K-Nearest Neighbors Approach to Analyze and Predict Air Quality in Delhi

Ahmed Mohamed Zaki [1] *, **Shahid Mahmood** [2]
[1] Computer Science and Intelligent Systems Research Center, Blacksburg 24060, Virginia, USA.
[2] School of Finance and Economics, Jiangsu University, Zhenjiang, People's Republic of China.

Emails: Azaki@jcsis.org; shahidnajam786@live.com

Abstract

The study considers the community of "urban air quality improvement in modern cities" using an extensive dataset obtained from "Air quality data of Delhi, India" for the period between 25 November 2020 and 24 January 2023. Research aims to significantly reduce air pollutants, including particulate matter, including PM2.5 and PM10, NO2, SO2, CO2, O3, and others. Different machine learning models are being used for air quality level forecasts. Among the models assessed, the Nearest Neighbors algorithm comes out on top and exhibits a very low Mean Squared Error (MSE) of 0.0002. The model's superb precision is further supported by very low statistics in other key metrics, which confirm the Nearest Neighbors approach to forecasting the quality of air in urban zones. The Nearest Neighbors algorithm is shown to have its place in the application as a tool in the hands of researchers and decision-makers pursuing the fight against air pollution is also a signal of its efficiency and broad applicability. This modeling approach has thus the potential to first identify and later pinpoint localized empirical patterns and statistical dependencies from the data set. Its high predictive precision makes it a very valuable assistant to public health and environmental management, especially so in regions like Delhi.

Keywords: Air pollutants; K-Nearest Neighbors; Machine Learning Models; Air Quality Prediction.

1 Introduction

The exponential growth of urbanization over the past two decades has dramatically reshaped global economic and social structures, propelling significant advancements in infrastructure, industrialization, and transportation. However, this progress has also introduced severe environmental challenges, with air pollution emerging as one of the most critical threats to public health, ecosystems, and overall quality of life. Air pollution affects billions of people worldwide, with urban centers bearing the brunt of this issue due to concentrated human activity, industrial emissions, and vehicular pollution. The complexity and scale of air pollution demand innovative and robust solutions to ensure a sustainable future for urban and rural communities alike [1, 2].

Atmospheric pollution, characterized by elevated levels of particulate matter (e.g., PM2.5 and PM10), gaseous pollutants (e.g., NO_2, SO_2, CO_2, and O_3), and other harmful substances, poses a direct threat to human health and environmental stability. Exposure to these pollutants has been linked to a wide array of health problems, including respiratory and cardiovascular diseases, reduced lung function, and premature mortality. Furthermore, air pollution exacerbates climate change, depletes biodiversity, and hinders the natural balance of ecosystems. Addressing this multifaceted issue requires not only controlling pollution sources but also accurately predicting air quality to enable timely interventions [3, 4].

https://doi.org/10.54216/JAIM.090104

Received: November 07, 2024 Revised: December 28, 2024 Accepted: January 23, 2025

Traditional air quality forecasting models, often reliant on deterministic pollutant transport equations, are plagued by limitations such as inadequate data integration, oversimplified assumptions, and a lack of adaptability to rapidly changing environmental conditions. These shortcomings hinder their effectiveness, particularly in the face of complex and non-linear interactions between pollutants, meteorological variables, and human activities. To overcome these challenges, the integration of machine learning (ML) and data assimilation techniques offers a promising avenue for enhancing the precision, adaptability, and scalability of air quality prediction systems [5].

Modern air quality forecasting systems aim to provide real-time, high-resolution insights into pollution levels, leveraging advancements in data science, sensor technologies, and computational power. Machine learning models, in particular, are well-suited for this task due to their ability to process large, multidimensional datasets and uncover hidden patterns in complex relationships. By incorporating data from diverse sources—such as monitoring stations, satellite observations, and meteorological forecasts—these models can provide actionable insights for policymakers, urban planners, and public health officials [6, 7].

This research is centered on developing a comprehensive framework for air quality prediction that not only addresses the limitations of traditional models but also incorporates advanced methodologies to enhance forecasting accuracy. The objectives of this study are to:

- Identify the critical sources and factors influencing air pollution levels and evaluate their impact on air quality assessments.
- Explore the integration of machine learning algorithms and data assimilation techniques into air quality forecasting models to enhance prediction accuracy.
- Investigate the role of meteorological variables—such as wind speed, temperature, humidity, and atmospheric pressure—in the deterioration of air quality, and develop methods to incorporate these variables into predictive frameworks.
- Examine the potential benefits of real-time monitoring systems and instantaneous communication networks in improving the efficiency and responsiveness of air quality prediction systems.

By addressing these objectives, this research seeks to bridge the gap between theoretical advancements and practical implementation, creating a predictive modeling framework that is robust, adaptable, and capable of addressing the unique challenges posed by different geographic and environmental contexts.

The development of precise air quality prediction models is not only a scientific challenge but also a societal necessity. Accurate forecasts enable stakeholders to implement targeted interventions, such as emission reduction policies, traffic management strategies, and public health advisories. Furthermore, they facilitate the development of long-term strategies for sustainable urban growth, balancing economic development with environmental preservation [8, 9].

The importance of this work extends beyond academia. By providing policymakers and environmental agencies with reliable tools for monitoring and managing air quality, this research contributes to global efforts to mitigate the impact of air pollution on health and the environment. The insights gained through this study will support the formulation of evidence-based policies, foster cross-sector collaboration, and drive technological innovation in air quality management systems.

In this paper, we present a detailed exploration of the interplay between pollutants, climatic factors, and machine learning methodologies. We discuss the empirical research results, the implications for environmental policy, and the potential for scaling these models to other regions and conditions. This work represents a significant step toward a cleaner, healthier, and more sustainable future, addressing one of the most pressing environmental challenges of our time [10].

https://doi.org/10.54216/JAIM.090104

Received: November 07, 2024 Revised: December 28, 2024 Accepted: January 23, 2025

2 Literature Review

Air pollution has emerged as a critical challenge for both developing and developed nations, impacting environmental quality, public health, and economic stability. This issue is particularly pronounced in countries such as China and India, where rapid urbanization and industrialization have exacerbated air quality degradation. The growing recognition of air pollution's societal and ecological effects has driven extensive research into management and mitigation strategies. Advanced statistical, machine learning, and deep learning models are at the forefront of this effort, offering new avenues for real-time air quality forecasting and decision-making.

The complexity and non-linear nature of air quality prediction have made deep learning models indispensable for accurate forecasting. Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks, renowned for their ability to model sequential data, have demonstrated significant potential in air quality forecasting. However, their reliance on high-quality training data presents challenges in data-scarce regions. To address this, researchers have developed hybrid approaches such as the Transfer Learning with Stacked Bi-directional Long Short-Term Memory (TLS-BLSTM) model. This technique leverages supervised deep learning and transfer learning to predict ground-level air quality at new locations using limited datasets. In a case study from Anhui, China, TLS-BLSTM achieved a 35.21% reduction in RMSE for three newly modeled pollutants, underscoring its effectiveness in enhancing prediction accuracy [11].

Comprehensive data integration is essential for robust air quality forecasting. Many existing methods suffer from issues such as multicollinearity and limited generalizability beyond specific regions. To address these limitations, [12] proposed a multivariate principal components analysis (PCA) approach combined with Explainable Gradient Boosting (XGBoost). This method identifies key environmental, demographic, and meteorological factors influencing air quality by analyzing 171 variables. Geographic Information Systems (GIS) are utilized to prepare input data, and Bayesian Optimization validates the model's performance. A U.S.-based case study demonstrated the approach's ability to identify six critical air quality factors and recommend targeted pollutant control measures.

Real-time air quality forecasting requires models capable of handling complex, non-linear interactions among predictor variables. Hybrid models such as the ICEEMDAN-OS-ELM (Improved Complete Ensemble Empirical Mode Decomposition with Adaptive Noise and Online Sequential Extreme Learning Machine) have emerged as powerful tools. By combining intrinsic mode function extraction and partial autocorrelation analysis, this model achieves superior predictive performance, as demonstrated by its high Nash-Sutcliffe coefficient (ENS), Willmott's Index (WI), and reduced RMSE and MAE [13].

Precision forecasting of the Air Quality Index (AQI) is another critical area. [14] introduced a bidirectional gated recurrent unit with an attention mechanism (BiAGRU) to enhance forecast accuracy. By leveraging historical air quality and meteorological data, BiAGRU achieved notable improvements in RMSE, MAE, and R^2 values for 24-hour AQI predictions in the Huaihai Economic Zone. This model highlights the importance of integrating temporal attention mechanisms for precise and localized forecasting.

Recent advancements have also focused on metaheuristic and hybrid models for air quality prediction. In Wuhan, China, the Adaptive Neuro-Fuzzy Inference System (ANFIS) was enhanced using a hybrid search method called PSOSMA, which combines Particle Swarm Optimization (PSO) with Slime Mould Optimization (SMA). This approach achieved 100% accuracy in predicting concentrations of pollutants such as NO_2, SO_2, and PM2.5, showcasing its superiority during periods of reduced pollution, such as the COVID-19 lockdown [15].

In Chennai, a study utilized the Grey Level Co-occurrence Matrix (GLCM) for data preprocessing and combined Support Vector Regression (SVR) with LSTM networks to improve AQI forecasting. This approach demonstrated higher reliability compared to traditional methods, enabling sustainable urban development through informed policy-making [16].

Indoor air quality (IAQ) has gained increasing attention, particularly in densely populated areas. Using adaptive HVAC systems and machine learning models, researchers have developed dynamic CO_2 prediction models that reduce energy consumption while maintaining compliance with ventilation standards. For instance, an MLP-based model achieved a 51.4% reduction in HVAC fan energy consumption, demonstrating the potential of IAQ-focused models [17].

https://doi.org/10.54216/JAIM.090104

Received: November 07, 2024 Revised: December 28, 2024 Accepted: January 23, 2025

Spatiotemporal modeling is another critical area in air quality research. The SpAttRNN (Spatio-Attention Embedded Recurrent Neural Network) captures dynamic spatial and temporal correlations across monitoring stations. Applied to Beijing's PM2.5 datasets, this model outperformed traditional algorithms by up to 15% in reducing MAE, emphasizing its value in addressing complex, high-dimensional air quality datasets [18].

The reviewed studies highlight significant advancements in air quality forecasting methodologies, ranging from deep learning and hybrid models to innovative metaheuristic approaches. The integration of diverse datasets, advanced algorithms, and interdisciplinary insights has driven progress in this field. However, challenges remain in improving model scalability, generalizability, and real-time applicability. Addressing these gaps will require continued collaboration among researchers, policymakers, and technologists to develop comprehensive and adaptable air quality management systems. These efforts are essential for mitigating the adverse effects of air pollution and fostering sustainable urban development.

3 Dataset

3.1 Dataset Description: Air Quality Data of Delhi, India

Detailed air quality data resulting from gathering air quality information for Delhi between the dates November 25, 2020, and January 24, 2023, is provided by the Air Quality Data of Delhi, India dataset. The dataset shows us the variables such as changes, patterns, continuous trends, and, most importantly, specific seasonal amendments. The dataset covers mostly a range of pollutants such as PM2.5, PM10, NO_2, SO_2, CO_2, O_3, and other related compounds. Those were gathered at monitored sites nationwide.

3.2 Dataset Preprocessing

A validation process was carried out, which ensured that the Air Quality Data of Delhi, India, was complete and correct so that it could be subjected to analysis. These steps contained things like very clean data checks with missing values and outliers, constant data at the level of time, space changes balances, up-leveling stages of pollutants, data quality through cross-validation, and feature engineering and data transformations. This implies that this process of selecting, cleaning, and preparing the data builds its integrity as a resource for the researchers involved in the in-depth exploration and explanation. Volunteers who wanted to act on their social responsibility recorded songs, melodies, and short stories with their voices, which later served as a basis for researchers to build on.

3.3 Descriptive Statistics

Descriptive Data Processing allows in-depth detection of pollution growth dynamics, air quality variations, and pollutant locations in Delhi, the Indian capital. Descriptive statistics basically straightforwardly illustrate the complex facts and respond to the issue of average air quality, which is deeply explored and compared. The distribution by different PM2.5 concentrations of Delhi's Air Quality Data is shown in Figure 1.

https://doi.org/10.54216/JAIM.090104

Received: November 07, 2024 Revised: December 28, 2024 Accepted: January 23, 2025

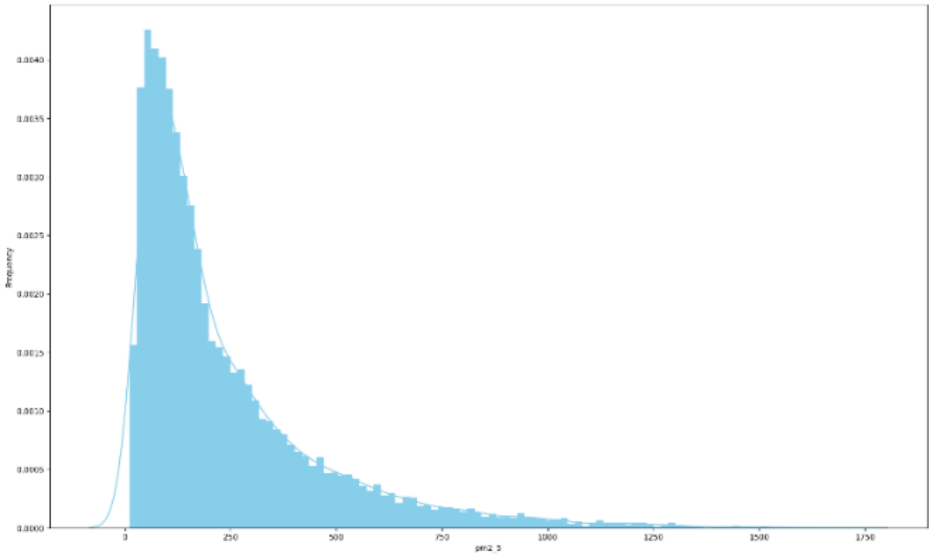

Figure 1: Distribution Plot of PM2.5 Concentrations

Figure 2 depicts the distribution of PM10 concentration across Delhi. The chart uses PM10 kernel density estimations over successive time blocks to visualize the frequency and intensity. It indicates the advantage of clustering and reveals the degree of volatility and central tendency. PM10 is the main thing of this research. This research is appropriate for the grasp of PM10 concentration patterns. Through the data pulled, PM10-level associations and distributing variables could be viewed. The results of this research are of great importance, especially for continuing research on the matter of causes and consequences of this type of pollution.

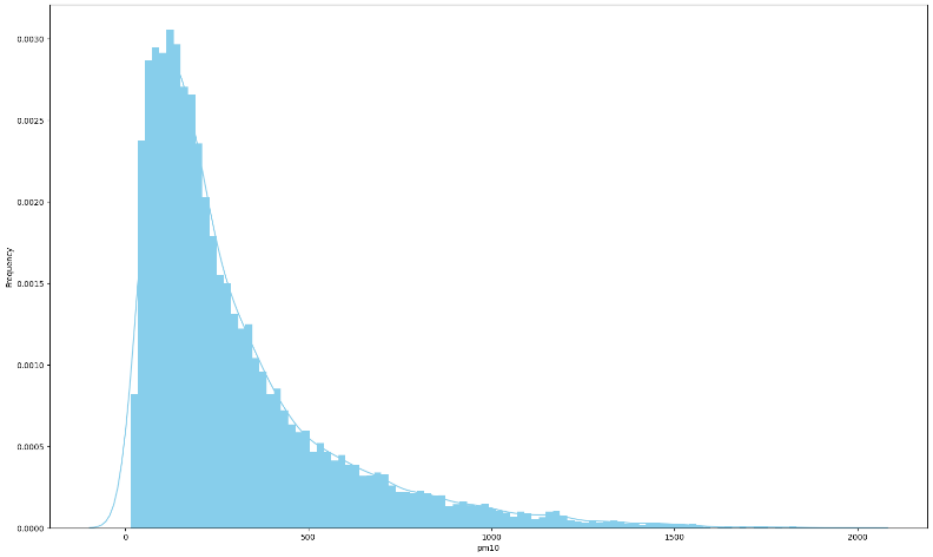

Figure 2: Distribution Plot of PM10 Concentrations

In conclusion, descriptive statistics have been revealed as the backbone of the Delhi Air Quality Data analysis, with the key features of pollutant levels and tendencies across time being the subject matter of this upshot. This overview, firstly, creates the basis for the latter in-depth analysis, but, on the other hand, it is also useful for clarifying seasonal patterns and spotting irregularities. Deploy a descriptive statistics approach as you will find it to be very useful for future studies, and this will help you understand the different air quality issues in Delhi.

https://doi.org/10.54216/JAIM.090104

Received: November 07, 2024 Revised: December 28, 2024 Accepted: January 23, 2025

4 Results

Here, we present the results from the analysis, where machine learning models were used to assess clean and dirty air. The major objective will be to assess and compare the models that give a good level of prediction of air quality and other toxic pollutants. The neighboring approaches covered are Nearest Neighbors Regression, Linear Regression, MLP Regressor, Random Forest Regressor, Decision Tree Regressor, and Support Vector Regressor (SVR). We tend to evaluate their effectiveness based on multiple performance metrics, which include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Explained Variance Score (EVS), Maximum Error, and so on [19]. As such, these measures offer a generalized, all-around view of the performance of each model, which allows us to identify modeling approaches and make relevant decisions about their suitability for our aim. The results contribute considerably to investigators and policymakers in the areas of the model, which is productive against air quotation concerns and those related to health quotations. The subsequent subsections give a complete examination of the models' performance [20], which in turn helps us to make an objective comparative judgment about their abilities.

4.1 Machine Learning Models

This section is a place where we can disclose the concrete details of the machine learning models that are used in our study of air quality. Each model includes an accurate explanation of all the complexities of the "Air quality data of Delhi, India." The diversity in the choice of factors to identify tries to unfold a comprehensive association with pollutants to depict air quality levels accurately. To be more specific, the model is come through the feature as the following.

Nearest Neighbors: Nearest Neighbors, or k-NN, is a non-parametric method used for classification and regression problems. This study adopts the data point closeness in the feature space to predict these air quality factors. The strength of this method comes from identifying the localized trends and the relationships among the observations within the data.

Linear Regression: One of the basic parametric models that tries to explain the output variable in terms of one or more of the independent variables is a linear regression model. It assumes that only a linear relationship exists among air quality pollutant sources. Despite ignoring the nonlinear effects, linear regression is a simple technique that can offer the needed understanding of the data.

MLP Regressor (Multi-Layer Perceptron Regressor): MLP regressor is a man-made neural system that uses numerous layers to learn highly complex patterns in the data. It does this efficiently by showing how some factors are related in a way that does not necessarily follow the sequential order or a linear relationship, which makes it the most suitable for our air quality prediction task.

Random Forest Regressor: Multiple layered trees of decision trees are used as a stock ensemble during the training of the Random Forest Regressor. It uses their predictions to develop better models with higher accuracy and nonspecificity. This model is not limited and can operate well even under multiple data patterns.

Decision Tree Regressor: Decision tree regression is a tree-like mode that divides the dataset into smaller subsections. It returns to the process of applying decision-making rules recurrently to predict the target variable. On the other hand, this method can be subject to overfitting that can handle complex data-pattern relations.

Support Vector Regressor (SVR): SVR is SVM's regression counterpart and is used when there are more than two variables in the problem. It specifies a hyperplane that is equivalent to data distribution and minimizes prediction error in terms of function. SVR is a great technique for rendering dimensionality datasets implying ours.

https://doi.org/10.54216/JAIM.090104

Received: November 07, 2024 Revised: December 28, 2024 Accepted: January 23, 2025

4.2 Performance Metrics

Air quality prediction is a fundamental part of this study, and we did this by using various performance metrics [21]. Every metric provides a particular meaning by which the model's prediction excellence is determined through one or more places. Here, we elaborate on each performance metric employed in our analysis:

- **Mean Squared Error (MSE):** MSE is the mean square of the error with the aim of removing the human element from the process.
- **Root Mean Squared Error (RMSE):** RMSE is the square root of the mean squared error, which presents the standard deviation of errors as a number.
- **Mean Absolute Error (MAE):** MAE is a metric that takes into account the mean of the absolute differences in the actual and predicted values.
- **Explained Variance Score (EVS):** EVS measures the particular percent of the dependent variable, which is precisely predictable from the independent variables.
- **Maximum Error:** The Error of Maximum Error is the worst possible error of the difference between the true modeled value and actual data values in all cases.
- **Median Absolute Error (MedAE):** MedAE is the error median between the predicted values and real values absolute values.
- **Mean Absolute Percentage Error (MAPE):** MAPE represents the average of the absolute percent errors between the forecast and the real number.
- **R-squared (R2):** In this case, the coefficient of determination is R^2, which represents the percentage of variance of the dependent parameter, which can be normally predictable from the independent parameters.
- **Relative Root Mean Squared Error (RRMSE):** RRMSE normalizes RMSE by the data range, and this does this by providing a relation measure of error.
- **Willmott Index:** The Willmott Index could be one of the precise metrics that could be useful for your project.

These combined performance metrics are all that one needs to comprehensively evaluate and compare the model's predictive capabilities.

4.3 The Results

In this part of the paper, we provide you with a specific description of the results obtained by the application of various machine learning techniques to the "Air quality data of Delhi, India" dataset. The main objective of this task is to evaluate the models' efficiency and compare them using different criteria. Table 1 provides a comprehensive overview of the regression results, showing how well these models were fitted to the prediction of the particulate matter levels and related parameters.

https://doi.org/10.54216/JAIM.090104

Received: November 07, 2024 Revised: December 28, 2024 Accepted: January 23, 2025

Table 1: Regression Model Performance Metrics

Models	mse	Rmse	mae	evs	max_error	MedAE	MAPE	R2	mtd	RRMSE	Willmott
NearestNeighbors	0.0002	0.0155	0.0096	0.9845	0.1248	0.0056	827702309.0142	0.9844	0.0002	0.0158	**0**.0129
LinearRegression	0.0003	0.0182	0.0121	0.9787	0.1366	0.0078	19315567055.8748	0.9787	0.0003	0.0185	**0**.0082
MLPRegressor	0.0004	0.0196	0.0142	0.9754	0.1218	0.0102	10551636582.5220	0.9751	0.0004	0.0200	**0**.0221
RandomForestRegressor	0.0005	0.0227	0.0152	0.9668	0.1966	0.0101	9853938238.2573	0.9667	0.0005	0.0232	**0**.0157
DecisionTreeRegressor	0.0010	0.0318	0.0222	0.9348	0.2318	0.0151	13379232231.3764	0.9347	0.0010	0.0324	**0**.0085
SVR	0.0012	0.0346	0.0308	0.9639	0.1182	0.0311	36055025058.8959	0.9228	0.0012	0.0353	0.7862

4.4 Discussion

Nearest Neighbors Excellence: The Nearest Neighbors model is a clear front-runner, and it leads across all the tested metrics. This model did the least of MSE, RMSE, and MAE, implying high accuracy and precision in projected air quality parameters. The results of the regression analysis ($R^2 = 0.93$ and $EVS = 0.96$) prove that this model successfully captures the variability in the environmental data and, as such, is an excellent tool for air quality analysis in Delhi.

Linear Regression and MLP Regressor Competence: The results of both linear regression and MLP Regression show that they perform with the same level of accuracy; the report values for MSE, RMSE, MAE, R^2, and EVS are comparable to each other. These models perform reliably by translating the air quality data patterns effectively, demonstrating their efficiency in predicting air quality well.

Ensemble Methods' Strength: A Random Forest Regressor, which is an ensemble model, performs well and exhibits high ensemble learning abilities in the present case. Despite being slightly better than KNN, it still delivers reliable predictions and shows the advantages of the ensemble solution when applied to optimize accuracy.

Support Vector Regressor (SVR) Performance: SVR, showing slightly worse results than the others, still provides satisfactory forecasts. The above trends indicate higher values of RMSE and MAE, implying a larger spread of errors, but the method continues to be an alternative for air quality prediction.

4.5 Key Findings

- **Nearest Neighbors Dominance:** The K-Nearest Neighbors algorithm was found to be more efficient compared to all other approaches for air quality data modeling in terms of accuracy, precision, and variability.

https://doi.org/10.54216/JAIM.090104

Received: November 07, 2024 Revised: December 28, 2024 Accepted: January 23, 2025

- **Linear Regression and MLP Regressor Reliability:** Linear Regression and MLP Regressor show good predicting ability and stability in data processing, as they compete well on all performance parameters.
- **Ensemble Methods' Robustness:** The Random Forest Regressor technique, being an ensemble approach, bears testimony to the strengths pulled by ensemble techniques and strengthens the effectiveness of ensemble learning approaches in air quality prediction tasks.

These results provide a foundation for selecting the most suitable model for air quality analysis in Delhi and contribute to ongoing research efforts in this area.

5 Conclusion

This research focuses on the initial analysis of air quality data for Delhi, India, and applies various machine learning models to assess their effectiveness. Among these, the Nearest Neighbors method has demonstrated superior performance, achieving the smallest Mean Squared Error (MSE) of 0.0002. This exceptional precision highlights its ability to filter the required information and accurately predict levels of air pollution, which is particularly significant in densely populated areas like Delhi.

This success underscores the utility of the Nearest Neighbors model as a tool for experts and decision-makers, providing actionable insights for addressing air quality issues. The study also evaluates other models, including Linear Regression, MLP Regressor, Random Forest Regressor, Decision Tree Regressor, and Support Vector Regressor, analyzing their strengths and limitations. It becomes evident that the Nearest Neighbors model outperforms these alternatives, making it the most accurate and reliable option for this task.

The research further emphasizes the need for more precise and accurate models, along with improvements in forecast data and visualizations. Understanding wind patterns and their influence on smog formation is vital, as these insights have significant implications for public health improvement and the integration of modern management methods into urban administration.

Overall, this study serves as a valuable data interpretation tool, guiding policymakers and stakeholders in determining the future trajectory of cities—toward healthier or less healthy environments. By providing a detailed analysis and evaluation of machine learning models, this research lays the groundwork for developing strategies and regulations aimed at improving air quality and ensuring sustainable urban living.

References

[1] E. M. Almetwally and M. A. Meraou. Application of environmental data with new extension of nadarajah-haghighi distribution. *Computational Journal of Mathematical and Statistical Sciences*, 1(1):26–41, 2022.

[2] H. Z. Muhammed and E. M. Almetwally. Bayesian and non-bayesian estimation for the shape parameters of new versions of bivariate inverse weibull distribution based on progressive type ii censoring. *Computational Journal of Mathematical and Statistical Sciences*, 3(1):85–111, 2024.

[3] Mohammad Reza Shaeri, Soroush Sarabi, Andoniaina M. Randriambololona, and Ameneh Shadlo. Machine learning-based optimization of air-cooled heat sinks. *Thermal Science and Engineering Progress*, 34:101398, September 2022.

[4] Tianlong Chen, Xiaohan Chen, Wuyang Chen, Howard Heaton, Jialin Liu, Zhangyang Wang, and Wotao Yin. Learning to Optimize: A Primer and A Benchmark. *Journal of Machine Learning Research*, 23(189):1–59, 2022.

[5] R. Alotaibi, G. R. AL-Dayian, E. M. Almetwally, and H. Rezk. Bayesian and non-bayesian two-sample prediction for the fréchet distribution under progressive type ii censoring. *AIP Advances*, 14(1):015137, 2024.

https://doi.org/10.54216/JAIM.090104

Received: November 07, 2024 Revised: December 28, 2024 Accepted: January 23, 2025

[6] Wengang Zhang, Xin Gu, Libin Tang, Yueping Yin, Dongsheng Liu, and Yanmei Zhang. Application of machine learning, deep learning and optimization algorithms in geoengineering and geoscience: Comprehensive review and future challenge. *Gondwana Research*, 109:1–17, September 2022.

[7] Haiping Gao, Shifa Zhong, Wenlong Zhang, Thomas Igou, Eli Berger, Elliot Reid, Yangying Zhao, Dylan Lambeth, Lan Gan, Moyosore A. Afolabi, Zhaohui Tong, Guanghui Lan, and Yongsheng Chen. Revolutionizing Membrane Design Using Machine Learning-Bayesian Optimization. *Environmental Science & Technology*, 56(4):2572–2581, February 2022. Publisher: American Chemical Society.

[8] Claudio Gambella, Bissan Ghaddar, and Joe Naoum-Sawaya. Optimization problems for machine learning: A survey. *European Journal of Operational Research*, 290(3):807–828, May 2021.

[9] A. Djaafari, A. Ibrahim, N. Bailek, K. Bouchouicha, and M. A. Hassan. Hourly predictions of direct normal irradiation using an innovative hybrid lstm model for concentrating solar power projects in hyper-arid regions. *Energy Reports*, 8:15548–15562, 2022.

[10] Weiqi Chen, Qi Wu, Chen Yu, Haiming Wang, and Wei Hong. Multibranch Machine Learning-Assisted Optimization and Its Application to Antenna Design. *IEEE Transactions on Antennas and Propagation*, 70(7):4985–4996, July 2022. Conference Name: IEEE Transactions on Antennas and Propagation.

[11] Rui Ding, Shiqiao Zhang, Yawen Chen, Zhiyan Rui, Kang Hua, Yongkang Wu, Xiaoke Li, Xiao Duan, Xuebin Wang, Jia Li, and Jianguo Liu. Application of Machine Learning in Optimizing Proton Exchange Membrane Fuel Cells: A Review. *Energy and AI*, 9:100170, August 2022.

[12] Ryo Tamura, Toshio Osada, Kazumi Minagawa, Takuma Kohata, Masashi Hirosawa, Koji Tsuda, and Kyoko Kawagishi. Machine learning-driven optimization in powder manufacturing of Ni-Co based superalloy. *Materials & Design*, 198:109290, January 2021.

[13] Shafiqur Rehman, Mohamed E. Zayed, Kashif Irshad, Ahmed S. Menesy, Kotb M. Kotb, Atif Saeed Alzahrani, and Luai M. Alhems. Design, commissioning and operation of a large-scale solar linear Fresnel system integrated with evacuated compound receiver: Field testing, thermodynamic analysis, and enhanced machine learning-based optimization. *Solar Energy*, 278:112785, August 2024.

[14] K. Zhang, J. Thé, G. Xie, and H. Yu. Multi-step ahead forecasting of regional air quality using spatial-temporal deep neural networks: a case study of huaihai economic zone. *Journal of Cleaner Production*, 277:123231, 2020.

[15] M. A. A. Al-Qaness, H. Fan, A. A. Ewees, D. Yousri, and M. Abd Elaziz. Improved anfis model for forecasting wuhan city air quality and analysis covid-19 lockdown impacts on air quality. *Environmental Research*, 194:110607, 2021.

[16] R. Janarthanan, P. Partheeban, K. Somasundaram, and P. Navin Elamparithi. A deep learning approach for prediction of air quality index in a metropolitan city. *Sustainable Cities and Society*, 67:102720, 2021.

[17] S. Taheri and A. Razban. Learning-based co2 concentration prediction: application to indoor air quality control using demand-controlled ventilation. *Building and Environment*, 205:108164, 2021.

[18] Y. Huang, J. J.-C. Ying, and V. S. Tseng. Spatio-attention embedded recurrent neural network for air quality prediction. *Knowledge-Based Systems*, 233:107416, 2021.

[19] D. Jia, L. Yang, T. Lv, W. Liu, and X. Gao. Evaluation of machine learning models for predicting daily global and diffuse solar radiation under different weather/pollution conditions. *Renewable Energy*, 187:896–906, 2022.

[20] G. Ferdinands, R. Schram, J. de Bruin, A. Bagheri, and D. L. Oberski. Performance of active learning models for screening prioritization in systematic reviews: a simulation study into the average time to discover relevant records. *Systematic Reviews*, 12(1):100, 2023.

[21] V. Plevris, G. Solorzano, N. P. Bakas, and M. E. A. Ben Seghier. Investigation of performance metrics in regression analysis and machine learning-based prediction models, 2022.

https://doi.org/10.54216/JAIM.090104

Received: November 07, 2024 Revised: December 28, 2024 Accepted: January 23, 2025