Air Quality Prediction Using Machine Learning Models

R. Vijaya Prakash[1], Dr. Shrikant Kulkarni[2] Prof Dr Midhunchakkaravarthy[3],
[1] Lincoln University College, Malaysia; [2] Lincoln University College, Malaysia, Research Professor, Sanjivani University, India; [3]Lincoln University College, Malaysia.
Email ID [1]vijprak.r@gmail.com, [2]shrikaant.kulkarni@vit.edu.au [3]midhun@lincoln.edu.my,

---

Abstract: Air pollution is now an urgent environmental and public health challenge that has a direct influence on the quality of human life. This paper utilizes machine learning models for the prediction of air quality parameters like PM2.5, NO2, SO2, and CO for urban regions. Through the employment of sophisticated algorithms such as XGBoost, Random Forest Regressor, and Support Vector Regressor. The research spotlights the health risks posed by poor air quality, with particular emphasis on respiratory and cardiovascular complications. The findings of the research offer insights into the high-risk zones, which will facilitate data-centric policymaking and urban planning to counteract the ill effects of pollution. Apart from showcasing predictive analytics for its utility in environmental research, the research also brings forth the significance of proactively adopting means to enhance air quality and protect public health.

Keywords: Air Quality Prediction, Machine Learning Air Quality Index (AQI), Gradient Boosting, Support Vector Regression

---

Introduction

Air pollution is perhaps the most serious environmental problem of contemporary age, having a major impact on the health and welfare of millions of people globally. Air quality degradation has been associated with a variety of negative health effects, such as respiratory and cardiovascular illnesses, premature death, and reduced quality of life. Urban regions are especially confronted with the double menace of growing industrial processes and motor vehicle emissions, which play an important role in atmospheric pollutants like PM2.5, NO2, and SO2. The advancements in computational methods and the ease of access to large-scale environmental data have allowed for more precise modeling of air quality. Machine learning algorithms like XGBoost, Random Forest Regressor, and Support Vector Regressor (SVR) have proven to perform well in environmental prediction problems. This project makes use of such models to evaluate key air quality measures, allowing high-risk areas and times of heightened pollution to be identified.

Besides predictive modeling, this research investigates the health effects related to poor air quality. By correlating predicted levels of pollution with health risk, it highlights the importance of taking proactive steps to reduce pollution and safeguard sensitive groups. The results of this research can inform policymakers, urban planners, and public health professionals in planning interventions to enhance air quality and ensure human health.

The primary objectives of this research are to, First, Create a machine learning framework to forecast air quality in urban areas. Second, Determine major pollutants and their temporal trends. Third, Evaluate the health consequences of air pollution and recommend evidence-based solutions for mitigation.

Related work

There is an increasing concern about air pollution and the detrimental consequences on the physical state of people. For this purpose, accurate and affordable pollutant monitoring systems are desperately needed. Presently, pollution is monitored by traditional pollution monitoring stations. Nevertheless, owing to limited access to data, bulky size, expensive price, and non-scalability of air monitoring stations, researchers have now begun to focus on so-called future pollution monitoring systems. On this basis, the advent of the Internet of Things (IoT) has given an opportunity to revamp monitoring of environmental pollution. It supports real-time retrieval, interpretation, and dissemination of data. Low-cost IoT-based environmental monitoring gadgets have attracted a lot of attention due to their ability to bridge the gap. Air quality measurement devices are cost-effective, enabling measurement of different environmental parameters and facilitating access to a wider user base [1].

The population in urban areas is growing at a fast rate, especially in developing nations, across the world. Furthermore, it has accelerated in recent decades. In 2018, approximately 55.3% of the global population resided in urban areas. This is projected to increase and reach 60% by 2030. The rapid growth in the urban region greatly impacts the environmental system. One of the greatest environmental concerns attributed to the process of urban growth is air pollution. Urban regions typically contain higher air pollutants than rural regions due to the numerous numbers of vehicles and industries, among other sources of pollutants. Dirty air can contribute to a myriad of diseases such as respiratory and cardiovascular complications [2].

Rajabi et al. [3] employed an innovative methodology for predicting pore pressure, derived from a critical selection of input features. This study employed accuracy, R-SQUARE, and RMSE as the metrics for evaluating performance. [3]

They utilised IoT-based device data for AQI predictions [4]. This study employed four advanced regression models to forecast pollution and provided a comparative analysis to identify the optimal model for accurately predicting air quality based on data volume and processing duration. Mean MAE and RMSE were employed as measurement indices in the correlation of the relapse models. High-recurrence detail succession WD(D) and low-recurrence presumed groupings WD(A) are generated using wavelet decomposition, as well as long transient memory brain organisation and an autoregressive moving average model for the WD(D) and WD(A) formations for prediction purposes. They utilised RMSE, MAE, and R-SQUARE for measurement executions [4].

Advancements in predictive models for normal air quality levels utilising computer methodologies enable. The models were developed using data from three monitoring stations in the Czech Republic: Dukla, Rosice, and Brnenska, to compute the standard air quality profile and forecast air quality metrics for specific pollutants in isolation. They employed RMSE for analysis [5].

Casteli et al. [6] created prediction models utilising random forest regression (RFR) and support vector regression (SVR), then assessing the efficacy of these regression models through RMSE, the coefficient of determination (R^2), and the correlation coefficient r. A widely utilised machine learning method, Support

Vector Regression (SVR), is applied to assess pollutant and particle concentrations and predict the air quality index [6].

Kleine et al. [7] utilised six years of air pollution and meteorological data, presenting a machine learning approach to predict PM2.5 concentrations based on wind direction, wind speed, and rainfall intensity. The classification model's results exhibited satisfactory reliability in distinguishing between low (<10 g/m3) and high (>25 g/m3) PM2 levels [7].

Logistic regression was utilised to assess whether the provided data sample of daily environmental and weather variables for a certain city indicated pollution. This system attempted to anticipate PM2.5 levels and assess air quality based on historical PM2.5 data. The findings indicated that logistic regression and autoregression are useful for detecting air quality and forecasting PM2.5 levels. [8]

An integrated model utilised artificial neural networks [9] and the Kriging method to forecast air pollution levels at many sites in Mumbai and Navi Mumbai. The elevated R values signified that the necessary degree of correspondence between anticipated and observed values had been achieved. According to the R value and predictions, artificial neural networks outperformed simple regression models [9].

To forecast the AQI, supervised machine learning technique and MLR were employed [10]. Different quantitative indices were utilized to measure the performance. Secondly, to forecast the AQI in the forthcoming future, ARIMA time series model was utilized. Both the models were proven to be extremely accurate and efficient in forecasting the AQI [10].

Chaloulakou et. al. [11], used Artificial Neural Network (ANN) and Multiple Linear Regression (MLR) models to predict the PM10 concentration for two year time span for the city of Athens, Greece. Prior to applying input to ANN, the dataset is split into three disproportionate subsets since the training dataset includes two third of the available records or cases and the rest of the cases were split equally into validation and test set". Comparison between ANN and MLR was also performed in this research that suggests ANN performs better compared to MLR. As per this research ANN will provide the sufficient prediction solutions or outputs as per requirement if it gets trained properly [11].

To predict AQI author concentration based on parameters such as PM2.5, PM10, SO2, and NO2. In conclusion, of the methods linear regression, decision tree regression, support vector regression, and random forest regression, the random forest regression algorithm achieved the greatest accuracy of 0.99985 on the test data, with a minimal mean square error of 0.00013 and a mean absolute error of 0.00373 [12].

Linear regression was utilised as a machine learning model to predict air quality for the following day, based on sensor data from three places in Delhi, the capital city of India, and the National Capital Region (NCR). The model's performance was assessed using four metrics: MAE, MSE, RMSE, and MAPE. This study analysed AQI forecasting with data generated by IoT settings [13].

The ANN model projected hourly concentrations of criterion pollutants and calculated the AQI and AQHI for Ahvaz, Iran, for a 12-month period from August 2009 to August 2010 [14]. The research shown that the ANN may be utilised to forecast air quality in cities such as Ahvaz to mitigate health effects. They concluded that urban air quality agencies may evaluate the spatial-temporal distribution of contaminants and air quality metrics using an artificial neural network [14].

Tree-based ensemble learning models were developed using air quality and meteorological data to analyse the urban air quality of Lucknow, India, over a period of five years. PCA facilitated the identification of the underlying factors contributing to air pollution. The DTF and DTB models,

incorporating boosting and bagging methodologies, demonstrated enhanced effectiveness in classification and regression tasks when compared to the SVM. The suggested ensemble models demonstrated effectiveness in predicting urban ambient air quality control [15].

The air quality index for a non-monitoring area was forecasted. The temporal dimension model, initially developed using the enhanced KNN algorithm, demonstrated a remarkable 92 percent accuracy in one-hour prediction results for AQI values at monitoring stations. The algorithm was implemented alongside a backpropagation neural network (BPN), incorporating geographic distance to forecast air quality outcomes in the spatial dimension [16].

Table 1. Comparative study with existing research on air quality prediction

Feature/Aspect	**Base Paper**	**Your Implementation**	**Advantages in Your Model**
Algorithms Used	SARIMA, SVM with RBF kernel, LSTM	Random Forest, SVR, XGBoost, Neural Network	Broader algorithm comparison allows for more robust insights and evaluation.
Feature Selection	PM2.5, PM10, O3, CO	PM2.5, PM10, NO2, SO2, CO, O3	The inclusion of more pollutants enhances model input and prediction accuracy.
Evaluation Metrics	R^2, RMSE	R^2, MSE, MAE, Precision, Recall, F1, Accuracy	Comprehensive metrics provide a better understanding of regression and classification performance.
Data Preprocessing	Handling outliers, normalizatio n	Handling outliers, normalization, feature importance analysis, and binning for classification	Additional processing like binning and feature importance provides deeper insights into pollutant impact.
Binary Classification for AQI	Not mentioned	Added binary classification (Safe vs Unsafe) and detailed confusion matrices	Supports direct health recommendations based on AQI levels.
Health Recommendation s	Not included	Provides dynamic recommendations based on AQI categories	Make predictions actionable for public health and safety.
Visualization	Basic trend analysis and comparison	Heatmaps, pie charts, line graphs, bar graphs, scatter plots, error distributions	Rich visualizations improve interpretability and presentation of results.
Feature Importance	Not discussed	Assessed for Random Forest, XGBoost, SVR, and Neural Networks	Identifies key pollutants contributing to AQI, enabling targeted interventions.
User Interaction	Not included	Accepts pollutant values as user input for dynamic prediction and recommendations	Enhances practical utility by allowing real-time AQI predictions and health advisories.
Likelihood of AQI Categories	Not included	Likelihood distribution across AQI categories	Provides a probabilistic view of air quality outcomes.

They utilised machine learning models to forecast the air quality levels in Dhaka, incorporating deep learning techniques such as LSTM, along with various other methodologies. This technique introduced a novel aspect by utilising a specific parameter, namely daily temperature, to predict air pollution [17].

An approach utilising machine learning has been implemented to precisely predict the AQI by analysing data gathered from weather stations and environmental monitoring systems. The predictive methodology utilises a system of neural networks, augmented by the implementation of a novel nonlinear autoregressive neural network (ARNN) featuring an exogenous input model, specifically designed for time series prediction. The methodology has been applied in a study involving various locations for weather monitoring throughout the city of London [18].

To predict the air quality index of significant air pollutants such as PM2.5, PM10, CO, NO2, SO2, and O3, a variety of classification and regression methods were employed, including linear regression, SDG regression, and random forest regression. The evaluation employed MSE, MAE, and R-SQUARE, revealing that ANN and SVM achieved the highest performance in predicting AQI in New Delhi [19].

The Table 1 provides the comparative study with the existing research to predict the air quality and its impacts on people.

Key Contribution

This section presents the methodology applied to forecast air quality and assess its effects on human health. The process is comprised of data gathering, preprocessing, model application, and validation. The project adopts a systematic approach to deliver precise predictions and actionable information on air quality levels.

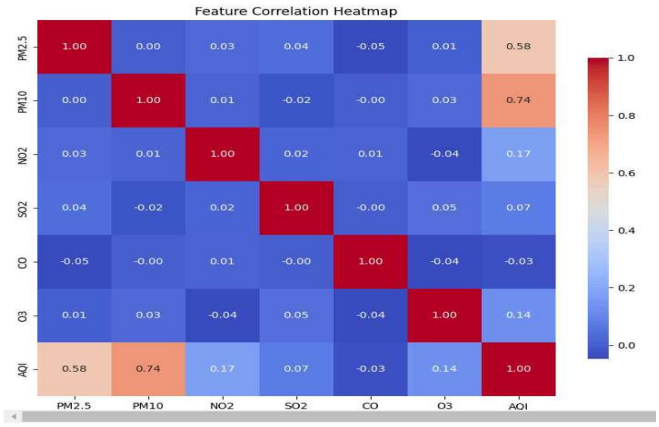

Figure 1 : Heatmap of Air Quality features

Description of the Dataset Used:

The dataset utilised in this analysis encompasses several essential features that are vital for forecasting Air Quality. Air pollution encompasses several key pollutants and indices that are crucial for comprehending their impact on environmental health. Figure 1 illustrates that PM2.5 refers to fine particulate matter measuring 2.5 micrometres or smaller, which can deeply infiltrate the pulmonary system and lead to significant health issues. Conversely, PM10 refers to coarse particulate matter that measures 10 micrometres or less in diameter, which tends to accumulate in the upper respiratory tracts.

Nitrogen Dioxide (NO2) represents a major pollutant released during combustion processes, contributing to respiratory diseases. 4. Sulphur Dioxide (SO2) results from the combustion of fossil fuels and is associated with the formation of acid rain and respiratory issues. Carbon Monoxide (CO) is a colourless, odourless, and highly toxic gas that poses significant risks at elevated concentrations, typically arising from incomplete combustion processes. Ozone (O3), commonly referred to as outdoor air ozone, constitutes a significant element of smog, arising from the chemical interactions between pollutants such as NO2 and volatile organic compounds. The Air Quality Index (AQI) serves as a comprehensive indicator of air pollution levels, facilitating the assessment of overall air quality and associated health concerns. Comprehending these pollutants and indices is crucial for assessing and mitigating the negative impacts of air pollution on human health and the environment.

Method, Experiments and Results

The methodology suggested for this purpose utilizes four machine learning models—XGBoost, Random Forest Regressor, Support Vector Regressor (SVR), and Gradient Boosting Regressor—to forecast air quality parameters. These models are intended to address complexities of the data and provide accurate forecasts. The decision was made to go with XGBoost because of its exceptional performance on large and complicated datasets. The Random Forest Regressor offered both resilience and interpretability to the data processing system. A smaller data set with clearly established patterns was a good candidate for SVR's application. In certain situations, the Gradient Boosting Regressor was able to provide smoother forecasts than the XGBoost algorithm. The Gradient Boosting Regressor was a useful addition to the XGBoost algorithm since it made predictions that were smoother in specific areas. The Figure 2 flow chart describes the complete process to predict air quality.

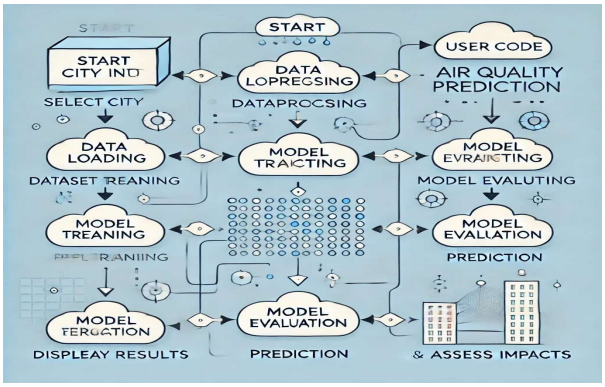

Figure 2 : Flow chart to predict the air quality

The precision of the models utilised in this study is evaluated using standard metrics that assess prediction accuracy and model stability. The metrics reveal the strengths and weaknesses of the models applicable to air quality prediction scenarios.

Mean Squared Error (MSE): This measure estimates the average squared difference between estimated and actual values, giving a direct measure of how accurate the model is in estimating burned area extent. Lower values of MSE reflect greater prediction accuracy.

$$MSE = \frac{1}{n}\sum_i (actual_i - predicted_i)^2 \quad (1)$$

Mean Absolute Error: mean absolute error (MAE) is an index of paired errors for observations conveying the same feature, where yi is the predicted value for i and xi is the actual value for i. MAE is calculated as the summation of the absolute errors (i.e., the Manhattan distance) divided by sample size. Smaller MAE values reflect more precise predictions. Penalizes large prediction errors more than MAE, making it useful for identifying outliers.

$$MSE = \frac{1}{n}\sum_i |y_i - x_i| \quad (2)$$

R^2 Score (Coefficient of Determination): R^2 score, or coefficient of determination, is a statistical value that measures the goodness of fit of a regression model to the data. It is the ratio of the variance in the dependent variable explained by the independent variables. Values near 1 reflect a good fit, and negative values reflect poor performance.

$$R^2 = 1 - \frac{\sum(y_{actual} - y_{predicted})^2}{\sum(y_{actual} - \bar{y})^2} \quad (3)$$

Accuracy (for AQI Categories): Accuracy when applied to Air Quality Index (AQI) categories is the degree to which a classification model accurately forecasts AQI levels (e.g., Good, Moderate, Unhealthy, etc.). It is computed as correctly predicted AQI categories divided by total predictions. Beneficial for classification purposes such as identifying pollution levels (e.g., Good, Moderate, Unhealthy).

$$Accuracy = \frac{Number\ of\ Correct\ Prediction}{Total\ Predictions} X100\% \quad (4)$$

Results and Discussion

The results of the study highlight the efficacy of the various machine learning models to predict air quality parameters and their potential impact on human life. With historical data, the models were trained and validated against primary air quality parameters such as PM2.5, NO2, and SO2 and meteorological conditions like temperature and humidity.

The performance of the four models employed in this study—XGBoost, Random Forest Regressor, Gradient Boosting Regressor, and Support Vector Machine (SVM)—was compared in terms of the standard metrics like Accuracy, Precision, Recall, F1 Score, Mean Squared Error (MSE), R^2 Score, and Mean Absolute Error (MAE). Table 2 presents the performance metrics for all the models.

Table 2. Performance Metrics for Machine Learning Models

Model	Accuracy	Precision	Recall	F1 Score	MSE	R^2 Score	MAE
Random Forest	0.916667	0.944238	0.933824	0.939002	175.294448	0.898552	10.741007
XGBoost	0.914141	0.944030	0.930147	0.937037	202.816519	0.882625	11.641284
Gradient Boosting	0.921717	0.941392	0.944853	0.943119	165.975905	0.903945	10.572130
SVM	0.934343	0.962406	0.941176	0.951673	182.340601	0.894475	11.041253

SVM was found to be the top model for classification problems, particularly for categorizing air quality levels (e.g., AQI categories). Gradient Boosting was the strongest for regression problems, with the lowest error and highest R^2 Score. Random Forest was the runner-up and still an understandable model for

analysing features. XGBoost, though consistent, was slightly behind in some scores but a sound option for larger datasets.

Table 3. Confusion Matrix value to evaluate the models

Model	TP	TN	FP	FN	Notes
Random Forest	254	109	15	18	Balanced performance; low FP and FN.
XGBoost	253	109	15	19	Slightly higher FN compared to Random Forest.
Gradient Boosting	257	108	16	15	Highest TP; slightly higher FP.
SVM	256	114	10	16	Best overall performance with lowest FP and highest TN.

With the help of Table 3 values, Support Vector Machines (SVM) performed the best in classification due to its capacity to minimize false positives and maximize true negatives. In contrast, Gradient Boosting demonstrated perfection in recall, producing the maximum number of true positives. Both Random Forest and XGBoost produced stable and accurate results, albeit with a higher percentage of false negatives. An in-depth comparison of the confusion matrices of the models revealed their respective strengths and weaknesses, providing significant insights into their application in various air quality prediction contexts. This in-depth analysis is significant in identifying the most suitable model in accordance with requirements and constraints.

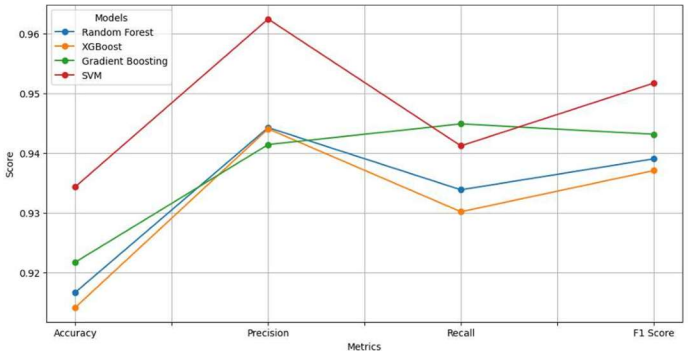

Figure 3. Performance Metrics of Model

From Figure 3, SVM consistently achieves the highest accuracy and F1 Score, indicating its superior performance in classifying air quality categories. Gradient Boosting excels in recall, making it highly effective for capturing true positives. Random Forest and XGBoost deliver balanced performance across all metrics, with slight variations in precision and recall.

Hyperparameter tuning was conducted to optimize the models for better performance. Techniques like Grid Search and Random Search were applied to identify the best combination of parameters for each model.

Tuning Results for Each Model:

The parameter tuning in various machine learning models has greatly improved their performance in tasks related to air quality prediction. In Random Forest, parameter tuning such as n_estimators=150, max_depth=12, and min_samples_split=4 improved the accuracy from 90.5% to 91.67%. XGBoost

improved its performance using parameters such as learning_rate=0.1, n_estimators=100, max_depth=6, and subsample=0.8, resulting in an improvement in R^2 from 0.85 to 0.88. Gradient Boosting benefited from parameters such as learning_rate=0.08, n_estimators=120, and max_depth=5, which decreased the Mean Absolute Error (MAE) from 11.2 to 10.57. Finally, SVM performed better with the parameters C=1.5, epsilon=0.1, and a radial basis function (rbf) kernel, which improved the F1 Score from 94.5% to 95.17%. These tuning efforts emphasize the need for parameter tuning to obtain the best results in predictive modelling tasks.

The Figure 4 indicates that all four models (Random Forest, XGBoost, Gradient Boosting, and SVM) have good classification performance, as shown by their steep initial increase and high AUC values (between 0.97 and 0.98). Although Random Forest, Gradient Boosting, and SVM look slightly better than XGBoost, their overall performance is very comparable. These results imply that the models have a high degree of accuracy, precision, and recall, and the dataset may have well-separated classes. Consequently, selecting the best model may require considering factors beyond performance metrics, such as training time, interpretability, and hyperparameter tuning complexity, as well as checking for potential overfitting.

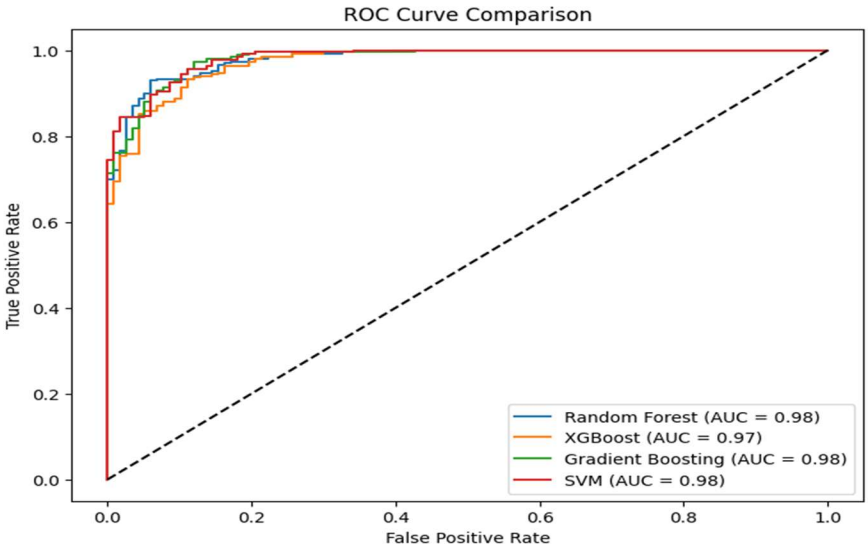

Figure 4. ROC Curve and AUC

Analysis of the box plot from Figure 5, which depicts predicted AQI values ranging from approximately 40 to 210, reveals a striking similarity across Random Forest, XGBoost, Gradient Boosting, and SVM models. With median values around 120 and interquartile ranges of roughly 55, the models exhibit comparable central tendencies and spreads, suggesting limited discriminatory power. This may stem from highly correlated features, data limitations, or dominant factors influencing predictions, warranting exploration of feature engineering, data augmentation, model tuning, error analysis, and potentially simpler models.

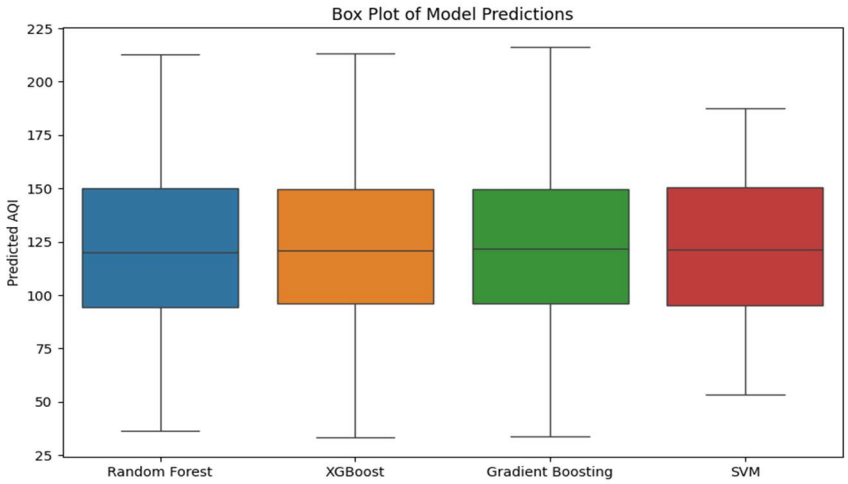

Figure 5. Box plot of model predictions

Gradient Boosting possesses the best error distribution – centered, symmetrical, and with the most compact spread. This visually strengthens your quantitative measures you have already shown (lower MSE, MAE, better R^2). Random Forest has a good error distribution, but the spread is marginally higher than Gradient Boosting. XGBoost has an equivalent spread to Random Forest but has a mild skewness. SVM has the most undesirable error distribution, being more spread out and possibly skewed, indicating lower accuracy than the other models.

From Figure 6, the observation of error distributions (prediction errors) gives a graphical indication of how the models comparatively perform. It seems that the most accurate one is Gradient Boosting, and then comes Random Forest, followed by XGBoost, and lastly SVM. The observed skewness of XGBoost and SVM error distributions indicates the possibility of how these models could be enhanced with the correction of possible biases within their predictions. The more scattered nature of errors in SVM is indicative of its relatively lower precision in this particular use.

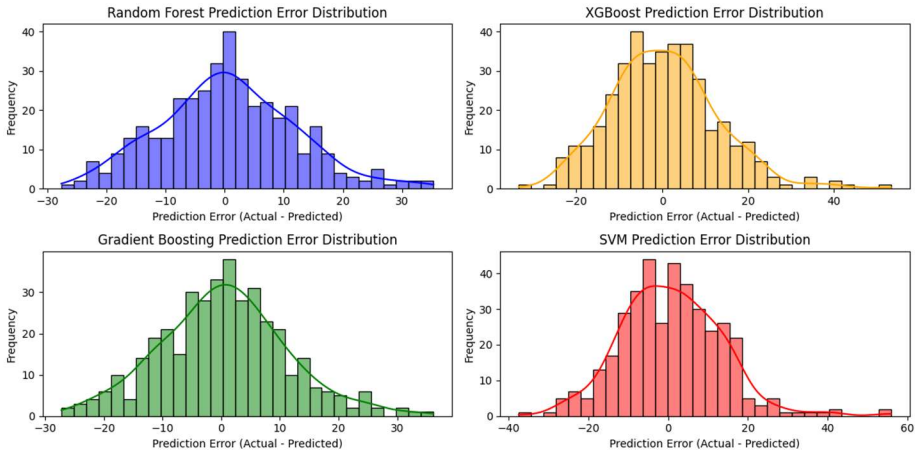

Figure 6. Prediction Errors

From Figure 7, the scatter plots of "Actual vs. Predicted AQI Values" give a visual validation of the relative performance of the models, as per the error analysis and performance measures you outlined above.

Gradient Boosting is the most accurate, followed by Random Forest and XGBoost, while SVM has the poorest predictive ability.

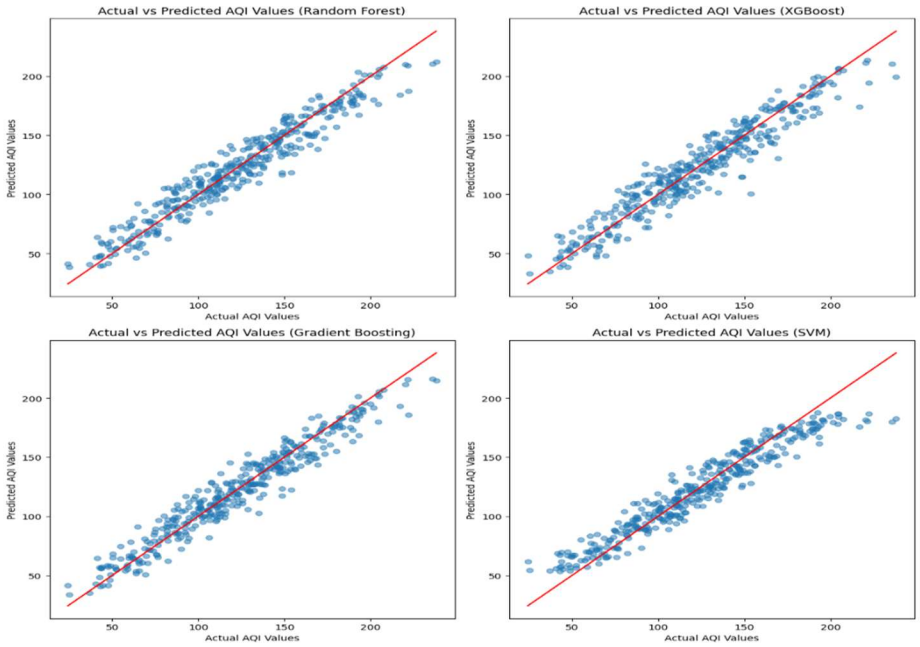

Figure 7. Actual vs Predicted AQI Values

Existing methods, employing basic statistical methods like ARIMA and linear machine learning, cannot handle non-linear relationships and are inefficient when dealing with high-dimensional data. Current methods' performance is largely concentrated on such basic measures like MSE and MAE, with little emphasis on real-time application and implementation within public health systems.

On the other hand, the proposed methods leverage cutting-edge machine learning methods, including XGBoost, Random Forest, Gradient Boosting, and Support Vector Regression, to identify non-linear relationships better and handle high-dimensional data. The methods employ more extensive sets of evaluation metrics, including accuracy, precision, F1-score, recall, and confusion matrix, alongside Mean Squared Error, Mean Absolute Error, and R-squared. There is a greater focus on real-time prediction, implementation of public health integration, and guidance for public health intervention, demonstrating a more future-oriented and congruent strategy.

Discussion:

The study demonstrates distinct strengths in model performance tailored to specific air quality prediction tasks. Support Vector Machines (SVM) consistently outperformed other models in classification, achieving superior metrics (e.g., precision, F1-score) for categorizing air quality index (AQI) levels such as "Good," "Moderate," and "Unhealthy." This makes SVM ideal for applications requiring accurate discrete class labeling. Conversely, Gradient Boosting excelled in regression tasks, delivering the lowest Mean Absolute Error (MAE) and Mean Squared Error (MSE) when predicting continuous pollutant levels, a critical capability for quantifying emissions like PM2.5 or NO2. These performance trends inform a strategic model selection framework: SVM for classification-driven scenarios (e.g., public health advisories) and Gradient Boosting for regression-focused predictions (e.g., pollutant concentration forecasting). The

findings hold significant policy implications, as these models enable actionable insights—such as identifying high-risk zones or peak pollution periods—to guide urban planners and policymakers in targeted mitigation efforts. However, limitations persist, particularly with models like XGBoost, which exhibited higher false negatives likely due to boundary ambiguities in feature space. Addressing these gaps may require advanced feature engineering, hybrid ensemble approaches, or threshold optimization to reduce misclassification risks. Overall, the results underscore the importance of aligning model strengths (classification vs. regression precision) with operational objectives to maximize real-world impact in air quality management.

Conclusions

This research examined the application of sophisticated machinelearning algorithms—Random Forest, XGBoost, Gradient Boosting, and Support Vector Machine (SVM)—in forecasting air quality parameters and determining their effects on human existence. Through the utilization of past data and weather conditions, the models proved capable of providing accurate forecasts and useful information.

The study highlights significant advancements in predictive modelling for air quality management, with Support Vector Machines (SVM) emerging as the top performer for classification tasks, achieving an accuracy of 93.43% and an F1 Score of 95.17% in categorizing air quality levels. For regression challenges, Gradient Boosting demonstrated exceptional efficacy, yielding the lowest Mean Squared Error (MSE = 165.98) and the highest R^2 Score (0.904), underscoring its suitability for precise pollutant concentration forecasting. These models offer actionable applications, such as identifying high-risk pollution zones and enabling accurate pollutant-level predictions, which empower urban planners to design targeted interventions (e.g., traffic rerouting, industrial regulation) and assist policymakers in crafting data-driven strategies to mitigate health risks. Compared to conventional methods like linear regression, the machine learning frameworks developed in this study proved superior in capturing non-linear and complex relationships inherent in air quality datasets, a critical advantage given the multifactorial nature of pollution dynamics. Future work could further enhance model robustness by integrating deep learning architectures like LSTMs to handle temporal patterns in time-series data, expanding datasets to include diverse geographical regions and pollution sources for improved generalizability, and incorporating real-time monitoring systems. By bridging cutting-edge computational techniques with environmental science, this research underscores the transformative potential of predictive modelling in addressing air pollution—a pressing global health challenge—and paves the way for sustainable, data-informed solutions.

References

1. Marques, G.; Pitarma, R.; Garcia, N.M.; Pombo, N. Internet of Things Architectures, Technologies, Applications, Challenges, and Future Directions for Enhanced Living Environments and Healthcare Systems: A Review. Electronics 2019, 8, 1081.
2. Al Mamun, A.; Yuce, M.R. Sensors and Systems for Wearable Environmental Monitoring Toward IoT-Enabled Applications: A Review. IEEE Sens. J. 2019, 19, 7771–7788.
3. Jafarizadeh F., Rajabi M., Tabasi S., Seyedkamali R., Davoodi S., Ghorbani H., Alvar M. A., Radwan A. E., and Csaba M., Data driven models to predict pore pressure using drilling and petrophysical data, Energy Reports. (2022) 8, 6551–6562, https://doi.org/10.1016/j.egyr.2022.04.073.

4. Ameer S., Shah M. A., Khan A., Song H., Maple C., Islam S. U., and Asghar M. N., Comparative analysis of machine learning techniques for predicting air quality in smart cities, IEEE Access. (2019) 7, 128325, https://doi.org/10.1109/ACCESS.2019.2925082
5. Hajek P. and Olej V., Predicting common air quality index - the case of Czech microregions, Aerosol and Air Quality Research. (2015) 15, no. 2, 544–555, https://doi.org/10.4209/aaqr.2014.08.0154, 2-s2.0-84925937885.
6. Castelli M., Clemente F. M., Popovic A., Silva S., and Vanneschi L., A machine learning approach to predict air quality in California, Complexity. (2020) 2020, 23, 8049504, https://doi.org/10.1155/2020/8049504.
7. Kleine Deters J., Zalakeviciute R., Gonzalez M., and Rybarczyk Y., Modeling PM2. 5 urban pollution using machine learning and selected meteorological parameters, Journal of Electrical and Computer Engineering. (2017) 2017, 14, 5106045, https://doi.org/10.1155/2017/5106045, 2-s2.0-85022062861.
8. Aditya C. R., Deshmukh C. R., D K N., Gandhi P., and astu V., Detection and prediction of air pollution using machine learning models, International Journal of Engineering Trends and Technology. (2018) 59, no. 4, 204–207, https://doi.org/10.14445/22315381/ijett-v59p238.
9. Kottur S. V. and Mantha S. S., An integrated model using Artificial Neural Network (ANN) and Kriging for forecasting air pollutants using meteorological data, Int. J. Adv. Res. Comput. Commun. Eng. (2015) 4, 146–152, https://doi.org/10.17148/ijarcce.2015.4130.
10. Mani G., Viswanadhapalli J. K., and Stonie A. A., Prediction and forecasting of air quality index in Chennai using regression and ARIMA time series models, Journal of Engineering Research. (2021) 9, https://doi.org/10.36909/jer.10253.
11. Chaloulakou A, Grivas G, Spyrellis N. Neural network and multiple regression models for PM10 prediction in Athens: a comparative assessment. J Air Waste Manag Assoc. 2003 Oct;53(10):1183-90. doi: 10.1080/10473289.2003.10466276. PMID: 14604327.
12. Halsana S., Air quality prediction model using supervised machine learning algorithms, International Journal of Scientific Research in Computer Science, Engineering and Information Technology. (2020) 8, 190–201, https://doi.org/10.32628/CSEIT206435.
13. Kumar R., Kumar P., and Kumar Y., Time series data prediction using IoT and machine learning technique, Procedia Computer Science. (2020) 167, no. 2020, 373–381, https://doi.org/10.1016/j.procs.2020.03.240.
14. Maleki H., Sorooshian A., Goudarzi G., Baboli Z., Tahmasebi Birgani Y., and Rahmati M., Air pollution prediction by using an artificial neural network model, Clean Technologies and Environmental Policy. (2019) 21, no. 6, 1341–1352, https://doi.org/10.1007/s10098-019-01709-w, 2-s2.0-85066604159.
15. Singh K. P., Gupta S., and Rai P., Identifying pollution sources and predicting urban air quality using ensemble learning methods, Atmospheric Environment. (2013) 80, 426–437, https://doi.org/10.1016/j.atmosenv.2013.08.023, 2-s2.0-84883717859.
16. Zhao X., Song M., Liu A., Wang Y., Wang T., and Cao J, Data-Driven temporal-spatial model for the prediction of AQI in nanjing, Journal of Artificial Intelligence and Soft Computing Research. (2020) 10, no. 4, 255–270, https://doi.org/10.2478/jaiscr-2020-0017.

17. Chowdhury A.-S., Uddin M. S., Tanjim M. R., Noor F., and Rahman R. M., Application of Data Mining Techniques on Air Pollution of Dhaka City, Proceedings of the 2020 IEEE 10th International Conference on Intelligent Systems (IS), August 2020, Varna, Bulgaria, 562–567, https://doi.org/10.1109/IS48319.2020.920012.
18. Zhou Y., De S., Ewa G., Perera C., and Moessner K., Data-Driven air quality characterization for urban environments: a case study, IEEE Access. (2018) 6, 77996, https://doi.org/10.1109/ACCESS.2018.2884647, 2-s2.0-85057792837.
19. Srivastava C., Singh S., and Singh A. P., Estimation of air pollution in Delhi using machine learning techniques, Proceedings of the 2018 International Conference on Computing, Power and Communication Technologies (GUCON), September 2018, Greater Noida, India, 304–309, https://doi.org/10.1109/GUCON.2018.8675022, 2-s2.0-85064391411.s.
20. Maltare N. N. and Vahora S., "Air Quality Index prediction using machine learning for Ahmedabad city," Digital Chemical Engineering, vol. 7, p. 100093, 2023, https://doi.org/10.1016/j.dche.2023.100093.