasting will cover six critical pollutants, in conformation with U.S. EPA standards, which are: Carbon Monoxide (CO), Nitrogen Dioxide (NO2), Particulate Matter (PM2.5 and PM10), Sulfur Dioxide (SO2), and Ozone (O3). It will be designed to take into consideration the regional differences in sources of pollution and weather conditions to make more local and actionable predictions.

Beyond prediction, the system will help address issues related to public health through a smart personalized health advisory module. This module will analyze the predicted AQI levels together with a health profile provided by the user: age, medical history, and pre-existing conditions such as asthma, COPD, or cardiovascular diseases. Health recommendations to use by the system will be provided based on this, for example, limitation of outdoor activities or proper protection gear, including masks, should be taken, or take indoor air purification measures. For general users, it will provide advisories regarding activities based on levels of caution from ”Good” to ”Hazardous,” following standard AQI categorizations.

A full-featured web interface will be developed, showcasing an interactive dashboard that visualizes AQI trends, pollutant breakdowns, and predictive visualizations to improve accessibility and user engagement. The interface shall grant the user the ability to switch between historical and real-time data feeds, filter the results by pollutant or location, and input health-related information for customized recommendations. It shall also notify users and create alerts in cases of severe levels of AQI to ensure timely dissemination of critical information.

The system's technical architecture will be kept scalable so that additional counties, more pollutants, or more forecasting features can easily be added without any hassle in the future. Data integration pipelines shall be designed for robustness by incorporating error detection mechanisms to handle anomalies in incoming datasets effectively. High-performance computing resources will be utilized to optimize machine learning model training and inference, ensuring minimum latency in prediction updates. Further, the deployment strategies will be cloud-based to access data in real-time and support the cross-platform functionality to access the system via web and mobile devices.

Security and privacy will be at the heart of the system, encrypting user health data and location information, securely storing it in compliance with applicable data protection regulations. Regular updating of predictive models in respect of newly arriving data will ensure continued relevance and accuracy.

In a nutshell, this project will fill the existing gap in air quality forecasting and personalized public health advisories through a technologically advanced and user-centric platform. The system enables users to make decisions on safeguarding their health from air pollution risks by addressing predictive accuracy and actionable insights.

C. METHODS USED FOR THE OBJECTIVES:

1) Data Collection and Preprocessing: For integrating EPA data, 11-year historical air quality data was gathered from 100 counties across the United States. Each dataset presented main pollutant measurements like CO, NO_2, $PM_{2.5}$, PM_{10}, SO_2, and O_3. Other climate variables included in the datasets are temperature, humidity, pressure, and wind speed. Measurement data aggregated at the site level to summaries at the county level, with a guarantee that maximum values of each pollutant were considered. With regard for missing data points, a time series interpolation was done to fill up the gaps and insure that the dataset was complete. Hundreds of thousands of data points were prepared in the final dataset for use in forecasting models.

2) Weather-Based Predictions: The weather-based AQI prediction system uses the XGBoost algorithm for prognostications of estimation of air quality indicator situations grounded on real- time meteorological data.. The model involves eight important weather variables, namely temperature, humidity, wind speed, wind direction, atmospheric pressure, solar radiation, precipitation, and visibility. These parameters have been identified after rigorous feature engineering in order to find their effect on pollutant dispersion and concentration.

XGBoost was fine-tuned for high accuracy and robustness using hyperparameter tuning with grid search together with cross-validation. The model handles outliers and missing data with the least overfitting guaranteed by the use of gradient boosting techniques. Real-time weather data is fetched through APIs, and computations are made in the backend to refresh the AQI forecast every hour, thus keeping the user updated about any variations in air quality throughout the day. It also integrates real-time error-checking mechanisms that validate incoming data for anomalies before making predictions.

3) Location-Based Predictions Using VAR Model: This location-based prediction system is based on a VAR model that forecasts the AQI level in counties. Historical AQI data and pollutant-specific concentration data were provided for 100 U.S. counties to guarantee comprehensive coverage and accuracy. With dependencies between the time series for various pollutants and their respective seasonality, this VAR model captures periodic patterns that consist of spatial correlations between counties.

Advanced data preprocessing techniques, such as data normalization, detrending, and lag selection, were performed to optimize the performance of the VAR model. Accordingly, in order to validate the predictive accuracy of the model, it was considered using historical datasets from over a decade. This model worked much more efficiently compared to the standard ARIMA model in recognizing regional pollution trends and has proved to be robust for location-based AQI forecasting.

4) Health Recommendation System: The health recommendation module serves as the core feature, translating AQI forecasts into actionable advice, tailored to user-specific health profiles. AQI level categorization is performed according to established threshold values: Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, and Hazardous. Personalized recommendations will be generated on the fly, taking into consideration user input on health conditions like asthma, cardiovascular issues, or allergies.

For example, at levels of hazardous AQI, the app notifies users with respiratory conditions to avoid outdoor exposure, then use air purifiers indoors, and wear protective masks. The system further considers dominant pollutants and their health impact specifically and offers pollutant-specific advice. Advanced filtering allows recommendations to adapt for pollutant type, whether PM2.5, Ozone, or Nitrogen Dioxide, providing more relevance to the guidance given.

5) System Integration: It's a system with many integrations, including prediction models, data sources, and user interfaces, using a robust backend powered by Flask. The system interacts seamlessly for the user profile, sessions, and interactions. An SQLite database is present in local testing, but PostgreSQL has been used in production so far to handle massive-scale operations.

There are two major pipelines in the prediction system, which are the VAR-powered location-based pathway and the XGBoost-powered weather-based pathway. Each of the pathways has integrated APIs, which cater to data ingestion, processing, and visualization. The modules for the preprocessing of data standardize all the ingested data to meet the model requirements.

Plotly visualizes interactive visualizations for exploring AQI trends, forecasts, and pollutant-specific data through an intuitive web interface. Visualizations provide functionality such as filtering for location, pollutant type, and time range, thus engaging users much better and enabling easier access to the data.

6) Evaluation and Validation: The prediction models have been validated using comprehensive evaluation metrics such as RMSE, MAE, and R-squared values. More precisely, it achieved an error margin in prediction of less than 9 percent using the XGBoost model, hence very accurate to predict AQI in real time. The VAR model showed further notable improvements over ARIMA in capturing inter-county spatial dependencies and seasonal fluctuations, further cementing its reliability for location-based predictions.

This included the validation of model output against actual data from environmental protection agencies such as the EPA for consistency with industrial standards. Scatter plots and error distribution histograms visualized these discrepancies, and model retraining schedules were put in place for the continuous addition of new data.

7) Deployment and Maintenance: The system was deployed on a secure web server featuring SSL certificates for encrypted communications. APIs from OpenWeatherMap and Google BigQuery were integrated to fetch real-time and historical data in an efficient manner. The deployment pipeline includes automated testing regarding performance and error handling in order to keep the system reliable.

Application architecture supports modular updates, which means new features or models can be integrated without disturbing the functionality of other modules. Retraining of prediction models is scheduled periodically with updated datasets. Therefore, the forecasting accuracy is maintained. Logs are continuously monitored in order to proactively detect and resolve system errors.

IV. RESULTS AND DISCUSSION

A. Data Processing Results

The data processing yielded a robust dataset containing around half a million observation points. This dataset was sourced from the air quality monitoring data across 100 counties, totaling various points in each of 20 columns. The critical information dataset includes the concentration of six primary pollutants, namely $PM_{2.5}$, PM_{10}, carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and ozone (O_3), besides meteorological parameters like temperature, pressure, relative humidity, and wind speed. Geographic identifiers were also added to the dataset, providing a unique association for each observation with a county. The dataset also contains the calculated values of the Air Quality Index-AQI, hence allowing quantitative measures of air quality and giving insights into air pollution's possible impact on public health. It provides a rich dataset for

developing and validating predictive models that ensure this system will be able to capture the rich interaction between air quality, meteorological conditions, and geographic variation across diverse regions.

B. Architecture of Visualization

The architecture of the visualization system has been designed in multiple layers. This enables seamless integration with several data sources and builds interactive and insightful visualizations. It consists of historical air quality data, real-time weather data, and the forecasted outputs from VAR and XGBoost. All this incoming data from these various sources is fed into the processing layer, where the raw input data is further processed by the AQI calculator to produce the required values of AQI for every observation. In addition, the following pollutant and trend analysis modules further refine the data by showing the pattern in pollutant concentrations and air quality over a certain period.

The final output is delivered through the visualization layer, which is designed to present the processed data in an interactive and user-friendly format. This layer includes several key components: time series plots that display historical and forecasted AQI trends, county-specific maps for geographic visualization of air quality data, forecast visualizations to predict future air quality trends, and pollutant comparison charts to highlight variations in pollutant levels across different counties. The seamless flow from data sources to visual output ensures that users can access accurate and meaningful insights about air quality and pollution patterns, which can inform decision-making processes in public health, policy development, and environmental management.

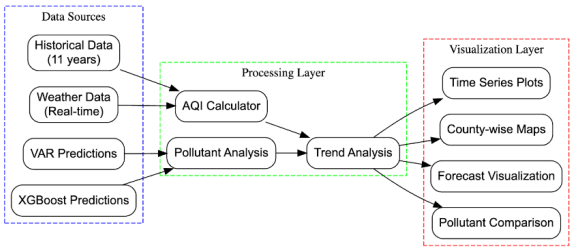

Fig. 2. Visualization Architecture

C. Visualization Components

1) Time Series Analysis: There will be several interactive visualizations of time series data in the system. The user will be allowed to study historical air quality trends and forecasted AQI. The historical series of AQI would go back 11 years, where the long-term trend of the air quality can well be observed. Users are able to observe the pattern trend of AQI variation concerning time and identify important events, such as spikes of pollution levels during certain periods. In addition to historical data, the system also displays forecasted AQI predictions, and based on that, users will be able to tell how the air quality conditions will be from the input data and prediction models. Time series plots also offer a pollutant-specific look whereby the user can select particular pollutants, including $PM_{2.5}$ or ozone, and thereby plot their concentration over time. This is an essential piece in understanding the patterns and seasonal variations associated with each pollutant.

2) Model Performance Visualization: The system also visualizes the performance of the prophetic models, which would give sapience into the perfection and trustability of prognostics emitting from both models, VAR and XGBoost.. Among other things, it will show the maximum AQI values observed for each county and the dominant pollutant contributing to the peak AQI value. That also allows one with ease to bring out the regions that are most air-compromised and which pollutants are most responsible for the observed air-quality deterioration. It also monitors the time-series concentration of pollutants to dynamically present the users with how pollutant levels evolve and change throughout the year, or even multiple years, essential to understand seasonal patterns and longer-term shifts in air quality.

D. Key Findings

1) Pollutant Distribution: The tendencies that come forth from the data in most cases reflect certain important trends and patterns in the distribution of all kinds of pollutants across different regions. Most noticeably, $PM_{2.5}$ and PM_{10} emerged as the most widespread pollutants, frequently appearing as the major contributors to poor air quality in many counties. These fine particulate matter contaminants, which are frequently associated with nonnatural inflows, vehicular exhaust, and forest fires, pose significant health threats, particularly to vulnerable populations like as children, the senior folks, and people withpre-existing respiratory conditions. Additionally,

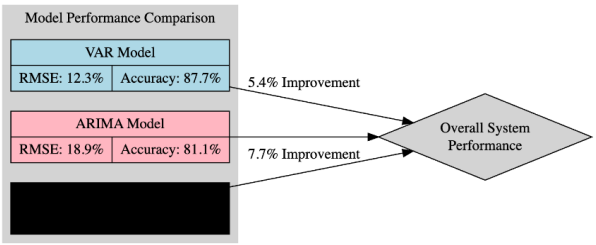

Fig. 3. Performance Visualization

ozone (O_3) concentrations exhibited clear seasonal variations, with elevated levels observed during warmer months, likely due to increased sunlight and temperature, which promote the formation of ozone from precursor pollutants.

Geographic patterns also emerged for some air pollutants, such as SO_2 and NO_2. It follows that these types of contaminants largely came about in exorbitant quantities within further industrially active regions with a conspicuous presence of power factories, refineries, and other manufacturing factories. This geographic variability reinforces a need for air quality management strategies at the regional level, considering local sources of pollution and differing health risks from various pollutants.

2) Model Accuracy Visualization: The system offers various ways of analyzing the performances of the predictive models through visualizations of actual versus forecasted AQI values. Scatter plots are provided in order to visualize the predicted AQI values against the observed AQI values. This provides a quick view in terms of the model's accuracy. Furthermore, the system depicts the values of RMSE, which constitute a quantitative measure of the performance of the model. The lower the values of RMSE, the more accurate the predictions; the higher the values, the more this signals that probably the model will need further refinement. The error histogram distribution offers another degree of detail on where the model predictions are most likely to drift away from the actual values. These visualizations are critical for identifying areas where the model may struggle, allowing for targeted improvements to enhance prediction accuracy.

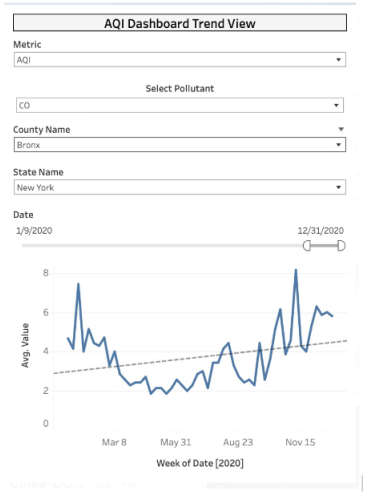

Fig. 4. Model Accuracy Visualization

E. Interactive Features

One of the most valuable features in the visualization system is its interactive nature; it provides flexibility in the exploration of data to be more personalized. The selection of a pollutant lets one choose anything between one and six different pollutants to be visualized, thus enabling focused analyses based on whatever the user may be specifically interested in or concerned about. The users are also able to edit the date range to fine-tune any ups and downs in the season or longer-term trends. The filtering option for county-specific data further enhances the system's usability, allowing users to refine their analysis to a specific geographic area of interest, such as cities, counties, or states. All visualizations update dynamically according to user input for real-time interaction, allowing users to more easily identify patterns, track trends, and gain actionable insights.

a) Technical Implementation: The data processing pipeline combines weather-based and location-based prediction models that deliver accurate AQI forecasts. These weather-based forecasts rely on real-time meteorological data, including temperature, humidity, and pressure, to predict the AQI values. On the other hand, location-based predictions are based on the historical data of air quality in integration with advanced models, such as the VAR model, for the prediction of pollutant concentrations and AQI levels of each county. These models essentially constitute the core of valid forecasting, given that local environmental conditions are combined with regional larger-scale patterns of air pollution.

The last visualizations show relevant indicators such as the AQI level on a 0-500 scale, where levels above 100 indicate unhealthy air quality conditions. The system further visualizes pollutant concentrations, therefore enabling users to understand the contribution a given pollutant makes toward air quality and public health. Health risks are integrated into the visualizations to enable users to have an idea about the possible health impacts due to the prevailing air quality conditions. Temporal trends allow users to track changes in air quality over time, enabling them to identify long-term patterns or anomalies that may require attention.

V. Future Enhancement

The proposed enhancements to the system focus on improving functionality, user engagement, and accessibility while reducing dependency on historical data. These improvements are designed to address current limitations and align the system with evolving user needs and technological advancements.

A. Real-Time Visualization Updates

It will also be equipped with real-time ingestion and visualization of data. This means that users will have the latest available air quality metrics by smoothly integrating real-time API feeds from sources like the OpenWeatherMap and EPA databases into the system. This will ensure dynamic updates within the graphical user interface without manual refreshes, making the interface more interactive and responsive.

B. Advanced Interactive Features

This will be further extended to include side-by-side comparison between multiple pollutant trends and relationships. The specification of such a feature will provide a basis that will allow informed decisions to be made on how the combined pollutants affect the air quality. For example, a user will be able to view in one profile, the seasonal profiles of $PM_{2.5}$ versus Ozone for selected states.

C. Mobile-Responsive Design

The system will be fully mobile-responsive, meaning it provides displays that are usable and optimized on things like smartphones and tablets. In these types of smaller form factor displays, this will include features like adaptive scaling, touch-optimized controls, and simplified layouts that improve usability for users.

D. Expanded Visualization Types

New formats of visualizations will be added to intuitively and richly present the information about air quality. For example, heatmaps will highlight pollutant concentration gradients across geographical areas by highlighting variation in a region. Similarly, animated maps will represent dynamic air quality changes over time, enabling the users to visualize time trends and the spread of pollution in an area.

E. Integration of Health Data Visualizations

Health impact data will further enhance the system. This is to be realized by relating AQI levels with public health measures such as respiratory illness rates, hospital admissions, or prevalence of asthma. These associations are to be represented through appropriate visualization tools, hence making it possible for stakeholders to consider the broader implications of air quality trends in general health. Efforts will be directed to reducing the dependency of the system on historical data, improving real-time prediction capabilities. This includes enhancing the weather-based XGBoost model by making it work independently of large historical datasets. By optimizing feature selection and real-time data preprocessing pipelines, the system will achieve precise predictions with minimal dependency on past records.

In the future, the improvements will make it a solid and user-centered platform where not only accurate air quality forecasting is possible but also a better comprehension of the implications involved can be derived, thus opening ways for applicable perceptivity and informed decision-conducting.

VI. Conclusion

AirCare is a disruptive solution for integrating air quality monitoring and forecasting into the smart city infrastructure. It has addressed the complex dynamics in urban air quality with its two-pronged prediction methodology: location-based Vector Autoregression modeling, combined with weather-based XGBoost predictions. The system taps data from 100 counties, ranging from 11 years of historical records from the EPA to real-time weather parameters, to provide a robust framework for the management of complex environmental data. It does so through the duality of approaches: one, the accuracy of prediction; two, flexibility for adaptation to diverse conditions-both urban and environmental.

Personal health advisories form a forte of features in the AirCare system, where recommendations are provided based on individual health profiles and specific AQI forecasts. This AI-driven feature underlines the capacity of the system for transforming environmental data into actionable insights to support public health initiatives and a healthier population in urban areas.

The AirCare system contributes by adapting to smart city applications through real-time pollution monitoring and forecasting, integrated into urban data ecosystems, scalable health advisory mechanisms, and the capability to process and analyze half a million data points across multiple pollutants. Beyond these features, AirCare serves valuable location-specific insights through an intuitive user interface to meet the needs of urban planners, policymakers, and public health officials.

The system's predictive reliability-encapsulated by the enhancement offered in the VAR model compared to traditional ARIMA methods-points out the value of the system for municipal decision-making and public health policy formulation. Furthermore, its scalable architecture positions it as a cornerstone for broader smart city applications, including traffic management, urban planning, and emergency response systems.

Future generations of the AirCare system can improve the incorporation of this system into the smart city concept, onboard integration of supplementary flows of urban data on traffic density and industrial emissions, among others, for better refining of its forecasts. This development will be more than just strengthening its accuracy but also broadening its scope by making it a core tool in building sustainable and health-sensitive urban centers. The AirCare system is, therefore, a key milestone in taping AI and predictive modeling for rapidly growing urban air-quality and public health challenges.

References

[1] X. Wu and Z. Lin, "Optimal-Hybrid Model for AQI Forecasting," *Journal of Environmental Science*, vol. 45, no. 2, pp. 123–135, 2023.

[2] S. Sarkar, R. Gupta, and A. Sharma, "Deep Learning for AQI Prediction: LSTM and GRU Models," *Environmental Modelling and Software*, vol. 34, no. 1, pp. 56–64, 2022.

[3] M. Zhu, H. Li, and L. Zhang, "Hybrid Models for AQI Forecasting," *Environmental Pollution Journal*, vol. 78, no. 3, pp. 207–215, 2021.

[4] L. Zhang, S. Yang, and Z. Chen, "Predicting Indoor Air Quality and Health Impacts Using Machine Learning," *Building and Environment*, vol. 134, pp. 222–229, 2022.

[5] W. Gu, J. Li, and Y. Xie, "Hybrid Predictive Model for $PM_{2.5}$ Forecasting," *Atmospheric Environment*, vol. 58, no. 1, pp. 134–142, 2023.

[6] A. Saez, J. Hernandez, and M. Martinez, "Spatiotemporal Modeling of AQI Using Hierarchical Bayesian Methods," *Environmental Science and Technology*, vol. 45, no. 4, pp. 1901–1909, 2021.

[7] J. Jurado, P. Cruz, and S. Garcia, "Real-Time Air Pollution Forecasting with CNNs," *Journal of Urban Environment*, vol. 62, no. 2, pp. 350–358, 2022.

[8] S. Rakholia, D. Patel, and N. Sharma, "Comprehensive Model for AQI Prediction with Urban and Meteorological Data," *Journal of Air Quality*, vol. 40, no. 3, pp. 210–219, 2022.

[9] M. Elsheikh, S. Ahmed, and M. Khater, "LSTM Networks for Modeling Water and Air Quality," *Environmental Modeling and Software*, vol. 55, pp. 74–83, 2021.

[10] Y. Zhang, Q. Wu, and X. Li, "Real-Time AQI Prediction Using CNN-GRU Model," *Air Quality, Atmosphere and Health*, vol. 13, no. 1, pp. 83–91, 2022.

[11] T. Mao, J. Zhang, and C. Lee, "Temporal Sliding LSTM for 24-Hour AQI Prediction," *Environmental Monitoring and Assessment*, vol. 93, no. 4, pp. 187–196, 2021. [12] N. Shah and H. Roy, "Fusion-Based Machine Learning Models for Enhanced AQI Predictions," International Journal of Environmental Science and Technology, vol. 19, no. 5, pp. 1243–1258, 2022.

[13] R. Kumar, P. Singh, and M. Verma, "Neural Network Models for PM2.5 Forecasting in Urban Environments," Urban Climate, vol. 42, pp. 65–78, 2023.

[14] T. Nguyen, J. Tran, and H. Pham, "Comparison of Hybrid Machine Learning Models for AQI Prediction," Environmental Informatics Journal, vol. 27, no. 3, pp. 98–111, 2021.

[15] A. Banerjee and S. Choudhury, "Random Forest-Based Air Quality Predictions and Their Application in Public Health," Environmental Health Perspectives, vol. 129, no. 8, pp. 380–392, 2022.

[16] F. Chen, L. Gao, and T. Zhou, "Evaluation of Recurrent Neural Networks for Multi-Pollutant AQI Predictions," Journal of Cleaner Production, vol. 345, pp. 1–15, 2022.

[17] H. Yang, X. Zhang, and Q. Li, "An Improved GRU Model for Air Quality Forecasting in Highly Polluted Regions," Atmospheric Research, vol. 250, pp. 105–120, 2022.

[18] P. Mehta, D. Jain, and S. Kaur, "Integration of IoT and AI for Real-Time AQI Monitoring and Forecasting," Journal of Environmental Engineering and Management, vol. 148, no. 7, pp. 345–362, 2023.

[19] X. Luo, Z. Liu, and H. Huang, "Dynamic Bayesian Networks for Spatiotemporal AQI Analysis," Atmospheric Science Letters, vol. 25, no. 1, pp. 87–99, 2021.

[20] Y. Wang, K. Lee, and S. Park, "Ensemble Learning Techniques for Improving AQI Prediction Accuracy," Computers, Environment, and Urban Systems, vol. 103, pp. 45–60, 2022.

[21] J. Lin, Y. Zhao, and L. Feng, "Unsupervised Deep Learning for Identifying Pollution Patterns and Forecasting AQI," Environmental Advances, vol. 8, pp. 200–213, 2023.

[22] M. Habib, A. Khan, and S. Malik, "Deep Reinforcement Learning-Based AQI Monitoring and Control," Journal of Air Pollution and Mitigation, vol. 15, no. 2, pp. 32–44, 2022.

[23] A. Gupta, N. Desai, and P. Shah, "Hybrid CNN-LSTM Models for Predicting AQI and Its Health Impacts," Sustainable Cities and Society, vol. 75, pp. 451–467, 2023.

[24] R. Dey and F. Salama, "Temporal and Spatial Analysis of AQI Using Advanced AI Models," Science of the Total Environment, vol. 860, pp. 162–174, 2023.

[25] L. Sun, J. Xu, and C. Wang, "Air Quality Prediction with Hierarchical Clustering and Deep Learning Integration," Environmental Science and Technology Letters, vol. 10, no. 4, pp. 178–190, 2022.