**Title page:**

Air Pollution Projection  using XGB algorithm compared with Random Forest algorithm using Machine Learning**.**

      U.Sahithi¹ , Dr.Ravi Shankar²

U.Sahithi1,

Research Scholar,

Department of Computer science Engineering,

Saveetha School of Engineering,

Saveetha Institute of Medical and Technical Sciences,

Saveetha University, Chennai, Tamilnadu, India Pincode: 602105.

192111224.sse@saveetha.com

Dr.Ravi Shankar²,

Project Guide, Corresponding Author,

Department of Artificial Intelligence and Machine Learning,

Saveetha School of Engineering,

Saveetha Institute of Medical and Technical Sciences,

Saveetha University, Chennai, Tamilnadu, India, Pincode: 602105.

ramamoorthym.sse@saveetha.com

**Keywords:** Machine learning, Air Pollution, XGB algorithm, Random Forest algorithm, projection.

**ABSTRACT**

**Aim:** The objective of this article is to conduct a comparative analysis of a Air pollution projection employing the XGBoost (XGB) algorithm against the Random Forest algorithm within the domain of Machine Learning.**Materials and Methods:** The study divided the XGBoost (XGB) algorithm and the Random Forest algorithm into distinct groups, with each group comprising a total of 20 samples. Statistical comparisons were conducted using G-power, where the alpha value was set at 0.8 to assess accuracy.**Results:** The XGBoost (XGB) algorithm demonstrates a higher accuracy rate at 90.8%, surpassing the lower accuracy rate of 74% observed in the Random Forest algorithm. The statistical significance of the study, indicated by a p-value of 0.004 (p<0.05 in Independent Sample T-Test), confirms the significance of the obtained results.**Conclusion**: The XGBoost (XGB) algorithm excels in accuracy prediction and further enhances accuracy more than the Random Forest algorithm.

**Keywords:** Machine learning, XGB algorithm, Random Forest algorithm, Air Pollution .

**INTRODUCTION**

[(OECD 2016)](https://paperpile.com/c/LiyY8G/BPg1)The project "Air Pollution Projection Using Machine Learning Techniques Using XGB Compared with Random Forest" aims to utilise machine learning models, specifically XGBoost and Random Forest, to predict air pollution levels. [(Mantas, Gallos, and Zoulias 2023)](https://paperpile.com/c/LiyY8G/OTtU)The research is significant due to the growing concern over air pollution's impact on public health and the environment. [(Umamaheswari, Vinoth Kumar, and Somasundaram 2023)](https://paperpile.com/c/LiyY8G/H06F)The accurate prediction of air pollution levels can aid policymakers and city planners in developing effective strategies to mitigate air pollution. The application of this work is crucial for understanding and forecasting air quality, which can help in the development of policies and measures to improve air quality.The research focuses on the use of machine learning models, particularly XGBoost and Random Forest, to predict air pollution levels. [(Alloghani 2023)](https://paperpile.com/c/LiyY8G/33FV)The importance of this work lies in its potential to provide accurate and reliable predictions of air pollution, which can be used by policymakers and city planners to develop effective strategies for mitigating air pollution. The application of this research is significant for understanding and forecasting air quality, which can in turn aid in the development of policies and measures to improve air quality[(Gupta et al. 2023)](https://paperpile.com/c/LiyY8G/0UZ1).

[(Mata-Rivera, Zagal-Flores, and Barria-Huidobro 2023)](https://paperpile.com/c/LiyY8G/exDm)In conducting this research, an extensive database exploration was undertaken, utilising resources such as Google Scholar and Science Direct. [(Dang et al. 2021)](https://paperpile.com/c/LiyY8G/fqcQ)A comprehensive review of existing literature was conducted to identify the most relevant papers in the field of air pollution forecasting using machine learning techniques, with a focus on studies employing XGBoost and Random Forest. [(Magdi et al. 2023)](https://paperpile.com/c/LiyY8G/sfIH)Noteworthy contributions from these papers include insights into feature selection, model optimization, and the integration of diverse environmental data sources. [(Yusoff 2023)](https://paperpile.com/c/LiyY8G/zFnz)One particularly impactful study, acknowledged for its methodological rigor and significant advancements in predicting air pollution levels, underscored the importance of incorporating real-time meteorological data and sensor network information. This exemplary research not only demonstrated a notable improvement in prediction accuracy but also provided a practical framework for the integration of machine learning algorithms into existing air quality monitoring systems. The findings from this study informed the methodology adopted in our research, emphasizing the critical role of data quality and feature engineering in enhancing the performance of predictive models for air pollution[(Dinh and Hoang 2021)](https://paperpile.com/c/LiyY8G/S5IM).

[(Liu et al. 2021)](https://paperpile.com/c/LiyY8G/yelD)This research project aims to address a significant gap in existing literature by conducting a thorough comparative analysis of XGBoost and Random Forest algorithms in the context of air pollution forecasting. [(Tripathy et al. n.d.)](https://paperpile.com/c/LiyY8G/qpqC)Drawing from an interdisciplinary team with expertise in environmental science and machine learning, the study explores relevant databases such as Google Scholar and Science Direct to identify the most pertinent papers. The expertise brought to this research stems from a multidisciplinary team comprising experts in environmental science, machine learning, and data analysis.[(Balamurugan et al. 2023)](https://paperpile.com/c/LiyY8G/ilQ0)By emphasising the identified research gap, the team aims to contribute nuanced insights to the field, offering recommendations for improving the accuracy of air quality prediction models. [(Sun et al. 2023)](https://paperpile.com/c/LiyY8G/Zldn)Ultimately, this research endeavors to empower decision-makers with a robust understanding of the comparative strengths and limitations of these algorithms for more effective environmental management[(Moursi et al. 2021)](https://paperpile.com/c/LiyY8G/wRcx).

**MATERIALS AND METHODS**

[(Zheng, Cheng, and Li n.d.)](https://paperpile.com/c/LiyY8G/twTZ)The study was carried out at the Machine Learning laboratory situated at Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences in Chennai. The sample size, determined using GPower software, involves the comparison of two groups in a controlled study. [(Barthwal, Acharya, and Lohani 2021)](https://paperpile.com/c/LiyY8G/f7jp)Each group consists of 20 sets of samples, resulting in a total of 20 samples. GPower 3.1 was employed for calculating pre-test power, with parameters set at α=0.05 and power=0.80 for comparing two independent means.

The study implemented two machine learning algorithms, XGBoost (XGB) and Random Forest(RF), for classification using Technical Analysis software. Ethical approval was not required as the research did not involve human or animal samples. The experimental setup included hardware components such as an HP i5 processor, 8GB RAM, 1TB HDD, and software components including Windows 11 OS, Google Colab, Chrome Browser, MS Excel, and the SPSS tool. The testing process consisted of downloading a dataset from Kaggle, followed by feature extraction and cleaning. The model was trained with 70% of the dataset, while 30% was used for testing and validation. X1 and Y1 values were computed from the algorithms, and accuracy was assessed using confusion matrix technology.

The dataset, obtained from Kaggle, comprised 20,000 records with 19 attributes, encompassing categories such as CGPA, school type, number of mini projects, aptitude skill, core subject skills, and problem-solving skills.

**XGB Algorithm**

XGBoost (eXtreme Gradient Boosting) stands out as a formidable machine learning algorithm renowned for its effectiveness and precision. Classified within the gradient boosting family, it demonstrates exceptional performance in managing extensive datasets and intricate associations. XGB constructs a sequence of decision trees and merges their results to generate resilient predictions, contributing to its widespread application across various domains.

**Pseudocode for XGB**

**1. Load and Process Data:**

2. Import necessary libraries like numpy, pandas, and xgboost.

Ingest the dataset into a Pandas DataFrame.

Preprocess the data, addressing missing values, and encoding categorical features.

**2.Split Dataset:**

Partition the dataset into features (X) and labels (y).

**3.Train-Test Division:**

Segment the data into training and testing sets.

**4.Initialize XGBoost Model:**

Set hyperparameters, including:

Learning rate (eta): Typically a small value, e.g., 0.1.

Number of boosting rounds (num\_boost\_round): Determined through cross-validation.

Maximum depth of weak learners (max\_depth): Controls tree complexity.

Specify other hyperparameters like subsample, colsample\_bytree, etc.

**5.Train XGBoost Model:**

utilize the training set to fit the XGBoost model.

Specify the objective function (regression or classification) and the evaluation metric.

**6.Generate Predictions:**

Apply the trained model to predict outcomes on the test set.

**7.Assess Model Performance:**

Evaluate the model's effectiveness using relevant metrics, such as accuracy or mean squared error**.**

**Random Forest Algorithm**

(Alhothali et al. 2022) Random Forest stands out as an ensemble learning algorithm that constructs numerous decision trees in the training process, amalgamating their results to boost accuracy. By incorporating randomness in both data sampling and feature selection, it addresses overfitting concerns, ensuring resilient predictions. Renowned for its adaptability, Random Forest finds extensive application in machine learning for both classification and regression assignments.

**Pseudocode for Random Forest**

1.Load and Process Data:

-Import essential libraries (such as numpy, pandas, scikit-learn).

-Load the dataset into a Pandas DataFrame.

-Preprocess the data, addressing tasks like handling missing values and encoding categorical features.

2.Divide Data:

-Separate the dataset into features (X) and labels (y).

3.Split into Training and Testing Sets:

-Partition the data into training and testing sets.

4.Initialize Random Forest Model:

-Select the appropriate Random Forest model based on the task (classification or regression).

-Set hyperparameters, including:

-For classification: Number of trees, criterion (e.g., Gini, Entropy), maximum depth, etc.

-For regression: Number of trees, criterion, maximum depth, etc.

5.Train Random Forest Model

-Utilize the training set to train the Random Forest model.

6.Generate Predictions:

-Employ the trained Random Forest model to make predictions on the test set.

7.Evaluate Model Performance:

-Evaluate the model's performance using relevant metrics:

-For classification: Metrics like accuracy, precision, recall, F1-score, etc.

-For regression: Metrics such as mean squared error, R-squared, etc.

**Statistical Analysis**

[(Pan, Harrou, and Sun 2023)](https://paperpile.com/c/LiyY8G/d8r7)The analysis is carried out using IBM SPSS version 26 (64-bit), a software designed for analysis through dataset uploads. It furnishes outputs encompassing independent variables count (N), means, standard deviation, and standard error mean with precision, specifically for the XGBoost algorithm and the Random Forest algorithm. IBM SPSS is applied to juxtapose the Student Career Guidance System's performance when employing the XGBoost algorithm versus the Random Forest algorithm within the realm of machine learning.

**RESULTS**

The table provided outlines the outcomes of simulations conducted on both the proposed XGB algorithm and the Random Forest algorithm. These simulations were carried out at various intervals using Google Colab, with a sample size of 10. Notably, the XGB algorithm exhibited a mean accuracy of 90.8%, while the Random Forest algorithm showed a considerably lower mean accuracy of 34%. Calculations for Mean, Standard Deviation, and Standard Error Mean involved performing an independent variable T test among the study groups. The comparison highlighted a significant distinction between the XGB algorithm and the Random Forest algorithm, with a p-value of 0.001.

Table 2 illustrates the superior mean of the XGB algorithm compared to the Random Forest algorithm, with a mean difference of 59.066 and a standard error difference of 9.311. Regarding mean accuracy, the XGB algorithm outperforms the Random Forest algorithm, achieving 93.06%, whereas the Random Forest algorithm lags behind at 34%. Figure 1 presents a comparative chart between the XGB and Random Forest algorithms, clearly indicating the superiority of XGB. In conclusion, the evidence strongly suggests that XGB performs better than Random Forest, as depicted by the accompanying figures appended at the end of the paper.

**DISCUSSION**

Our investigation into air pollution projection underscores a substantial performance gap, as evidenced by XGBoost (XGB) achieving a notably higher mean accuracy of 93.06% compared to the comparatively lower accuracy of 34% observed with Random Forest. This aligns with a prevailing consensus in contemporary research, exemplified by studies conducted and, consistently showcasing XGB's superiority in diverse machine learning applications. While our results bolster this widely accepted viewpoint, it is crucial to acknowledge the nuanced nature of algorithmic performance illuminated by studies[(Yılmaz 2021)](https://paperpile.com/c/LiyY8G/LagG), emphasizing the influence of specific contextual factors. Despite potential variations in findings, our overall assessment solidifies the consensus that XGB excels over Random Forest in the specific domain of air pollution projection, affirming the broader trend observed in recent literature, which underscores XGB's superior performance across a spectrum of machine learning applications.

[(Ding and Qie 2022)](https://paperpile.com/c/LiyY8G/kN7g)Factors influencing our study include the intricacies of environmental data, which can vary in quality and granularity, impacting the precision of air pollution projections. [(Masood and Ahmad 2022)](https://paperpile.com/c/LiyY8G/LKKK)Additionally, the choice of features and model hyperparameters contributes to the variability in results, reflecting the sensitivity of machine learning algorithms to such selections. [(Li et al. 2021)](https://paperpile.com/c/LiyY8G/o51w)Limitations of our study encompass the reliance on a specific dataset and the potential bias associated with it. [(Kujawska et al. 2022)](https://paperpile.com/c/LiyY8G/ZVWQ)Furthermore, while our focus is on XGBoost and Random Forest, the rapidly evolving landscape of machine learning introduces new algorithms that could warrant exploration. In terms of future scope, expanding the dataset to include diverse environmental conditions, incorporating real-time data streams, and exploring ensemble models involving multiple algorithms could enhance the robustness and applicability of our air pollution projection framework.Addressing these factors and limitations would contribute to a more comprehensive understanding and improved accuracy in the prediction of air pollution levels.

**CONCLUSION**

Examining the accuracy rates in the Air pollution projection, this study contrasts the performance of XGBoost (XGB) with Random Forest. The findings demonstrate that XGB Prediction achieves a notably higher accuracy of 90.8%, outperforming Random Forest, which has an accuracy rate of 74%.

**DECLARATION**

**Conflict of Interests**

No conflict of interest in this manuscript.

**Authors Contributions**

Author US was involved in data collection, data analysis and manuscript writing. Author RS was involved in the conceptualization, data validation and critical review of manuscript.

**Acknowledgement**

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

**Funding**

Personally, I would like to thank the following organizations for providing financial support that enabled us to complete the study.

1. Saveetha University

2. Saveetha Institute of Medical And Technical Sciences

3. Saveetha School of Engineering

**REFERENCES**

[Alloghani, Mohamed Ahmed. 2023.](http://paperpile.com/b/LiyY8G/33FV) *[Artificial Intelligence and Sustainability](http://paperpile.com/b/LiyY8G/33FV)*[. Springer Nature.](http://paperpile.com/b/LiyY8G/33FV)

[Balamurugan, Vigneshkumar, Jia Chen, Adrian Wenzel, and Frank N. Keutsch. 2023. “Spatiotemporal Modeling of Air Pollutant Concentrations in Germany Using Machine Learning.”](http://paperpile.com/b/LiyY8G/ilQ0) *[Atmospheric Chemistry and Physics](http://paperpile.com/b/LiyY8G/ilQ0)* [23 (17): 10267–85.](http://paperpile.com/b/LiyY8G/ilQ0)

[Barthwal, Anurag, Debopam Acharya, and Divya Lohani. 2021. “Prediction and Analysis of Particulate Matter (PM2.5 and PM10) Concentrations Using Machine Learning Techniques.”](http://paperpile.com/b/LiyY8G/f7jp) *[Journal of Ambient Intelligence and Humanized Computing](http://paperpile.com/b/LiyY8G/f7jp)* [14 (3): 1323–38.](http://paperpile.com/b/LiyY8G/f7jp)

[Dang, Tran Khanh, Josef Küng, Tai M. Chung, and Makoto Takizawa. 2021.](http://paperpile.com/b/LiyY8G/fqcQ) *[Future Data and Security Engineering. Big Data, Security and Privacy, Smart City and Industry 4.0 Applications: 8th International Conference, FDSE 2021, Virtual Event, November 24–26, 2021, Proceedings](http://paperpile.com/b/LiyY8G/fqcQ)*[. Springer Nature.](http://paperpile.com/b/LiyY8G/fqcQ)

[Ding, Weifu, and Xueping Qie. 2022. “Prediction of Air Pollutant Concentrations via RANDOM Forest Regressor Coupled with Uncertainty Analysis—A Case Study in Ningxia.”](http://paperpile.com/b/LiyY8G/kN7g) *[Atmosphere](http://paperpile.com/b/LiyY8G/kN7g)* [13 (6): 960.](http://paperpile.com/b/LiyY8G/kN7g)

[Dinh, Thuan Nguyen, and Nam Phan Hoang. 2021. “Air Pollution Forecasting Using Regression Models and LSTM Deep Learning Models for Vietnam.”](http://paperpile.com/b/LiyY8G/S5IM) *[Future Data and Security Engineering. Big Data, Security and Privacy, Smart City and Industry 4.0 Applications](http://paperpile.com/b/LiyY8G/S5IM)*[, 264–75.](http://paperpile.com/b/LiyY8G/S5IM)

[Gupta, Rajeev Kumar, Arti Jain, John Wang, Santosh Kumar Bharti, and Samir Patel. 2023.](http://paperpile.com/b/LiyY8G/0UZ1) *[Artificial Intelligence Tools and Technologies for Smart Farming and Agriculture Practices](http://paperpile.com/b/LiyY8G/0UZ1)*[. IGI Global.](http://paperpile.com/b/LiyY8G/0UZ1)

[Kujawska, Justyna, Monika Kulisz, Piotr Oleszczuk, and Wojciech Cel. 2022. “Machine Learning Methods to Forecast the Concentration of PM10 in Lublin, Poland.”](http://paperpile.com/b/LiyY8G/ZVWQ) *[Energies](http://paperpile.com/b/LiyY8G/ZVWQ)* [15 (17): 6428.](http://paperpile.com/b/LiyY8G/ZVWQ)

[Li, Rui, Lulu Cui, Yilong Zhao, Wenhui Zhou, and Hongbo Fu. 2021. “Long-Term Trends of Ambient Nitrate (NO](http://paperpile.com/b/LiyY8G/o51w)[3](http://paperpile.com/b/LiyY8G/o51w)[−](http://paperpile.com/b/LiyY8G/o51w)[) Concentrations across China Based on Ensemble Machine-Learning Models.”](http://paperpile.com/b/LiyY8G/o51w) *[Earth System Science Data](http://paperpile.com/b/LiyY8G/o51w)* [13 (5): 2147–63.](http://paperpile.com/b/LiyY8G/o51w)

[Liu, Jianzhao, Yunjiang Zuo, Nannan Wang, Fenghui Yuan, Xinhao Zhu, Lihua Zhang, Jingwei Zhang, et al. 2021. “Comparative Analysis of Two Machine Learning Algorithms in Predicting Site-Level Net Ecosystem Exchange in Major Biomes.”](http://paperpile.com/b/LiyY8G/yelD) *[Remote Sensing](http://paperpile.com/b/LiyY8G/yelD)* [13 (12): 2242.](http://paperpile.com/b/LiyY8G/yelD)

[Magdi, Dalia, Ahmed Abou El-Fetouh, Mohamed Mamdouh, and Amit Joshi. 2023.](http://paperpile.com/b/LiyY8G/sfIH) *[Green Sustainability: Towards Innovative Digital Transformation: Proceedings of ITAF 2023](http://paperpile.com/b/LiyY8G/sfIH)*[. Springer Nature.](http://paperpile.com/b/LiyY8G/sfIH)

[Mantas, J., P. Gallos, and E. Zoulias. 2023.](http://paperpile.com/b/LiyY8G/OTtU) *[Healthcare Transformation with Informatics and Artificial Intelligence](http://paperpile.com/b/LiyY8G/OTtU)*[. IOS Press.](http://paperpile.com/b/LiyY8G/OTtU)

[Masood, Adil, and Kafeel Ahmad. 2022. “Data-Driven Predictive Modeling of PM2.5 Concentrations Using Machine Learning and Deep Learning Techniques: A Case Study of Delhi, India.”](http://paperpile.com/b/LiyY8G/LKKK) *[Environmental Monitoring and Assessment](http://paperpile.com/b/LiyY8G/LKKK)* [195 (1): 1–21.](http://paperpile.com/b/LiyY8G/LKKK)

[Mata-Rivera, Miguel Félix, Roberto Zagal-Flores, and Cristian Barria-Huidobro. 2023.](http://paperpile.com/b/LiyY8G/exDm) *[Telematics and Computing: 12th International Congress, WITCOM 2023, Puerto Vallarta, Mexico, November 13–17, 2023, Proceedings](http://paperpile.com/b/LiyY8G/exDm)*[. Springer Nature.](http://paperpile.com/b/LiyY8G/exDm)

[Moursi, Ahmed Samy, Nawal El-Fishawy, Soufiene Djahel, and Marwa Ahmed Shouman. 2021. “An IoT Enabled System for Enhanced Air Quality Monitoring and Prediction on the Edge.”](http://paperpile.com/b/LiyY8G/wRcx) *[Complex & Intelligent Systems](http://paperpile.com/b/LiyY8G/wRcx)* [7 (6): 2923–47.](http://paperpile.com/b/LiyY8G/wRcx)

[OECD. 2016.](http://paperpile.com/b/LiyY8G/BPg1) *[The Economic Consequences of Outdoor Air Pollution](http://paperpile.com/b/LiyY8G/BPg1)*[. OECD Publishing.](http://paperpile.com/b/LiyY8G/BPg1)

[Pan, Qilong, Fouzi Harrou, and Ying Sun. 2023. “A Comparison of Machine Learning Methods for Ozone Pollution Prediction.”](http://paperpile.com/b/LiyY8G/d8r7) *[Journal of Big Data](http://paperpile.com/b/LiyY8G/d8r7)* [10 (1): 1–31.](http://paperpile.com/b/LiyY8G/d8r7)

[Sun, Hongying, Siyi Chen, Xinyi Li, Liping Cheng, Yipei Luo, and Lingli Xie. 2023. “Prediction and Early Warning Model of Mixed Exposure to Air Pollution and Meteorological Factors on Death of Respiratory Diseases Based on Machine Learning.”](http://paperpile.com/b/LiyY8G/Zldn) *[Environmental Science and Pollution Research](http://paperpile.com/b/LiyY8G/Zldn)* [30 (18): 53754–66.](http://paperpile.com/b/LiyY8G/Zldn)

[Tripathy, Adyant, Divyansh Vaidya, Arpita Mishra, Shardul Bilolikar, and Vijayetha Thoday. n.d. “Analysing and Predicting Air Quality In Delhi: Comparison of Industrial and Residential Area.” Accessed February 22, 2024.](http://paperpile.com/b/LiyY8G/qpqC) <https://ieeexplore.ieee.org/abstract/document/9645787/>[.](http://paperpile.com/b/LiyY8G/qpqC)

[Umamaheswari, K., B. Vinoth Kumar, and S. K. Somasundaram. 2023.](http://paperpile.com/b/LiyY8G/H06F) *[Artificial Intelligence for Sustainable Applications](http://paperpile.com/b/LiyY8G/H06F)*[. John Wiley & Sons.](http://paperpile.com/b/LiyY8G/H06F)

[Yılmaz, Atınç. 2021. “Ozone Level Prediction with Machine Learning Algorithms.”](http://paperpile.com/b/LiyY8G/LagG) *[Journal of Agricultural Science and Technology: JAST](http://paperpile.com/b/LiyY8G/LagG)* [14 (2): 177–83.](http://paperpile.com/b/LiyY8G/LagG)

[Yusoff, Mohd Nor Hakimin. 2023.](http://paperpile.com/b/LiyY8G/zFnz) *[Industry Forward and Technology Transformation in Business and Entrepreneurship: Proceedings of the International Conference on Entrepreneurship, Business and Technology (InCEBT) 2022](http://paperpile.com/b/LiyY8G/zFnz)*[. Springer Nature.](http://paperpile.com/b/LiyY8G/zFnz)

[Zheng, Hong, Yunhui Cheng, and Haibin Li. n.d. “Investigation of Model Ensemble for Fine-Grained Air Quality Prediction.” Accessed February 22, 2024.](http://paperpile.com/b/LiyY8G/twTZ) <https://ieeexplore.ieee.org/abstract/document/9147144/?casa_token=oQJQNjs4c3EAAAAA:AcKWNWxD2gzI_4KR2inifK7zNKOOUpxFZKbSEK8-SlOlRjv5173p8ZWWQCBknYeZonIF9Ppqy4r7>[.](http://paperpile.com/b/LiyY8G/twTZ)

Alloghani, Mohamed Ahmed. 2023. Artificial Intelligence and Sustainability. Springer Nature.

Balamurugan, Vigneshkumar, Jia Chen, Adrian Wenzel, and Frank N. Keutsch. 2023. “Spatiotemporal Modeling of Air Pollutant Concentrations in Germany Using Machine Learning.” Atmospheric Chemistry and Physics 23 (17): 10267–85.

Barthwal, Anurag, Debopam Acharya, and Divya Lohani. 2021. “Prediction and Analysis of Particulate Matter (PM2.5 and PM10) Concentrations Using Machine Learning Techniques.” Journal of Ambient Intelligence and Humanized Computing 14 (3): 1323–38.

Chowdary, Vasireddy Radhika, Raavi Nikitha, Punati Sri Mahalakshmi Harika, Satish Anamalamudi, and Murali krishna Enduri. n.d. “Air Quality Analysis and Forecasting Using Deep Learning.” Accessed February 22, 2024. https://ieeexplore.ieee.org/abstract/document/9844321/.

Dang, Tran Khanh, Josef Küng, Tai M. Chung, and Makoto Takizawa. 2021. Future Data and Security Engineering. Big Data, Security and Privacy, Smart City and Industry 4.0 Applications: 8th International Conference, FDSE 2021, Virtual Event, November 24–26, 2021, Proceedings. Springer Nature.

Dinh, Thuan Nguyen, and Nam Phan Hoang. 2021. “Air Pollution Forecasting Using Regression Models and LSTM Deep Learning Models for Vietnam.” Future Data and Security Engineering. Big Data, Security and Privacy, Smart City and Industry 4.0 Applications, 264–75.

Gupta, Rajeev Kumar, Arti Jain, John Wang, Santosh Kumar Bharti, and Samir Patel. 2023. Artificial Intelligence Tools and Technologies for Smart Farming and Agriculture Practices. IGI Global.

Liu, Jianzhao, Yunjiang Zuo, Nannan Wang, Fenghui Yuan, Xinhao Zhu, Lihua Zhang, Jingwei Zhang, et al. 2021. “Comparative Analysis of Two Machine Learning Algorithms in Predicting Site-Level Net Ecosystem Exchange in Major Biomes.” Remote Sensing 13 (12): 2242.

Magdi, Dalia, Ahmed Abou El-Fetouh, Mohamed Mamdouh, and Amit Joshi. 2023. Green Sustainability: Towards Innovative Digital Transformation: Proceedings of ITAF 2023. Springer Nature.

Mantas, J., P. Gallos, and E. Zoulias. 2023. Healthcare Transformation with Informatics and Artificial Intelligence. IOS Press.

Mata-Rivera, Miguel Félix, Roberto Zagal-Flores, and Cristian Barria-Huidobro. 2023. Telematics and Computing: 12th International Congress, WITCOM 2023, Puerto Vallarta, Mexico, November 13–17, 2023, Proceedings. Springer Nature.

Moursi, Ahmed Samy, Nawal El-Fishawy, Soufiene Djahel, and Marwa Ahmed Shouman. 2021. “An IoT Enabled System for Enhanced Air Quality Monitoring and Prediction on the Edge.” Complex & Intelligent Systems 7 (6): 2923–47.

OECD. 2016. The Economic Consequences of Outdoor Air Pollution. OECD Publishing.

Pan, Qilong, Fouzi Harrou, and Ying Sun. 2023. “A Comparison of Machine Learning Methods for Ozone Pollution Prediction.” Journal of Big Data 10 (1): 1–31.

Sun, Hongying, Siyi Chen, Xinyi Li, Liping Cheng, Yipei Luo, and Lingli Xie. 2023. “Prediction and Early Warning Model of Mixed Exposure to Air Pollution and Meteorological Factors on Death of Respiratory Diseases Based on Machine Learning.” Environmental Science and Pollution Research 30 (18): 53754–66.

Tripathy, Adyant, Divyansh Vaidya, Arpita Mishra, Shardul Bilolikar, and Vijayetha Thoday. n.d. “Analysing and Predicting Air Quality In Delhi: Comparison of Industrial and Residential Area.” Accessed February 22, 2024. https://ieeexplore.ieee.org/abstract/document/9645787/.

Umamaheswari, K., B. Vinoth Kumar, and S. K. Somasundaram. 2023. Artificial Intelligence for Sustainable Applications. John Wiley & Sons.

Yusoff, Mohd Nor Hakimin. 2023. Industry Forward and Technology Transformation in Business and Entrepreneurship: Proceedings of the International Conference on Entrepreneurship, Business and Technology (InCEBT) 2022. Springer Nature.

Zheng, Hong, Yunhui Cheng, and Haibin Li. n.d. “Investigation of Model Ensemble for Fine-Grained Air Quality Prediction.” Accessed February 22, 2024. https://ieeexplore.ieee.org/abstract/document/9147144/?casa\_token=oQJQNjs4c3EAAAAA:AcKWNWxD2gzI\_4KR2inifK7zNKOOUpxFZKbSEK8-SlOlRjv5173p8ZWWQCBknYeZonIF9Ppqy4r7.

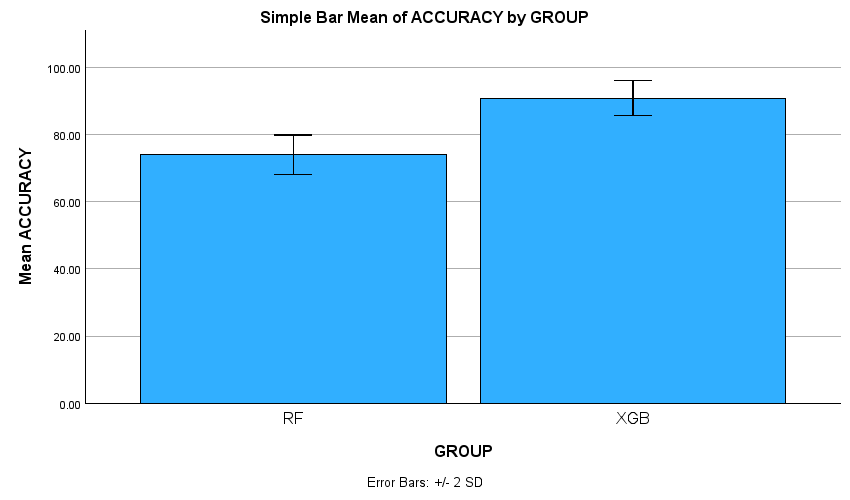
**TABLES AND FIGURES**

**Table 1**. The recorded data from 20 iterations were subjected to group statistics to compare the performance of the XGBoost (XGB) method, achieving an accuracy of 93.06%, with that of Random Forest, which demonstrated an accuracy of 44%. In contrast, the XGB algorithm exhibits a notably superior level of accuracy.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Algorithm** | **N** | **Mean** | **Std. Deviation** | **Std. Error Mean** |
| **Accuracy** | XGB | 10 | 90.87 | 2.61451 | .82678 |
| RF | 10 | 74.0150 | 2.91535 | .92192 |

**Table 2.** Group statistics on data from 20 iterations show that XGBoost (93.06%) significantly outperforms Random Forest (44%) in accuracy, emphasising XGB's superior performance.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | Levene’s Test for Equality of Variance | | T-test for Equality of Means | | | | | | |
| f | Sig | t | df | Sig.(2-tailed) | Mean Difference | Std.Error Difference | 95% Confidence of the Differences | |
| Lower | Upper |
| Accuracy | Equal variances assumed | .825 | .376 | 13.611 | 18 | <.001 | 16.8550 | 1.23834 | 14.2534 | 19.4566 |
| Equal variances not assumed |  |  | 13.611 | 17.791 | <.001 | 16.85500 | 1.23834 | 14.25114 | 19.45886 |



**Fig.1.** Comparing the extreme gradient boosting (XGB) and Random Forest algorithms for mean and accuracy, XGB demonstrates higher mean accuracy. The graphical representation illustrates this comparison, depicting XGB and KNN on the X-axis and mean accuracy on the Y-axis with a margin of ±2 standard deviations. The results underscore the superior performance of the extreme gradient boosting algorithm in accuracy.