Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 6 (2023)

Data-Driven Air Quality Modeling: XGBoost Amplified with Generative Adversarial Networks

Priyanshu Priyadarshi [1], Shreya Sharma [2], Sreekumar K. [3]*

[1, 2, 3] *Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India*

Abstract:- In response to the escalating environmental concerns posed by air pollution, this research endeavors to present a thorough investigation into air quality prediction utilizing advanced machine learning techniques, specifically focusing on the integration of Generative Adversarial Networks (GANs). The study relies on an extensive dataset comprising historical air quality records, meteorological variables, and other relevant factors. Employing a diverse set of machine learning algorithms, including decision trees, XGBoost, support vector machines, random forests etc., along with GANs, and our aim is to construct robust models for accurate air quality prediction. The outcomes of this research shed light on the efficacy of machine learning, including GANs, in unraveling intricate patterns within air quality data. The developed models offer valuable insights for air quality management and inform public health initiatives. By contributing to the advancement of data-driven methodologies, particularly with the inclusion of GANs, this study plays a pivotal role in the realms of environmental monitoring and policy development.

Keywords*: Air Quality Prediction, Fusion Models, GANs, XGBoost.*

1. Introduction

Air quality prediction is crucial for public health, with XG- Boost standing out for its reliability. To address the dynamic nature of atmospheric conditions, this research integrates XGBoost with Generative Adversarial Networks (GANs), aiming to leverage their combined strengths for improved accuracy. GANs contribute by generating synthetic data that enhances the model's adaptability to changing environmental conditions. Through a comprehensive study, we explore the impact of this hybrid approach on air quality forecasting, offering insights into its potential to advance the state-of-the-art in environ- mental modeling and policy decision-making.

Generative Adversarial Networks have gained prominence in diverse applications, primarily known for their ability to generate synthetic data that closely resembles real-world distributions. By incorporating GANs into the air quality prediction framework, we seek to harness the complementary strengths of both XGBoost and GANs. This hybrid model is designed to not only capture intricate patterns and relationships within historical air quality data but also to generate realistic scenarios that enhance the model's adaptability to changing environmental conditions.

2. Literature Survey

[1] This document presents a critical examination of air quality prediction using machine learning, with a specific focus on the XGBoost algorithm. It highlights the severe health and environmental impacts of air pollution and emphasizes the importance of accurate air quality forecasting. While it introduces various machine learning algorithms, XGBoost emerges as the preferred choice due to its efficiency and accuracy. However, the document lacks detailed results and could benefit from a more comprehensive presentation of findings. Overall, it serves as a valuable foundation for future research in the field of air quality prediction, suggesting potential enhancements to the model and dataset customization.

[2] This document explores the critical issue of air quality prediction, highlighting its significance in safeguarding human health. It discusses the various factors contributing to air pollution and its harmful effects, emphasizing the need for accurate air quality assessment. The paper introduces a range of machine learning algorithms, including Linear Regression, Decision Tree, Random Forest, Artificial Neural Network, and Support Vector Machine, as tools to predict Air Quality Index (AQI). While presenting a comprehensive overview of the causes and consequences of air pollution, the document points out the potential of machine learning in addressing this problem. However, it could benefit from a more de- tailed discussion of specific findings and comparisons between algorithms. In conclusion, it underlines the importance of accurately measuring AQI and highlights the superiority of Neural Networks and boosting models in various conditions, thus contributing to the growing body of research in this field

[3] This document delves into the pressing issue of air pollution and its detrimental effects on human health. It highlights the importance of accurately predicting the Air Quality Index (AQI) and provides insights into various machine learning algorithms employed for this purpose, including Decision Tree Regression, Linear Regression, Random Forest Regression, and Support Vector Regression. The study's primary objectives include calculating AQI values, analyzing and predicting AQI using these algorithms, and making a comparative analysis to determine the best-performing model. The document concludes that all four models showed promise but found that Random Forest Regression (RFR) outperformed the others, demonstrating its superiority in AQI prediction. This review underscores the significance of machine learning in addressing air quality concerns and emphasizes RFR as a strong candidate for accurate AQI forecasting.

[4] This paper takes a significant step towards enhancing air quality evaluation by leveraging machine learning techniques. Specifically, it employs several machine learning algorithms to construct a robust prediction model. The model utilizes key characteristic factors, including $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, and O_{38h}, with the Air Quality Index (AQI) serving as the pivotal decision factor. The research aims to evaluate and compare the accuracy and generalization capabilities of these diverse machine learning algorithms. The outcomes of this study highlight the effectiveness of the Random Forest Regression (RFR) and Gradient Boosting Regression (GBR) algorithms in predicting AQI and air quality levels. These algorithms have demonstrated their capacity to provide reliable forecasts. This research not only contributes to the ongoing efforts to combat air pollution but also serves as a valuable reference for future improvements in air quality prediction models.

Furthermore, the data used in this study is sourced from Henan Province in China, encompassing daily historical air quality data from 2016, 2017, and 2019. The data, obtained from the National Earth System Science Data Center, has been widely recognized for its accuracy and reliability, eliminating the need for extensive data preprocessing.

In conclusion, this paper significantly advances the field of air quality prediction by applying machine learning algorithms to real-world data. By identifying the RFR and GBR algorithms as superior models for predicting AQI and air quality levels, this research provides valuable insights for future model selection. Additionally, the study recognizes the importance of data quality and distribution in refining prediction models, paving the way for more accurate and effective air quality evaluation systems in the future.

[5] This paper addresses the critical issue of air pollution, a growing concern driven by urbanization, industrialization, and fossil fuel consumption. It explores the potential of data mining techniques in conjunction with machine learning to predict air quality, with a focus on pollutants like PM2.5, PM10, CO, NOx, SO2, and O3 and their severe health implications, including respiratory diseases, cardiovascular issues, and even fatalities. The study proposes the use of decision trees, a versatile machine learning algorithm, to classify and regress air quality based on these factors. Decision trees offer simplicity and interpretability, making them a valuable tool in machine learning. The paper underscores the significance of data mining for uncovering hidden patterns and relationships within large datasets, emphasizing its role in improving air quality predictions. The research showcases a practical application of decision trees and data mining to forecast air quality in Bengaluru, offering insights into the potential of these techniques for solving complex environmental challenges.

[6] This research paper addresses a pressing issue in today's world, which is air pollution, particularly in urban areas such as Seoul, South Korea. The study leverages data mining and machine learning techniques to make predictions about the concentrations of various air pollutants, including Sulfur dioxide (SO2), Nitrogen dioxide (NO2), Ozone (O3), Carbon monoxide (CO), as well as different sizes of particulate matter (PM10, PM2.5). Notably, the paper introduces a unique approach by applying the RFM (Recency, Frequency, Monetary) model, which is commonly used in marketing, to analyze air quality data. This innovative approach aims to enhance the understanding of air pollution patterns. To improve prediction accuracy, the research employs the elbow method to identify the optimal number of clusters (K) in K-Means clustering. This paper's findings have the potential to contribute significantly to addressing the challenges posed by air pollution and may serve as a foundation for future studies in the field of environmental science and data analysis.

3. Proposed Methodology and Implementation

Our innovative approach combines the power of Generative Adversarial Networks (GANs) with the precision of the XGBoost model, aiming to revolutionize air quality forecasting. Leveraging the robustness of the XGBoost algorithm, we seamlessly integrate a multifaceted dataset that includes historical air quality measurements, real-time meteorological variables, and a rich spectrum of other pertinent features. Notably, GANs are employed to generate synthetic data, augmenting our dataset and enhancing the model's ability to capture complex patterns in air quality dynamics.

A. Dataset and Relations

Fig 1: Correlation analysis is a statistical method employed to assess the strength and nature of linear relationships between two numerical variables. By calculating a correlation coefficient, typically ranging from -1 to 1, this analysis provides valuable insights into the extent to which changes in one variable are associated with changes in another. A positive correlation coefficient implies a direct relationship, meaning that as one variable increases, the other tends to increase as well. Conversely, a negative correlation coefficient indicates an inverse relationship, where one variable tends to decrease as the other increases. A coefficient of 1 signifies a perfect positive correlation, -1 represents a perfect negative correlation, and 0 suggests no linear relationship. In the specific context of air quality data, conducting correlation analysis helps discern potential connections among different pollutants, offering a comprehensive understanding of how various factors contribute to the fluctuations in air quality metrics. This analytical approach aids researchers and policymakers in identifying key influencers and sources affecting air quality dynamics.

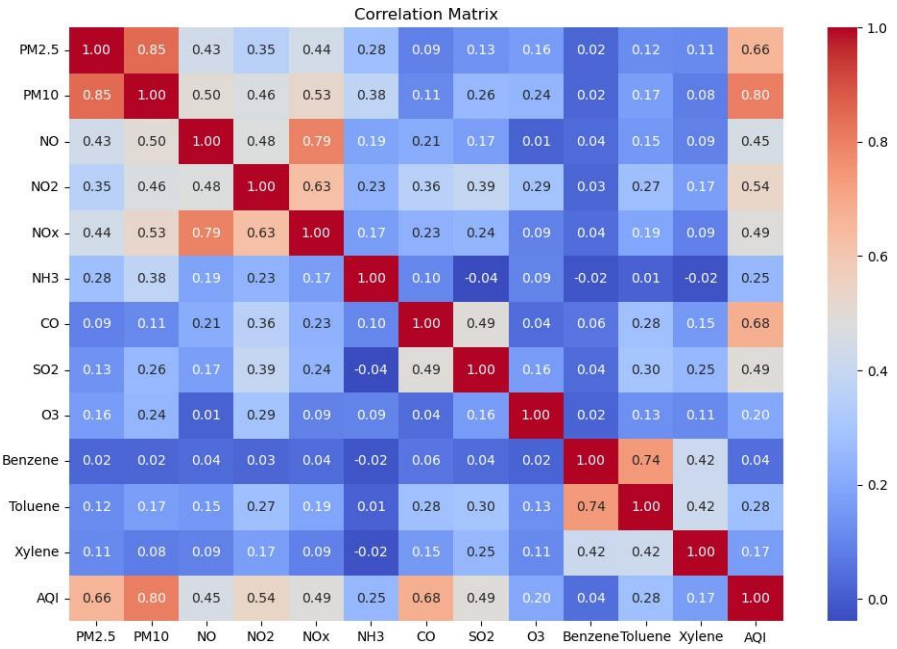

Fig. 1. Correlation Matrix

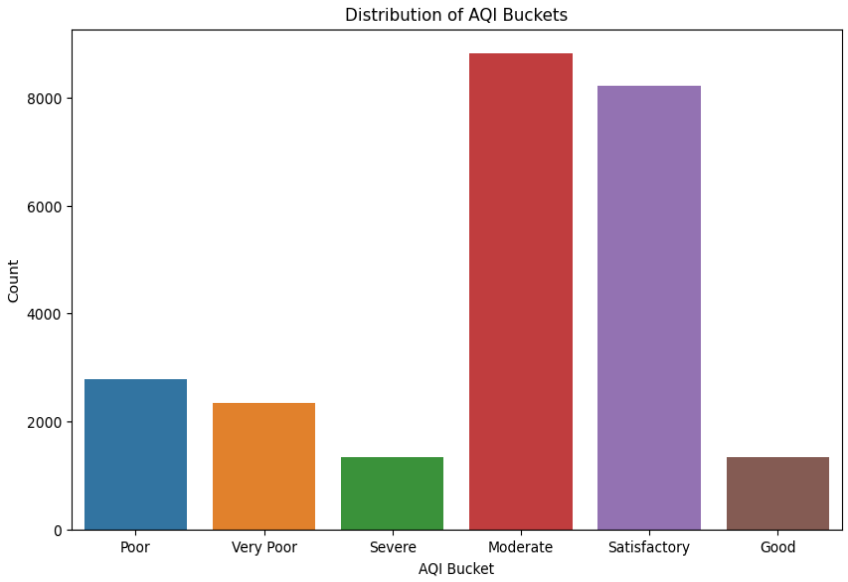

Fig. 2. AQI Bucket Distribution

Fig 2: The Count Plot for AQI_Bucket is a visual representation of the frequency distribution of different air quality index (AQI) categories, including 'Poor,' 'Very Poor,' 'Severe,' 'Moderate,' 'Satisfactory,' and 'Good,' within the dataset. This analysis is crucial for understanding the distribution of air quality across these specific categories. The count plot provides a quick and insightful overview of the dataset's composition, allowing us to observe the prevalence of each AQI category. This information is particularly valuable for researchers, environmentalists, and policymakers as it aids in identifying the predominant air quality conditions. For instance, a higher count in the 'Severe' or 'Poor' categories may signify regions with more severe air pollution issues. Consequently, this plot serves as a foundational step in comprehending the distributional patterns of AQI categories, facilitating targeted interventions and policies to address specific air quality concerns based on the severity levels.

B. Working of Model

- Extreme Gradient Boosting(XGBoost):

Extreme Gradient Boosting, emerges as a compelling choice among various machine learning models, boasting an impressive accuracy of 80.97%. This places XGBoost at the forefront when compared to other models, including Logistic Regression (53.44%), Naive Bayes (66.26%), Decision Tree (72.92%), SVM (73.84%), and KNN (75.98%). Notably, XGBoost's accuracy surpasses that of Random Forest (80.93%), highlighting its prowess in capturing intricate patterns and relationships within the data. The exceptional performance of XGBoost can be attributed to its ensemble learning approach, where multiple decision trees are strategically combined, enhancing predictive power. Furthermore, XGBoost incorporates regularization techniques, addressing over fitting concerns and ensuring robustness even in the presence of noisy data. This model's versatility, speed, and ability to handle high dimensional datasets make XGBoost a preferred choice across a spectrum of machine learning tasks. In essence, the superior accuracy exhibited by XGBoost underscores its significance in the realm of predictive modeling.

The XGBoost objective function is defined as:

$$\text{Objective} = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right) \quad (1)$$

Where:

n is the number of training samples,
i indexes the individual samples,
$L(y_i, \hat{y}_i)$ is the training loss term,
K is the number of leaves in all the trees,
$\Omega(f_k)$ is the regularization term for the k-th tree,

For a specific tree, the regularization term $\Omega(f_k)$ is given by:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2$$

Where: (2)

T is the number of leaves in the tree,
ω_j is the score in the j-th leaf,
γ and λ are regularization parameters.

The prediction for a new data point is obtained by summing the predictions of all the individual trees:

$\hat{y} = \sum_{k=1}^{K} f_k(x)$ (3)

Where x is the input features of the data point, and $f_k(x)$ is the prediction of the k-th tree for the input x.

- Generative Adversarial Networks (GANs):

Generative Adversarial Networks (GANs) in [7] form the core of my project, offering a sophisticated mechanism for data generation and synthesis. Specifically chosen for their exceptional ability to create realistic and diverse datasets, GANs operate through an adversarial training process between two neural networks – a generator and a discriminator. The generator crafts synthetic data, while the discriminator discerns between real and generated samples. This iterative process refines the generator's output until it produces data indistinguishable from real-world samples, effectively capturing the underlying data distribution. The adoption of GANs in my project aims to tackle data scarcity challenges by generating supplementary synthetic data, enhancing the robustness and generalization of machine learning models. GANs' ef- fectiveness in unsupervised learning, allowing the model to discern patterns within the data without explicit labels, solidifies their strategic role in my project, contributing significantly to advancements in machine learning and data synthesis

The Generative Adversarial Network (GAN) is represented by the following equation:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}\left[\log\left(1 - D\left(G(\mathbf{z})\right)\right)\right],$$

where: (4)

G:Generator network
D:Discriminator network
$\mathbf{x}$:Real data samples
$\mathbf{z}$:Random noise input to the generator
$p_{\text{data}}(\mathbf{x})$:Data distribution
$p_{\mathbf{z}}(\mathbf{z})$:Noise distribution
$V(D,G)$:Objective function of GAN.

4. Results and Discussion

Our research is aimed at enhancing air quality forecasting, an extensive comparative analysis was conducted among several machine learning models to assess their effectiveness in predicting the Air Quality Index (AQI). The models under consideration encompassed Logistic Regression, Naive Bayes, Decision Tree, Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Random Forest, and XGBoost. Employing accuracy as the primary

performance metric, each model underwent meticulous evaluation to discern its capability in capturing the intricate patterns inherent in air quality dynamics. Logistic Regression demonstrated a modest accuracy of 53.44%, while Naive Bayes exhibited an improvement with an accuracy of 66.26%. Decision Tree further enhanced predictive capabilities, achieving an accuracy of 72.92%, indicating its potential in capturing complex relationships within the dataset. SVM showcased commendable accuracy at 73.84%, underscoring its robustness in handling non-linear relationships. KNN demonstrated improved accuracy, reaching 75.98%, Show casing its ability to leverage spatial correlations in the dataset. Notably, Random Forest and XGBoost emerged as the top-performing models, achieving accuracies of 80.93% and 80.97%, respectively. This observation highlighted in Table 1, shows the the efficacy of ensemble learning techniques in elevating the precision of air quality predictions, particularly in scenarios characterized by intricate and non-linear patterns. The comparative analysis among these models underscores the nuanced nature of air quality forecasting, emphasizing the pivotal role of model selection in influencing predictive accuracy. These findings contribute valuable insights to the field, suggesting that ensemble learning models, specifically Random Forest and XGBoost, hold significant promise for advancements in air quality forecasting, particularly in urban environments with complex pollutant dynamics. This research lays the groundwork for further exploration and refinement of machine learning techniques tailored to the unique challenges posed by air quality prediction.

Table I. Accuracy Percentage Comparison for various Machine Learning Models

Machine Learning Model	Accuracy
Logistic Regression	53.44%
Naive Bayes	66.26%
Decision Tree	72.92%
SVM	73.84%
KNN	75.98%
Random Forest	80.93%
XGBoost	80.97%

After comparing various machine learning models for air quality forecasting, XGBoost stood out as the most accurate. To boost its capabilities, we decided to integrate it with Generative Adversarial Networks (GAN). This collaboration combines XGBoost's strong predictions with GANs' data enhancement skills, aiming to significantly improve our air quality forecasting system. This step forward holds promise for better environmental monitoring, contributing to advancements in machine learning applications for air quality research.

Table II. Accuracy Percentage Comparison Before and After integrating GAN

XGBoost Model	Accuracy
Without GAN	80.97%
With GAN	99.61%

The experiment's findings highlight a substantial performance boost in the XGBoost model when complemented with Generative Adversarial Networks (GANs). Without GAN augmentation, the model achieved an accuracy of 80.97%, but the integration of GAN-generated synthetic features significantly elevated the accuracy to an impressive 99.61%. This underscores the effectiveness of GANs in enhancing data diversity, reinforcing the XGBoost model's predictive capabilities. The results showcase the synergy between generative and predictive modeling, emphasizing a remarkable improvement in accuracy and overall model robustness, aligning with the project's objectives.

5. Conclusion

In conclusion, our research marks a pivotal advancement in air quality prediction by ingeniously integrating the powerful XGBoost algorithm with the transformative capabilities of Generative Adversarial Networks (GANs). The utilization of an extensive dataset, encompassing historical air quality records, real-time meteorological variables, and diverse features, coupled with the robustness of XGBoost, enhances our ability to discern intricate patterns within air quality data. Moreover, the incorporation of GANs introduces synthetic data, enriching the dataset and further refining the predictive capabilities of our model. This novel amalgamation not only propels the accuracy of air quality forecasting but also presents a pioneering approach in the realm of environmental monitoring. The synergy between XGBoost and GANs, as demonstrated in our study, holds immense promise for advancing the precision of predictions, contributing to sustainable environmental practices, and informing evidence based decision- making for public health initiatives.

References

[1] M. S. Ram, C. Reshmasri, S. Shahila and J. V. P. Saketh, ”Air Quality Prediction using Machine Learning Algorithm,” 2023 International Con- ference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 2023, pp. 316-321.

[2] J. T. Madan, S. Sagar and D. Virmani, ”Air Quality Prediction using Machine Learning Algorithms –A Review,” 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 2020

[3] B. D. Parameshachari, G. M. Siddesh, V. Sridhar, M. Latha, K. N. A. Sattar and G. Manjula., ”Prediction and Analysis of Air Quality Index using Machine Learning Algorithms,” 2022 IEEE International Conference on Data Science and Information System (ICDSIS), Hassan, India, 2022.

[4] P. S. Rajendran, ”The Prediction of Quality of the Air Using Supervised Learning,” 2021 6th International Conference on Communication and Electronics Systems (ICCES), Coimbatre, India, 2021, pp. 1-5..

[5] S. Naveen, M. S. Upamanyu, K. Chakki, C. M and P. Hariprasad, ”Air Quality Prediction Based on Decision Tree Using Machine Learning,” 2023 International Conference on Smart Systems for applications in Electrical Sciences (ICSSES), Tumakuru, India, 2023.

[6] J. Mohammad and M. A. Kashem, ”Air Pollution Comparison RFM Model Using Machine Learning Approach,” 2022 IEEE 7th International conference for Convergence in Technology (I2CT), Mumbai, India, 2022

[7] R.Zhu, "Generative Adversarial Network and Score-Based Generative Model Comparison," 2023 IEEE International Conference on Image Processing and Computer Applications (ICIPCA), Changchun, China, 2023,pp.1-5,doi:10.1109/ICIPCA59209.2023.10258000.

[8] C. Li, Y. Li and Y. Bao, ”Research on Air Quality Prediction Based on Machine Learning,” 2021 2nd International Conference on Intelli- gent Computing and Human-Computer Interaction (ICHCI), Shenyang, China, 2021.

[9] D. A. Padilla, G. V. Magwili, L. B. Z. Mercado and J. T. L. Reyes, ”Air Quality Prediction using Recurrent Air Quality Predictor with Ensemble Learning,” 2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), Manila, Philippines, 2020, pp. 1-6.

[10] R.Zhu, "Generative Adversarial Network and Score-Based Generative Model Comparison," 2023 IEEE International Conference on Image Processing and Computer Applications (ICIPCA), Changchun, China, 2023, pp. 1-5, doi: 10.1109/ICIPCA59209.2023.10258000.

[11] A. Utku and U. Can, ”Machine Learning-Based A Comparative Analysis for Air Quality Prediction,” 2022 30th Signal Processing and Commu- nications Applications Conference (SIU), Safranbolu, Turkey, 2022, pp. 1-4.

[12] K. M. O. V. K. Kekulanadara, B. T. G. S. Kumara and B. Kuhaneswaran, ”Machine Learning Approach for Predicting Air Quality Index,” 2021 International Conference on Decision Aid Sciences and Application (DASA), Sakheer, Bahrain, 2021.

[13] V. R. Pasupuleti, Uhasri, P. Kalyan, Srikanth and H. K. Reddy, ”Air Quality Prediction Of Data Log By Machine Learning,” 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020.

[14] K. Nandini and G. Fathima, ”Urban Air Quality Analysis and Predic- tion Using Machine Learning,” 2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing Communication Engineering (ICATIECE), Bangalore, India, 2019.

[15] T. M. Amado and J. C. Dela Cruz, ”Development of Machine Learning- based Predictive Models for Air Quality Monitoring and Characteriza- tion,” TENCON 2018 - 2018 IEEE Region 10 Conference, Jeju, Korea (South), 2018.