IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 1, January 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 1, January 2024 ||

| DOI:10.15680/IJIRSET.2024.1301054 |

AI for Clean Air: Advancing Smart Cities with Cutting-Edge Air Quality

Nitin Kumar[1, *], Vipin Kataria[2, *]

Marriott International, Dallas, Texas, USA[1]

Picarro Inc, Santa Clara, California, USA[2]

ABSTRACT: This paper presents a comprehensive machine learning framework for predicting the Air Quality Index (AQI) in urban environments, with the goal of promoting sustainable city development and environmental protection. The study employs a diverse set of models, including traditional regression techniques, tree-based ensemble methods, and deep learning architectures, to forecast AQI based on various air pollutants such as PM2.5, PM10, NOx, CO, SO2, and volatile organic compounds. A stacking ensemble model is used to combine the strengths of these models, with base learners like Gradient Boosting, XGBoost, and LightGBM contributing to a meta-model for enhanced predictive performance. Data preprocessing includes handling missing values, scaling numerical features, and encoding categorical data to ensure robust model performance. Evaluation of model accuracy is performed using the R^2 score, with results showing significant improvements in AQI prediction across different models. The stacking ensemble method achieved the highest test R^2 score, outperforming individual models. This research highlights the potential of AI and machine learning in addressing air quality challenges in smart cities, offering actionable insights for urban planners and policymakers. By leveraging AI, cities can more effectively monitor and improve air quality, contributing to environmental sustainability and public health.

KEYWORDS: Air Quality Index (AQI), Machine Learning, Smart Cities, Environmental Protection, Stacking Ensemble.

I. INTRODUCTION

The World Health Organization (WHO) has highlighted the significant impact of air quality on human health through various reports and initiatives. Air pollution is identified as a major public health threat, contributing to a substantial global disease burden and causing approximately seven million deaths annually [1][2]. The WHO's air quality database, updated in 2022, underscores the growing recognition of air pollution as a health priority, with data from 6,743 human settlements, a significant increase from previous years [3]. This database is crucial for understanding global exposure levels and the corresponding health impacts, particularly in high- and middle-income countries where most data is collected [3]. The WHO's air quality guidelines, revised in 2021, set stricter limits on pollutants like PM2.5, PM10, and NO2, reflecting increased confidence in their health effects even at low exposure levels [4]. Despite these guidelines, over 90% of the global population is exposed to air pollution levels exceeding WHO recommendations, with significant health impacts such as respiratory infections, cardiovascular diseases, and neurological effects [5][6][1]. The WHO's Global Air Quality Guidelines aim to address these issues by providing evidence-based recommendations to reduce emissions and improve air quality, thereby mitigating the health risks associated with air pollution [7]. Despite these efforts, the global community continues to face significant challenges in reducing air pollution levels, necessitating urgent interdisciplinary research and policy interventions to protect public health [7]. Detecting and predicting air pollution using traditional methods presents several challenges, primarily due to the complex, dynamic, and nonlinear nature of air quality data. Traditional statistical methods, such as multiple linear regression and the Kolmogorov–Zurbenko filter, have been widely used to separate the effects of emissions and meteorology on air pollutant concentrations. However, these methods often struggle to capture the nonlinear relationships between pollutant concentrations and their sources, leading to limitations in prediction accuracy [8][9]. Traditional time series prediction models, like ARIMA and EWMA, are also challenged by the highly dynamic and uncertain nature of air pollution signals, which are influenced by distinct latent physical processes and complex environmental changes over time [10]. Furthermore, traditional methods often rely on ambient concentration measurements from central-site monitors, which can introduce exposure prediction errors and misclassification, especially for spatially heterogeneous pollutants like those associated with traffic emissions [11]. The presence of outliers in data collected from surface sensors further complicates predictions, as these outliers can skew results if not accurately identified and removed [12]. Despite these challenges, traditional methods remain a foundational approach, but their limitations have spurred the integration of

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 1, January 2024 ||

| DOI:10.15680/IJIRSET.2024.1301054 |

machine learning and deep learning techniques, which offer improved capabilities in capturing complex patterns and providing more accurate predictions [13][14][15]. These advanced methods, including ensemble learning and neural networks, are increasingly being used to enhance prediction accuracy and address the shortcomings of traditional statistical approaches [16][17]. The integration of machine learning (ML) in air pollution detection and monitoring has become increasingly prevalent, offering significant advantages over traditional methods. This trend is driven by the need for more accurate, real-time, and scalable solutions to address the growing environmental and public health challenges posed by air pollution. Machine learning techniques, such as spatiotemporal modeling, data mining, and predictive analytics, have been effectively employed to enhance the detection, analysis, and forecasting of air pollution-related health issues [18]. Various ML models, including K-Nearest Neighbors, Random Forest, and Support Vector Machines, have been utilized to predict air quality, with studies indicating that these models can provide valuable insights into air quality prediction, enabling timely interventions and raising public awareness [19]. The use of IoT and sensor networks has further facilitated the collection of diverse air quality parameters, which are processed through ML algorithms to identify patterns and trends, thereby improving the accuracy of air quality predictions [20][21]. Moreover, ML-driven solutions have been shown to improve decision-making and accuracy in air quality monitoring, addressing deficiencies in traditional systems [22]. The application of ML in air pollution prediction also involves leveraging diverse data sources, such as satellite imagery and weather forecasts, to capture complex relationships between environmental variables and pollutant concentrations, thus enhancing forecasting accuracy and enabling proactive interventions [16]. Additionally, ML models have been employed to estimate contributions to air pollution from automobile emissions, highlighting the role of green technology in reducing pollution levels [23]. The development of smart air pollution detectors using ML further exemplifies the potential of these technologies to revolutionize air quality monitoring by providing real-time data and enabling targeted interventions [24]. Overall, the integration of machine learning in air pollution detection not only enhances the precision and timeliness of monitoring systems but also supports sustainable urban development and effective policy-making aimed at mitigating the adverse effects of pollution on human health and the environment [25].

II. LITERATURE SURVEY

Machine learning (ML) has emerged as a pivotal tool in detecting and predicting air pollution, offering innovative solutions to enhance air quality monitoring and forecasting. Various studies have demonstrated the effectiveness of ML models in improving the accuracy and efficiency of air quality predictions. For instance, the use of Sequential Forward Selection (SFS) with Random Forest (RF) models has shown superior performance in determining the Air Quality Index (AQI), achieving an accuracy of 99.99% and high precision metrics, thus underscoring the robustness of this approach in air quality determination [25]. Similarly, the calibration of low-cost sensors using ML models like Long Short-Term Memory (LSTM) has been effective in enhancing the accuracy of particulate matter (PM2.5) data, with LSTM models achieving high R^2 scores and low Root Mean Squared Errors (RMSEs) [26]. In urban settings, ML and deep learning techniques, including convolutional and recurrent neural networks, have been integrated with IoT-enabled sensors to provide scalable solutions for real-time monitoring and forecasting, despite challenges such as data sparsity and computational demands [27]. Moreover, the application of ML models such as Lasso Regression, Support Vector Regression, and XGBoost has been instrumental in predicting PM2.5 levels, with stack models offering the most accurate predictions [28]. The deployment of Edge AI and ML algorithms in decentralized systems has further improved real-time air quality monitoring by reducing latency and bandwidth usage while maintaining high detection accuracy [29]. Additionally, ML techniques have been applied to enhance predictive capabilities in large-scale environmental data analysis, addressing challenges related to data privacy and algorithmic bias [22]. The integration of ML with urban management strategies has also been explored, providing evidence-based recommendations for policy improvements in urban air quality management [30]. Overall, these advancements highlight the transformative potential of ML in air quality monitoring, enabling timely interventions and supporting sustainable urban development [31][32].The Self Tuning Deep Regression Neural Network (STDRNN) is one such model that utilizes the Air Quality Index (AQI) dataset to predict air pollution levels, demonstrating superior performance compared to existing methods through metrics like RMSE and R2 coefficient [33]. Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) units have also been employed to forecast PM10 levels, revealing that urban traffic is not the primary contributor to air pollution, as evidenced during COVID-19 lockdowns [34]. Another innovative approach is the Enhanced-A3T-GCN model, which uses graph-based deep learning to detect anomalies in PM2.5 concentrations by embedding spatial and temporal correlations from sensor networks [35]. The integration of atmospheric physics into deep learning models, as seen in AirPhyNet, further improves prediction accuracy by incorporating physical insights, outperforming traditional models in air quality forecasting [36]. Spectrogram images of time series data have been used with CNNs and LSTMs to capture frequency content variations, enhancing prediction accuracy by identifying intricate patterns related to air quality [37]. Time series models, including RNN, LSTM, and GRU, have been effective in

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 1, January 2024 ||

| DOI:10.15680/IJIRSET.2024.1301054 |

capturing temporal dependencies and seasonal patterns in air quality data, facilitating precise monitoring and forecasting [38]. AI-driven systems, such as AQNet-CNN, have been developed for real-time monitoring, utilizing sensor networks to quickly detect pollution sources and hotspots, thereby improving urban air quality management [31]. Dense Residual Convolutional Networks combined with Bi-LSTM have shown improved accuracy in predicting pollutants like CO, SO2, and NO2 by integrating spatial and temporal features from IoT sensor data [39]. Machine learning techniques, including regression models and ensemble methods, have been explored for their ability to capture complex relationships between environmental variables and pollutant concentrations, highlighting the potential for scalable monitoring systems [16]. Finally, machine learning algorithms like Decision Trees and Random Forest have been applied to analyze air and noise pollution, demonstrating effectiveness in urban environments such as New Delhi, India [40]. Collectively, these studies underscore the significant advancements in using deep learning for air pollution detection and prediction, offering promising solutions for environmental management and public health protection. Machine learning models such as Random Forest, Support Vector Regression, and ensemble methods like AdaBoost and Gradient Boosting have been effectively applied to predict air quality indices, with some studies highlighting the superior performance of ensemble models like the Stack Model and AdaBoost in forecasting PM2.5 levels and overall air quality [28][32]. The use of big data analytics frameworks, such as the Learning based Air Pollution Analysis (LbAPA), has been proposed to improve the automatic detection of air quality across multiple cities, demonstrating significant accuracy improvements, particularly with Random Forest models achieving an accuracy of 82.80% [41]. Additionally, the deployment of decentralized systems using Edge AI and IoT sensors has shown promise in reducing latency and bandwidth usage while maintaining high accuracy in pollutant detection, thus addressing the limitations of traditional centralized monitoring systems [29]. The integration of machine learning with sensor networks allows for real-time monitoring and prediction, providing actionable insights for pollution control and management [21][24]. Furthermore, the application of machine learning in air pollution epidemiology is gaining traction, with techniques like spatiotemporal analysis and geographic data mining enhancing the precision of pollution-related health risk assessments [18]. These advancements underscore the transformative potential of machine learning in creating scalable, efficient, and accurate air quality monitoring systems, paving the way for proactive environmental management and policy interventions [16].

III. METHODOLOGY

The proposed methodology s shown in figure 1, employs a comprehensive machine learning framework for Air Quality Index (AQI) prediction, integrating data preprocessing, feature engineering, and ensemble-based modeling techniques. The pipeline begins with data cleaning and preprocessing, including handling missing values through mean imputation for numerical features and appropriate encoding for categorical variables using one-hot encoding to preserve interpretability. Outliers are addressed using the interquartile range (IQR) method, ensuring that extreme values do not skew model predictions. Date features are transformed to extract meaningful temporal patterns, such as year-wise trends in air pollution levels. To enhance model efficiency, StandardScaler normalization is applied to numerical features, ensuring that algorithms sensitive to feature scaling (e.g., neural networks, gradient boosting, and linear regression) operate optimally. The dataset is then partitioned into training (80%), validation (10%), and test (10%) sets, preserving temporal dependencies to mimic real-world prediction conditions. The framework employs a diverse set of regression algorithms, including traditional models (Linear Regression, Ridge Regression, SGD Regressor), tree-based ensemble methods (Random Forest, Decision Tree, Gradient Boosting, XGBoost, LightGBM), and deep learning models (MLP Regressor with multiple hidden layers and adaptive learning rates). The stacking ensemble technique integrates these models, where base learners (Gradient Boosting, MLP, Random Forest, XGBoost, and LightGBM) generate initial predictions that serve as inputs for a meta-model (Linear Regression), optimizing the final prediction. This two-tier architecture ensures that diverse modeling strategies are leveraged while minimizing individual model weaknesses. sModel performance is evaluated using the R^2 score, a standard regression metric that assesses how well the predictions explain the variance in AQI values. The comparison of training, validation, and test R^2 scores provides insights into model generalization and potential overfitting risks, particularly in complex models such as Decision Trees. Additionally, residual analysis is performed to examine the distribution of errors, further validating model stability.

By leveraging ensemble learning techniques and rigorous preprocessing, this methodology provides a scalable and effective solution for AQI prediction, offering valuable insights for environmental monitoring, policymaking, and public health initiatives. Future extensions could explore spatio-temporal deep learning models such as LSTMs and hybrid stacking architectures to further improve air quality predictions in real-time applications.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 1, January 2024 ||

| DOI:10.15680/IJIRSET.2024.1301054 |

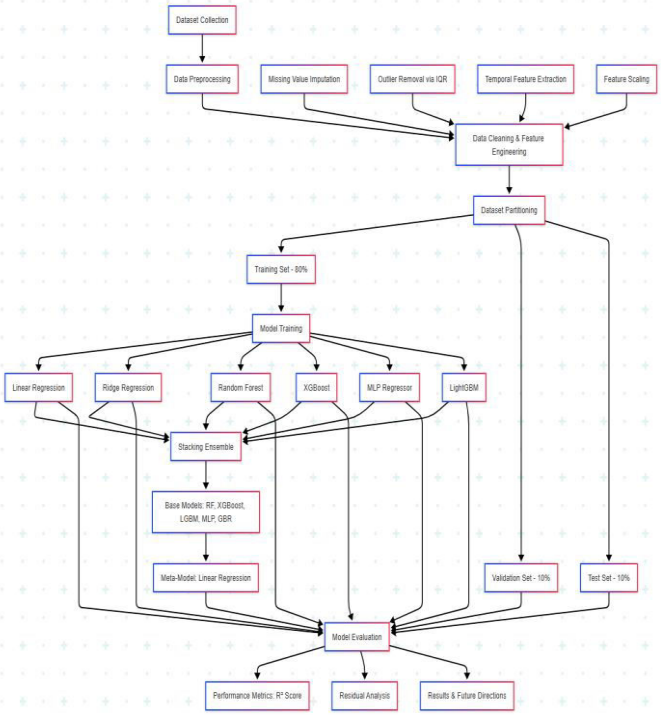

Figure 1: Process flow for the approach

The dataset used in this study is publicly[42] available and contains 29,531 records with 16 features related to air quality analysis. The features include measurements of various air pollutants, such as PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, and volatile organic compounds like Benzene, Toluene, and Xylene. Additionally, the dataset includes information about the location (City), the date of data collection (Date), and the target variable "AQI" (Air Quality Index), which serves as an indicator of overall air quality. The dataset contains some missing values, with certain features having a significant number of non-null entries, such as "City" and "Date" (both 29,531 non-null), while others, such as "PM10" (18,391 non-null) and "Xylene" (11,422 non-null), have more substantial missing data. The target variable "AQI" has 24,850 non-null entries, making it a crucial feature for predicting air quality levels.

IV. MODELS

Below is the detail of the models compared in the current study. The selection of these regression models (Linear Regression, Ridge Regression, SGD Regressor, Gradient Boosting Regressor, Decision Tree Regressor, Random Forest Regressor, MLP Regressor, XGBoost, and LightGBM) is based on their effectiveness in AQI prediction tasks. Decision tree-based models (Decision Tree, Random Forest, XGBoost, LightGBM) are included due to their ability to capture complex feature interactions and handle non-linearity effectively. Gradient Boosting Regressor is chosen for its strong predictive performance by iteratively minimizing errors. Linear Regression and Ridge Regression provide interpretable baselines, with Ridge mitigating overfitting through regularization. SGD Regressor enables scalability for large datasets using stochastic optimization. MLP Regressor, a neural network-based model, is included to capture intricate non-linear dependencies in air pollution data. By incorporating this diverse set of models, we aim to balance interpretability, accuracy, and computational efficiency for robust AQI prediction.

Linear Regression: A fundamental regression algorithm that models the relationship between dependent and independent variables using a linear equation. It estimates coefficients by minimizing the residual sum of squares, making it simple and interpretable. However, it assumes linearity and is sensitive to multicollinearity and outliers, limiting its applicability to complex datasets.

Ridge Regression: A regularized extension of Linear Regression that introduces an L2 penalty to shrink coefficients, reducing overfitting and improving generalization in high-dimensional data. Ridge Regression helps in handling multicollinearity by distributing weights more evenly among correlated features while maintaining interpretability.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 1, January 2024 ||

| DOI:10.15680/IJIRSET.2024.1301054 |

SGD Regressor: A linear regression model optimized using Stochastic Gradient Descent (SGD), which updates weights iteratively with small batches of data. This enables scalability to large datasets, making it computationally efficient. However, its performance is highly dependent on hyperparameter tuning, such as learning rate and regularization strength, to ensure convergence and prevent divergence.

Gradient Boosting Regressor: A powerful ensemble learning method that builds decision trees sequentially, with each tree learning from the errors of the previous ones. Unlike Random Forests, which build trees independently, Gradient Boosting focuses on minimizing a specific loss function using gradient descent. It excels in capturing complex non-linear relationships and is widely used in competitive machine learning but requires careful tuning to avoid overfitting.
Decision Tree Regressor: A non-parametric learning algorithm that partitions data into hierarchical structures based on feature splits, enabling interpretable, rule-based regression. The tree recursively splits data at each node based on impurity measures like mean squared error, forming leaf nodes that represent predicted values. While effective for small datasets, it is prone to overfitting and lacks robustness against noisy data.

Random Forest Regressor: An ensemble method that aggregates multiple Decision Trees, reducing variance and improving generalization. It constructs trees using bootstrapped samples and random feature selection at each split, mitigating overfitting and enhancing stability. Predictions are obtained by averaging outputs from all trees, making it a strong performer for structured data regression tasks.

MLP Regressor: A neural network-based model that consists of multiple layers of neurons, leveraging activation functions to capture non-linear dependencies in data. The model updates its weights using backpropagation and optimizers like Adam or SGD, making it highly flexible for complex regression tasks. However, it requires careful hyperparameter tuning, sufficient training data, and computational power to achieve optimal performance.

XGBoost Regressor: An advanced gradient boosting framework designed for efficiency and accuracy. It builds decision trees iteratively, correcting errors at each step, and incorporates techniques such as regularization, weighted quantile sketching, and parallel processing for improved speed and performance. XGBoost is widely used in machine learning competitions due to its robustness and superior predictive capability.

LightGBM Regressor: A highly optimized gradient boosting algorithm that enhances training speed and efficiency through leaf-wise tree growth, histogram-based feature selection, and exclusive feature bundling. It is well-suited for large-scale datasets with high-dimensional features and achieves competitive accuracy with reduced computational cost. Its ability to handle categorical variables natively and its scalability make it a popular choice for real-world regression problems.

Stacking Model: A powerful ensemble learning technique that combines multiple predictive models to enhance prediction accuracy. In this approach, base regressors such as Gradient Boosting, Multi-Layer Perceptron (Neural Network), Random Forest, XGBoost, and LightGBM generate independent predictions, which serve as inputs for a meta-model—Linear Regression in this case. The meta-model learns to optimally weight the contributions of each base model, capturing complex relationships within the data. By leveraging cross-validation, stacking mitigates overfitting and improves generalization, often outperforming individual models by synthesizing their strengths while minimizing their respective weaknesses.

Performance Metrics

The performance of the proposed air quality prediction model is evaluated using the R^2 score (coefficient of determination), a standard regression metric that measures how well the model's predictions align with actual values. R^2 represents the proportion of variance in the target variable (AQI) that is explained by the input features. It is calculated as follows:

$$R^2 = 1 - \frac{\sum(y_{true} - y_{pred})^2}{\sum(y_{true} - \bar{y})^2}$$

here y_{true}represents the actual AQI values, y_{pred} denotes the predicted AQI values, and $\bar{y}$ is the mean of the actual AQI values. An R^2 score of 1 indicates a perfect fit, where the model accounts for all variability in the data, while a score of 0 suggests that the model performs no better than predicting the mean of the target variable. A negative R^2 score implies that the model is worse than a naive prediction. In this study, models such as Gradient Boosting, Random Forest, XGBoost, and LightGBM exhibited high

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 1, January 2024 ||

| DOI:10.15680/IJIRSET.2024.1301054 |

R^2 scores, signifying their effectiveness in capturing complex relationships in the data. The Stacking Regressor achieved the highest R^2 score on the test set, demonstrating its superior generalization capability. Additionally, ensemble methods outperformed traditional linear regression models, which struggled to capture non-linearity in air pollution patterns. The evaluation of training, validation, and test R^2 scores provides insight into the model's generalization performance. A large gap between training and test scores, as observed in the Decision Tree model, indicates overfitting, while consistently low scores across all datasets suggest underfitting. By leveraging advanced ensemble techniques, this study mitigates these challenges, leading to an accurate and reliable AQI prediction framework for real-world applications.

V. EXPERIMENTAL RESULTS

Figure 2 The training R^2 scores reveal significant disparities in model complexity and learning capacity across the evaluated algorithms. DecisionTreeRegressor demonstrates the highest training performance at 99.958%, followed by RandomForestRegressor at 97.889% and XGBoost at 93.465%. This exceptionally high training score for the decision tree suggests potential overfitting, as the model appears to have memorized the training data almost perfectly. The ensemble methods, particularly RandomForestRegressor and XGBoost, also show strong training performance, which is expected given their ability to capture complex nonlinear relationships. In contrast, the linear models (LinearRegression and Ridge) exhibit identical training scores of 70.910%, indicating their limited capacity to capture complex patterns in the data. The neural network approach (MLPRegressor) achieves a moderate training score of 83.088%, suggesting reasonable learning but not excessive memorization of the training patterns.

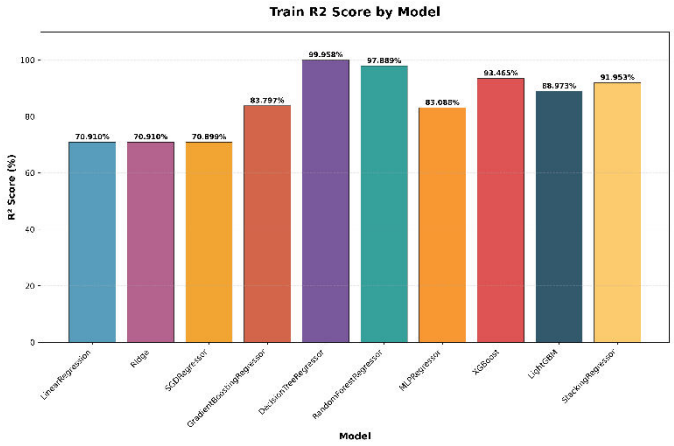

Figure 2: Train R^2 Score by Model

Figure 3 shows the validation R^2 scores provide crucial insights into model generalization capabilities and reveal the true predictive power beyond the training set. StackingRegressor emerges as the top performer with 89.281%, demonstrating superior generalization through its ensemble approach that combines predictions from multiple base learners. RandomForestRegressor and LightGBM follow closely at 85.674% and 85.574% respectively, showcasing the effectiveness of tree-based ensemble methods in maintaining consistent performance across different data splits. Notably, DecisionTreeRegressor's validation score drops dramatically to 71.392%, confirming severe overfitting as the model fails to generalize the patterns learned during training. The linear models maintain consistent performance between training and validation (around 70.5%), indicating stable but limited predictive capability. XGBoost, despite its high training score, shows a more modest validation performance at 84.566%, suggesting some degree of overfitting compared to other ensemble methods.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 1, January 2024 ||

| DOI:10.15680/IJIRSET.2024.1301054 |

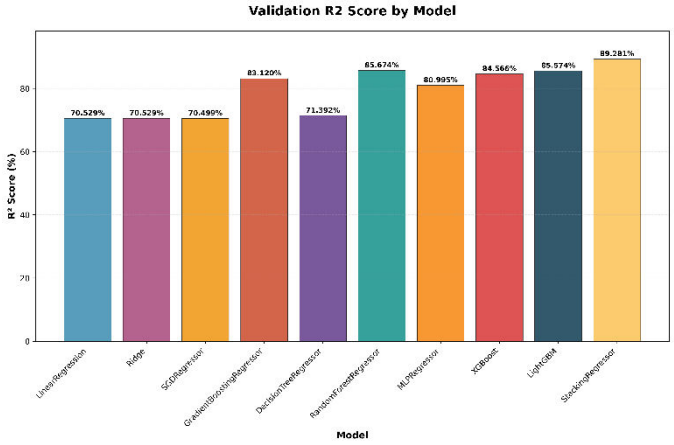

Figure 3: Validation R^2 Score by Model

Figure 4 shows the test R^2 scores represent the final assessment of model performance on completely unseen data, providing the most reliable indicator of real-world applicability. StackingRegressor maintains its superior performance with 89.962%, actually improving slightly from validation to test scores, which indicates robust generalization and suggests the model has learned genuine underlying patterns rather than dataset-specific artifacts. RandomForestRegressor and LightGBM continue to demonstrate strong performance at 85.081% and 85.210% respectively, with minimal degradation from validation to test performance. The consistency across validation and test scores for these ensemble methods indicates reliable and stable predictive capabilities. DecisionTreeRegressor shows slight improvement from validation to test (73.031%), but remains significantly below its training performance, reinforcing the overfitting concern. Linear models exhibit remarkable consistency across all three evaluation sets, with test scores virtually identical to training and validation performance, suggesting these models have reached their representational capacity for this particular dataset.

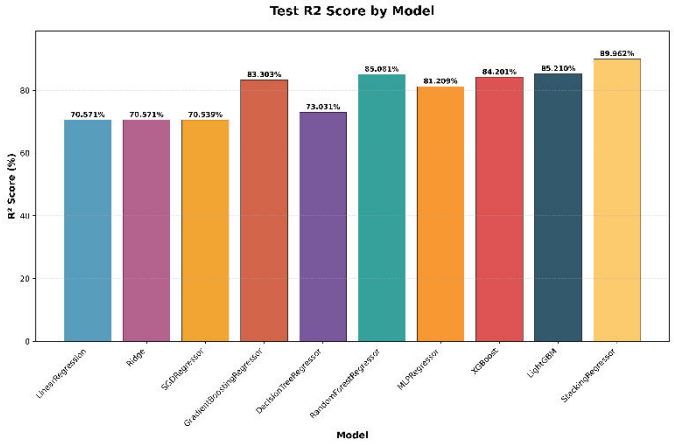

Figure 4: Test R^2 Score by Model

Figure 5 shows the comprehensive comparison across all three-evaluation metrics reveals distinct patterns in model behavior and highlights the critical importance of proper model selection based on generalization capability rather than training performance alone. StackingRegressor consistently outperforms all other models across validation and test sets, justifying its ensemble approach that leverages the strengths of multiple algorithms while mitigating individual weaknesses. The tree-based ensemble methods (RandomForestRegressor, LightGBM, and XGBoost) form a second tier of high-performing models, with relatively small gaps between training and generalization performance, indicating well-balanced bias-variance tradeoffs. The stark contrast between DecisionTreeRegressor's training and generalization performance serves as a textbook example of overfitting, where model complexity exceeds the optimal level for the given dataset size and complexity. Linear models, while showing the most consistent performance across all metrics,

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 1, January 2024 ||

| DOI:10.15680/IJIRSET.2024.1301054 |

clearly lack the representational power needed for this particular prediction task, suggesting the presence of nonlinear relationships in the underlying data that these models cannot capture effectively.

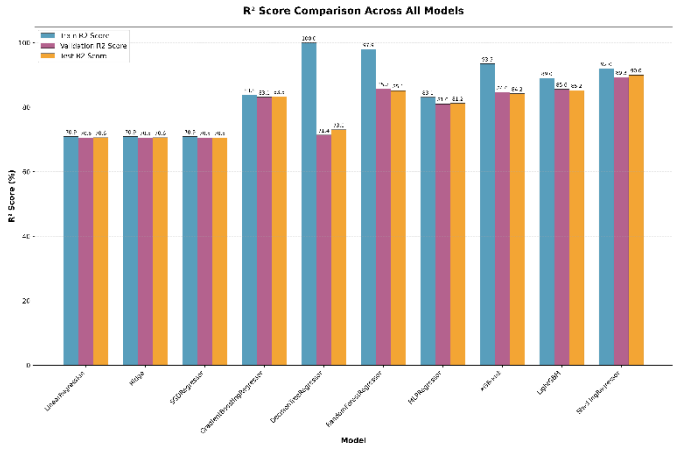

Figure 5: Comparison R^2 Score for Train, Validation and Test dataset

VI. CONCLUSION

This study introduces a machine learning framework for predicting the Air Quality Index (AQI). It combines advanced preprocessing, feature engineering, and various modeling techniques, including regressors, ensemble methods, and neural networks, to enhance predictive accuracy and robustness in air pollution data. Among the tested models, ensemble-based techniques such as **Random Forest, XGBoost, LightGBM, and Gradient Boosting** demonstrated superior performance compared to traditional regression approaches, highlighting their effectiveness in handling non-linearity and intricate feature interactions. However, the **stacking ensemble method** emerged as the most **accurate and generalized model**, outperforming all individual regressors by effectively combining their predictive strengths through a meta-model.The results underscore the importance of ensemble learning in AQI prediction, as stacking allowed the model to learn optimal weightings of base predictions, leading to **higher generalization across unseen data**. Additionally, the study highlights the role of robust **data preprocessing techniques,** including **outlier handling, feature scaling, and categorical encoding**, in improving model performance. MLP Regressor models showed promise but did not outperform ensemble methods for AQI prediction. Hybrid models can enhance results, but they have limitations like reliance on historical data and sensitivity to seasonal changes. Future research could investigate spatio-temporal modeling, LSTMs, and hybrid ensemble strategies to improve AQI predictions. Real-time IoT-based air quality monitoring can make models applicable in dynamic situations. This study offers a scalable framework for predicting air quality, aiding policymakers, environmental agencies, and researchers in refining pollution prediction and mitigation strategies.

REFERENCES

[1] L.-T. Le, K.-B. V. Quang, T.-V. Vo, T.-M. T. Nguyen, T.-V.-H. Dao, and X.-T. Bui, “Environmental and health impacts of air pollution: A mini-review,” Tạp chí Khoa học và Công nghệ Việt Nam, Mar. 2024, doi: 10.31276/vjste.66(1).120-128.
[2] R. Walters, “Air Pollution, Climate Change and International (in) Action,” Nov. 2020, doi: 10.1108/978-1-78769-355-520201028.
[3] K. K. Shairsingh et al., “WHO air quality database: relevance, history and future developments,” Bulletin of The World Health Organization, Oct. 2023, doi: 10.2471/blt.23.290188.
[4] T. Zhu, W. Wan, J. Liu, T. Xue, J. Gong, and S. Zhang, “Insights into the new <italic>WHO Global Air Quality Guidelines</italic>,” Kexue Tongbao, Mar. 2022, doi: 10.1360/tb-2021-1128.
[5] M. Iriti, P. Piscitelli, E. Missoni, and A. Miani, “Air Pollution and Health: The Need for a Medical Reading of Environmental Monitoring Data.,” International Journal of Environmental Research and Public Health, Mar. 2020, doi: 10.3390/IJERPH17072174.
[6] R. Singh, S. Ravichandran, and R. M. M. Sri, “Air pollution: A major threats to sustainable development,” International Journal of Clinical Biochemistry and Research, Oct. 2021, doi: 10.18231/J.IJCBR.2021.037.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 1, January 2024 ||

| DOI:10.15680/IJIRSET.2024.1301054 |

[7] W. Huang, X. Hong-bing, J. Wu, R. Minghui, K. Yang, and J. Qiao, "Toward cleaner air and better health: Current state, challenges, and priorities," Science, Jul. 2024, doi: 10.1126/science.adp7832.
[8] H. Zheng et al., "An intercomparison of weather normalization of PM2.5 concentration using traditional statistical methods, machine learning, and chemistry transport models," npj climate and atmospheric science, Dec. 2023, doi: 10.1038/s41612-023-00536-7.
[9] S. Karthikeyani and S. Rathi, "A Survey On Air Quality Prediction Using Traditional Statistics Method," Jul. 2020, doi: 10.32628/CSEIT2063197.
[10] "Learning Adaptive Probabilistic Models for Uncertainty-Aware Air Pollution Prediction," IEEE Access, Jan. 2023, doi: 10.1109/access.2023.3247956.
[11] H. Özkaynak, L. K. Baxter, K. L. Dionisio, and J. Burke, "Air pollution exposure prediction approaches used in air pollution epidemiology studies," Journal of Exposure Science and Environmental Epidemiology, May 2013, doi: 10.1038/JES.2013.15.
[12] M. Mahajan, S. Kumar, B. Pant, and R. Khan, "Improving Accuracy of Air Pollution Prediction by Two Step Outlier Detection," Feb. 2021, doi: 10.1109/ICAECT49130.2021.9392404.
[13] C. Chen, "A systematic survey of air quality prediction based on deep learning," alexandria engineering journal, Apr. 2024, doi: 10.1016/j.aej.2024.03.031.
[14] Y. Rybarczyk and R. Zalakeviciute, "Machine Learning Approaches for Outdoor Air Quality Modelling: A Systematic Review," Applied Sciences, Dec. 2018, doi: 10.3390/APP8122570.
[15] S. Subbiah and S. Paramasivan, "Prediction of air quality pollutants using artificial intelligence techniques: A review", doi: 10.1063/5.0183240.
[16] R. A. H. Khan and Mr. K. S. Bapu, "A Review : Air Pollution Prediction using Machine Learning Techniques," International journal of scientific research in computer science, engineering and information technology, Jun. 2024, doi: 10.32628/cseit241037.
[17] A. Mittal, S. Arora, S. Sharma, and A. Garg, "Advancements in Air Pollution Prediction and Classification Models: Exploring Emerging Frontiers," Mar. 2024, doi: 10.1109/icaccs60874.2024.10716970.
[18] D. A. Majeed et al., "Data analysis and machine learning applications in environmental management," Sep. 2024, doi: 10.22437/jiituj.v8i2.32769.
[19] A. Ahad and A. Alam, "Data Driven Air Quality Analysis and Predictions: A Machine Learning Approach," Sep. 2024, doi: 10.1109/peeiacon63629.2024.10800513.
[20] "A Pilot Study for the Detection of Air Pollution of Puducherry, India by Machine Learning Algorithms," Nanotechnology Perceptions, Jul. 2024, doi: 10.62441/nano-ntp.v20is7.8.
[21] N. U. Kumar, "Air quality monitoring using machine learning," Indian Scientific Journal Of Research In Engineering And Management, May 2024, doi: 10.55041/ijsrem32827.
[22] E. Varun, L. Rajesh, M. Lokeshwari, S. Ashwini, D. Bhat, and C. kumar S, "Machine Learning-Enhanced Water and Air Quality Monitoring Technologies and Applications," Advances in IT standards and standardization research series, Dec. 2024, doi: 10.4018/979-8-3693-4759-1.ch004.
[23] S. K. Vemula et al., "Air Pollution Detection from Automobile Emissions Using Machine Learning Model," Feb. 2024, doi: 10.1109/ic2pct60090.2024.10486576.
[24] Dr. M. S. Kumar, M. S. Vani, M. Saikumar, G. Bharath, and A. Sairam, "A Smart Air Pollution Detector Using Machine Learning," Nov. 2023, doi: 10.1109/ictacs59847.2023.10389898.
[25] N. G. Rezk, S. Alshathri, A. S. A. Mahmoud, E. E. Hemdan, and H. El-Behery, "Sustainable Air Quality Detection Using Sequential Forward Selection-Based ML Algorithms," Sustainability, Dec. 2024, doi: 10.3390/su162410835.
[26] S. Salim, I. Z. Hussain, J. Kaur, and P. Morita, "An Early Warning System for Air Pollution Surveillance: An IoT Based Big Data Framework to Monitor Risks Associated with Air Pollution," Dec. 2023, doi: 10.1109/gcaiot61060.2023.10385123.
[27] S. Salim, I. Z. Hussain, J. Kaur, and P. P. Morita, "Air Pollution Surveillance System: A Big Data Approach to Monitoring Adverse Health Outcomes for Public Health Interventions," Dec. 2022, doi: 10.1109/BigData55660.2022.10020691.
[28] N. V. Tinh and P. Q. Hieu, "Air pollution forecasting: Application of machine learning models to estimate PM2.5 index," Tạp chí Nghiên cứu Khoa học Kỹ thuật và Công nghệ Quân sự, Dec. 2024, doi: 10.54939/1859-1043.j.mst.fee.2024.286-294.
[29] V. Ramu, U. Naresh, P. A. Prakash, G. Shyamala, P. Saritha, and A. Atheeswaran, "Real-Time Air Quality Monitoring with Edge AI and Machine Learning Algorithm," Nov. 2024, doi: 10.1109/iceca63461.2024.10800783.
[30] D. Saputra, H. Nugroho, D. Julianingsih, and Z. Queen, "Understanding Air Pollution Through Machine Learning: Predictive Analytics for Urban Management," IAIC Transactions on Sustainable Digital Innovation (ITSDI), Oct. 2024, doi: 10.34306/itsdi.v6i1.679.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 1, January 2024 ||

| DOI:10.15680/IJIRSET.2024.1301054 |

[31] F. Aghazada, "Ai based approaches for real-time air quality monitoring in urban enviroments (ai based for aqm)," Oct. 2024, doi: 10.57033/tdtu.uz.279.
[32] M.-C. Vieru and M. Cărbureanu, "Machine learning methods applied in air quality prediction," Romanian Journal of Petroleum & Gas Technology, Aug. 2024, doi: 10.51865/jpgt.2024.01.01.
[33] M. B. S, "A Data-Driven Approach to Air Pollution Management with Self-Tuning Deep Regression Neural Network," International Journal of Advanced Trends in Engineering and Management, Dec. 2024, doi: 10.59544/qrqw2887/ijatemv03i11p3.
[34] M. Kostadinov et al., "Forecasting air pollution with deep learning with a focus on impact of urban traffic on PM10 and noise pollution," PLOS ONE, Dec. 2024, doi: 10.1371/journal.pone.0313356.
[35] S. Masmoudi et al., "Enhanced Sensor Environment Graph Based Deep Learning Approach for Air Quality Anomaly Detection," Aug. 2024, doi: 10.23919/eusipco63174.2024.10715088.
[36] T. Wang, "Air Quality Prediction based on Neural Network," Highlights in Science Engineering and Technology, Jun. 2024, doi: 10.54097/2fsfav47.
[37] M. Shakir, U. Kumaran, and N. Rakesh, "Deep learning approaches for air quality prediction using spectrogram images of time series," Nov. 2024, doi: 10.1201/9781003559092-45.
[38] M. D. Fathima, S. Donavalli, and H. Kambham, "Air Quality Prediction using Deep Learning models," Jun. 2024, doi: 10.1109/apci61480.2024.10616969.
[39] "A Real-time Environmental Air Pollution Predictor Model Using Dense Deep Learning Approach in IoT Infrastructure," Jan. 2024, doi: 10.30955/gnj.005666.
[40] A. Vijayalakshmi, E. Abishek.B, J. Rubi, J. Dhivya, Kavidoss. K, and A. R. A.S, "Machine Learning-Based Prediction and Analysis of Air and Noise Pollution in Urban Environments," Jul. 2024, doi: 10.1109/icscss60660.2024.10625644.
[41] V. Makula and A. Khare, "A Machine Learning Framework with Big Data Analytics for Predicting Air Pollution Across Cities," Jun. 2024, doi: 10.1201/9781003470939-8.
[42] https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details