IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 14 September 2023, accepted 29 September 2023, date of publication 10 October 2023,
date of current version 23 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3323447

RESEARCH ARTICLE

Comparative Analysis Study for Air Quality Prediction in Smart Cities Using Regression Techniques

SHOROUQ AL-EIDI[1], FATHI AMSAAD[2], OMAR DARWISH[3], (Senior Member, IEEE), YAHYA TASHTOUSH[4], ALI ALQAHTANI[5], AND NIVESHITHA NIVESHITHA[2], (Graduate Student Member, IEEE)

[1]Computer Science Department, Tafila Technical University, Tafila 66110, Jordan
[2]Computer Science and Engineering Department, Wright State University, Colonel, OH 45435, USA
[3]Information Security and Applied Computing Department, Eastern Michigan University, Ypsilanti, MI 48197, USA
[4]Department of Computer Science, Jordan University of Science and Technology, Irbid 22110, Jordan
[5]Department of Networks and Communication Engineering, Najran University, Najran 61441, Saudi Arabia

Corresponding author: Shorouq Al-Eidi (saleidi@ttu.edu.jo)

This work was supported by the Deanship of Scientific Research, Najran University, under the Research Groups Funding Program, under Grant NU/RG/SERC/12/9.

ABSTRACT In smart cities, air pollution has detrimental impacts on human physical health and the quality of living environment. Therefore, correctly predicting air quality plays an important effective action plan to mitigate air pollution and create healthier and more sustainable environments. Monitoring and predicting air pollution is crucial to empower individuals to make informed decisions that protect their health. This research presents a comprehensive comparative analysis focused on air quality prediction using three distinct regression techniques- Random Forest regression, Linear regression, and Decision Tree regression. The main goal of this study is to discern the most effective model by considering a range of evaluation criteria, including Mean Absolute Error and R^2 measures. Moreover, it considers the crucial aspects of minimizing prediction errors and enhancing computational efficiency by evaluating the regression models within two frameworks. The findings of this study underscore the superiority of the Decision Tree regression approach over the other models, demonstrating its exceptional accuracy with a high R^2 score and a minimal error rate. Moreover, integrating cloud computing technology has resulted in substantial improvements in the execution time of these approaches. This technology enhancement significantly affects the overall efficiency of the air quality prediction process. By leveraging distributed computing resources, real-time air quality forecasting becomes feasible, enabling timely decision-making and proactive measures to address air pollution episodes effectively.

INDEX TERMS Air pollution, machine learning, IoT, smart city, air quality index.

I. INTRODUCTION

Recently, the detrimental of air pollution have garnered significant global attention, as the World Health Organization studies underscored, which illuminate the impact on human health and the environment. It is alarming that air pollution was identified as a primary cause of various allergies, illnesses, and premature death, accounting for a staggering 12% of global deaths in 2019 [6]. Moreover, air pollution introduces dangerous substances into the atmosphere, including greenhouse gases and biological compounds [9], further exacerbating human-environmental challenges.

Specifically, the issue of air pollution in smart cities has gained significant attention in urban sustainability and

The associate editor coordinating the review of this manuscript and approving it for publication was Sajid Ali.

enhanced quality of life. While smart technologies have heralded remarkable efficiency and convenience, they have inadvertently become a source of air pollution. That is attributable to the concentration of industries and transportation networks within smart cities, which escalates air pollutants and harmful gases released into the atmosphere. Consequently, the urban planners within these smart cities recognize the need for innovative solutions to address this escalating problem. They leverage real-time monitoring, data analytics, and advanced approaches to accurately predict and proactively mitigate pollution levels, thereby safeguarding the well of their residents and ensuring the future sustainability of urban communities.

The Air Quality Index (AQI) emergency has recently assumed a vital role in predicting air quality. AQI clearly indicates poorer air quality and harmful gases based on predefined ranges of air pollutant concentrations [2]. Early prediction of AQI levels is instrumental in effective environmental management and preventing potential dangers of air pollution.

Given the situation of urgency, adopting sustainable solutions that effectively mitigate air pollution has become imperative, particularly when considering the well-being of future generations. Over recent years, various forecasting models have been proposed to predict pollution levels, with machine learning emerging as a noteworthy approach due to its ability to handle the intricate interplay of air quality parameters. Machine learning-based prediction systems are increasingly attractive for their precision in air quality management [8], [14], offering promising avenues for designing cleaner and healthier smart cities.

The primary objective of this study is to address the challenges of time and cost constraints in air quality prediction. It does so by leveraging the efficiency of machine learning techniques in conjunction with the AQI. To achieve this, the study compares three distinct regression approaches to provide the most accurate air quality prediction. To assess their effectiveness, well-established evaluation measures such as q Root Mean Square Error (RMSE), R^2 score, and Mean Absolute Error (MAE) are employed. The ultimate goal is identifying the most efficient and suitable regression model for predicting air quality. Beyond the accuracy, this study recognizes the real-time processing capabilities in smart cities, such as valuing the processing time associated with each regression technique. To reduce the execution time without compromising prediction accuracy, this work incorporates distributed computing techniques into its methodology. This means that optimization considerations encompass factors such as data size and processing time.

The implication of the study's finding holds significant practical relevance for formulating effective air pollution control strategies and contributes to advancements in air quality prediction methodologies. Particularly in urban environments where the monitoring of the AQI is crucial for public health and environmental management, these insights gleaned from the study can serve as invaluable inputs for decision-making processes. They can potentially guide the development of proactive measures that effectively address challenges posed by air pollution.

This paper is organized as follows. Section II provides a review of the air quality and pollution prediction literature. Section III details of air quality predication approaches, which illustrate the experimental setup pre-processing techniques and utilize regression techniques to predict air pollution levels. Section IV presents the experiment results. Section V offers a conclusion and potential future work.

II. LITERATURE REVIEWS

The field of air pollution prediction has experienced a notable rise in machine learning techniques to address the challenges associated with forecasting air quality levels. These techniques have demonstrated their effectiveness in predicting air pollution, thus contributing significantly to developing air quality management strategies. This section comprehensively explores the most notable models utilized for calculating and predicting the Air Quality Index (AQI) and the concentration levels of various air pollutants through different machine learning algorithms, such as regression techniques. These models hold considerable relevance and find practical utility in other application domains such as cloud computing.

Patil et al. [18] extensively reviewed different methodologies and techniques to analyze the concentration level of air pollution and the prediction of AQI. This study highlighted the performance of these analytical methods and presented the importance of calculating AQI as a significant measure for assessing pollution levels and how it dramatically influences human health and the environment. Similarly, Oliveri et al. [15] reviewed air quality models while discussing the effect of air pollution concentration on human health.

A noteworthy study by Ameer et al. [1] scrutinized the efficiency of four regression methods, namely Decision Tree, Gradient Boosting, Multilayer Perceptron, and Artificial Neural Network (ANN), in predicting air quality levels. These methods were evaluated based on tracking PM2.5 levels in the air and calculating the AQI. The findings of this study concluded the Random Forest regression method outperformed the others, achieving an adjusted MAE of 16% for Beijing City. This method also reduces the running time compared to Gradient Boosting and Multilayer Perceptron. Similarly, Maleki et al. [12] utilized the ANN approach to predict the concentration levels different air pollutants such as NO2 and SO2. This study applied in several monitoring areas including Naderi, Havashenasi, Behdasht, MohiteZist,and Iran. In this study the authors considered the effect of set parameters such as time, date, and meteorological data to offer a robust air quality predictive model.

Moreover, Zhang et al. [22] utilized the long short-term memory (LSTM) to proposed a deep learning approch for

air pollution detection. This study conducted a series of experiments using Detrended Cross-Correlation Analysis (DCCA) to explore the relationship between predicting levels of several air pollutants and meteorological data such as temperature and humidity, The results of this study was observed there were a negative correlation between AQI and meteorological data (temperature, humidity, and wind speed), while a strong positive correlation between pressure and AQI. Furthermore, Bougoudis [3] developed a hybrid computational method to identify the correlation between air pollutants and weather conditions to determine the actual cause of pollution. The study employed ANN and Random Forest as ensemble learning methods, claiming increased accuracy. However, the feedforward neural network faced challenges predicting continuous values due to insufficient data.

For using classification machine learning algorithms, Gore et al. [5] proposed a classification approach to study how air pollutant levels affect the health of humans. In their process, they employed Naive Bayes and Decision Tree algorithms and achieved a high accuracy using the Decision Tree model. Moreover, Simu et al. [21] presented a comparative study to compare the performance of several machine learning algorithms, such as Random Forest and Multi-linear Regression, in analyzing air pollutants and predicting air pollution levels. The study results concluded that the Multilayer Perceptron algorithm outperformed the other.

Moreover, In [19], Peng et al. utilized Multilayer Perceptron to enhance the air quality prediction accuracy. However, they noted limitations in data extension and the high computational cost because of the seasonal update of the model. Mahalingam et al. [10] proposed using ANN and SVM algorithms to predict the AQI in the smart city of Deldi with impressive accuracies, mainly the Medium Gaussian SVM function. To predict the AQI and air pollution levels, Sharma et al. [20] implied various algorithms, including Linear regression, ANNs, Lasso regression, and XGBoost regression. The study focused on tracking the values of several pollutants, including NO2, SO2, PM2.5, PM10, CO, and O3. The research findings indicated that the Random Forest algorithm outperformed the other algorithms, demonstrating its high performance in predicting the AQI and air pollution levels.

Nandini et al. [13] used Decision Trees and Multinomial Logistic Regression to forecast and analyze air quality pollutant levels, achieving better accuracy with Multinomial Logistic Regression compared to Decision Tree. Similarly, in a study by Mahanta et al., [11], a comprehensive comparison of several algorithms, including Linear regression, Decision Forest, XGBoost, ElasticNet, Boosted Decision Tree, KNN, Lasso regression, and Ridge regression to predict air pollutant levels. Among these algorithms, Extra Trees exhibited superior performance due to its technique of ranking the essential features to improve the accuracy of the predictions. Moreover, Pasupuleti et al. [17] conducted a study comparing Random Forest, Decision Tree, and Linear regression models for predicting air pollutants and meteorological conditions in the Arduino platform. The study found that the Random Forest model provided better performance by reducing errors caused by overfitting. However, it was noted that the Random Forest model required more memory and incurred higher costs.

For using the clustering approach, Kingsy et al. [7] enhanced the K-Means algorithm to analyze and identify the air pollution level. Their method calculates the correlation coefficient between pollutant data to determine the AQI value and find the air pollution level in a specific location. To validate their findings and evaluate the effectiveness of their approach, the authors compared their proposed algorithm with the Fuzzy C-Means algorithm. Their results demonstrated that the proposed K-Means clustering algorithm achieved higher accuracy and less execution time than the Fuzzy C-Means algorithm. Ganeshkumar et al. [4] presented an efficient and cost-effective classification model for environmental monitoring and air pollution prediction. Their study the authors used several artificial methods with a cloud platform for data processing, leading to significant time savings, reduced labor efforts, and producing high-quality outcomes. This research highlights the importance of integrating cloud platform solutions to enhance the efficiency and accuracy of monitoring and air quality prediction models, which is beneficial for addressing environment mentoring challenges. Similarly, Park et al. [16] used their own cloud computing technique to reduce the processing time of processing and visualization of urban air pollution data.

The literature review underscores the widespread prediction of air quality and air pollution utilizing machine learning algorithms, highlighting their potential to achieve accurate results, efficient computation, and effective prediction of air quality levels. However, certain limitations need to be addressed. These include the necessity for more extensive and more comprehensive datasets, challenges in accurately predicting continuous values, and the high computational cost associated with model updates. Additionally, the review identifies a research gap in the focus on predicting the AQI based solely on PM2.5 measurements, neglecting the inclusion of other important air pollutants. Incorporating data on multiple pollutants such as O3, NO2, SO2, and PM2.5 can significantly enhance the accuracy of air pollution prediction models. These insights provide valuable guidance for future research endeavors and for developing effective air quality management strategies, particularly in smart cities.

III. AIR QUALITY PREDICTION APPROACH

This section presents our air quality prediction approach and the stages of how to predict air pollution using regression techniques.

Our approach contains six main components: Dataset preprocessing, AQI calculation, Feature selection from

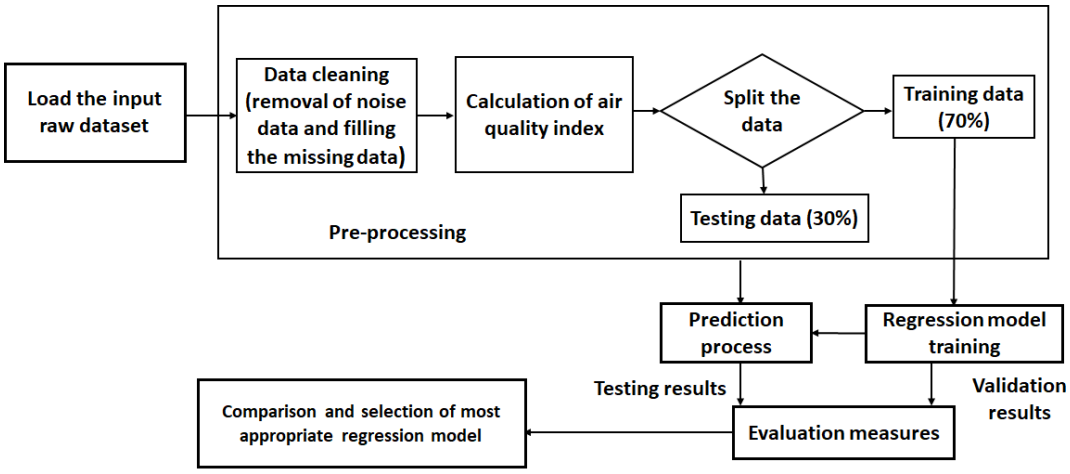


FIGURE 1. Air quality prediction model.

data, Splitting and Balancing data, and Regression model construction for air quality prediction, as shown in Figure 1. Air quality datasets were collected and loaded in the first stage for analysis. Next, preprocessing steps were applied to ensure data quality, including handling missing values and reducing outliers. Then, calculate the Air Quality Index (AQI) for air pollutants in the dataset. After processing data, our feature extraction module extracts the most relevant and essential features. This step helps reduce the dimensionality of the air dataset and focuses only on the significant variables. The dataset was then balanced to ensure equal representation of different classes, followed by splitting it into training and testing sets. Finally, our regression module takes the sets of essential features as input and constructs regression classifiers to predict air quality. Performance metrics were computed to identify a suitable and efficient model for predicting air quality. We describe the details of each module next.

A. DATASET DESCRIPTION

The dataset used in this study encompasses a comprehensive collection of 103,205 records, featuring data from monitoring stations situated across ten diverse locations within Pune City.[1] These areas include Bopadi Square 65, Karve Statue Square 5, Lullanagar Square 14, Hadapsar Gadital 01, PMPML Bus Depot Deccan 15, Goodluck Square Cafe 23, Chitale Bandhu Corner 41, Pune Railway Station 28, Rajashri Shahu Bus Stand 19, and Dr. Baba Saheb Ambedkar Sethu Junction 60. The dataset, compiled in 2019, resulted from a collaborative effort between the Pune smart city and the Indian Institute of Science, Bangalore.

Within the dataset, we focus on 28 distinct features related to air pollution, including NO2 (Nitrogen dioxide), O3 (Ozone), PM10 (Particulates) with a diameter of less than 10 microns, PM2.5 with a diameter of less than 2.5 microns, SO2 (Sulphur dioxide), CO (Carbon monoxide), and AQI. This study considers the pollutant concentration values as crucial features of the dataset, enabling a comprehensive understanding of pollution patterns in Pune's smart city.

[1] https://www.kaggle.com/datasets/akshman/pune-smartcity-test-dataset

B. DATA PRE-PROCESSING

Data pre-processing is an important step in data analysis to improve the quality and reliability of the dataset by reducing noise and inconsistencies.

The first stage of data pre-possessing is handling missing values in the raw data. The dataset used in this study comprised 103,205 entries containing several data types, such as objects, integers, and floats. Some of these entries had null or missing values, which must be addressed. To handle this issue, missing values were replaced with the mean values for pollutant parameters. This approach helped maintain the dataset by ensuring no crucial information was lost due to missing values. Moreover, the interquartile range (IQR) method addresses duplicate observations and outliers. The Interquartile range method utilizes three percentiles: quartiles Q1 (25th), Q2 (50th), and Q3 (75th), and considers the outlier as any values not in the range between $(Q1 - 1.5 * IQR)$ and $(Q3 + 1.5 * IQR)$. Instead of removing the outlier values, we used the lower and upper boundary values to replace them and retain important information while reducing the impact of data outliers on the data analysis process. Exploratory Data Analysis (EDA) has been used to gain insights into the dataset and understand its characteristics for cleaning and preparing the raw data for training purposes. EDA process conducted descriptive statistics of the dataset based on analyzing various statistical measures such as standard deviation, mean, minimum, and maximum values for each air pollutant. By calculating these statistics values, we obtained a comprehensive dataset overview, enabling us to identify potential anomalies that could affect the analysis.

TABLE 1. Basic characteristic of dataset.

Feature	Mean	Std	Min	Q1	Q2	Q3	Max
NO2	67.7	34.5	0	42.5	70.5	91.5	315.5
PM10	16.8	11.4	0	7.0	16	26	48.5
PM25	13.4	8.5	0	6.5	12.5	20	37.5
SO2	4.9	11.4	0	1.0	3	5.5	165
CO	71.2	27.8	13.5	50	75.5	89.5	144.5
OZONE	10.8	22.8	0	0	3.5	12	335

C. AQI CALCULATION

As mentioned before AQI is one of the most crucial parameter have been used for monitoring the air quality in particular cities. It provides a standard measure that quantifies air pollution and helps understand its effects on human health and environment. AQI is a numerical value within a defined range, typically from 0 to 500. A higher value of AQI indicates poorer air quality and the existence of harmful air pollutants. Each pollutant has specific constraints and specific averaging periods to ensure accurate assessment such as the period is 8-hour maximum for Q3 and 24-hour average concentrations for SO2, PM10, CO, NO2, PM2.5.

To calculate the AQI, the concentrations of these air pollutants are categorized into sub-indices. These sub-indices were defined based on predefined ranges that help to give the level of air quality, ranging from ''good'' to ''hazardous.'' Where the highest value of sub-index among the air pollutants represents the overall air quality index for a certain location. The computation of the AQI is based on Equation 1 combines the sub-indices of each pollutant [11], which considers the weightage assigned to each pollutant based on its potential health impacts, by incorporating multiple pollutants and their respective sub-indices, the AQI helps to assess the air quality [1].

$$I = \frac{I_{high} - I_{low}}{C_{high} - C_{low}}(C - I_{low}) + I_{low} \quad (1)$$

where, I is Air Quality Index, C is Pollutant concentration.

D. FEATURE SELECTION

Feature selection becomes crucial in our research following the data preprocessing and exploratory data analysis step. This process involves identifying and selecting the most relevant features related to the AQI, representing the overall air quality. The features in this study based on the preprocessed dataset contain several pollutant information such as CO, SO2, O3, OZONE, NO2, PM10, and PM2.5, along with their corresponding AQI values.

We used the correlation analysis to determine the relationship between the features and AQI. Correlation analysis can be used to find the linear relationship between two variables. By calculating the correlation coefficients between each feature and the AQI, we can assess their predictive value in understanding and predicting variations in air pollutant levels. The correlation values are compiled into a correlation matrix, which provides a view of the relationships between all dataset variables and identifies features with strong positive or negative correlations with the AQI, as shown in Figure 2. The results of this study highlighted that most values of air pollutants demonstrated a positive correlation with the AQI values, which indicates that higher concentrations of air pollutants are associated with higher AQI values, reflecting poorer air quality. This highlights how selecting important air pollutants as features that represent significant correlations with the AQI is essential in analyzing and predicting air quality variations in the study area.

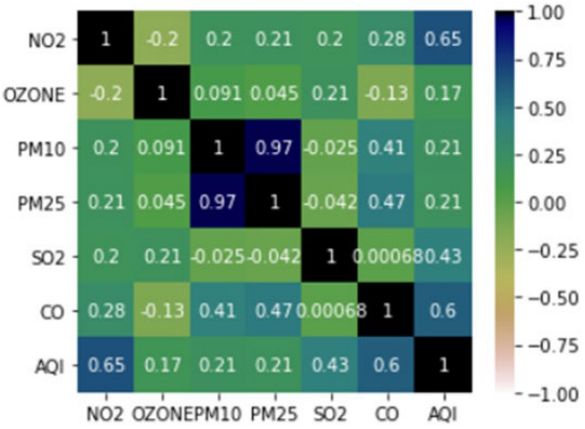


FIGURE 2. Correlation of AQI air pollutants.

E. SPLITTING DATA

In this stage, the train-test split() method was utilized to split the data into two parts with a ratio of 70:30 for training and testing sets. This means 70% of the total dataset was chosen for training, while the remaining 30% of data was assigned for testing data. With this splitting ratio, the model is trained on a large sufficient portion of the data and evaluated on test data to assess its performance.

F. BALANCING DATA

In machine learning tasks, addressing the issue of imbalanced data is a crucial process to ensure reliable and accurate prediction results. In this study, the distribution of AQI values exhibits an imbalance in the given dataset, where certain values occur more frequently than others. This can be observed by categorizing the AQI values into predefined ranges, as shown in Figure 3.

Using imbalanced data can significantly effect on the regression approaches. Biases can occur as approaches favor the majority class and overlook minority classes with fewer instance data. To get over this issue, SMOTER (Synthetic Minority Over-sampling technique for Regression with Gaussian Noise) is one of the most common techniques has been used to improve the model's performance.

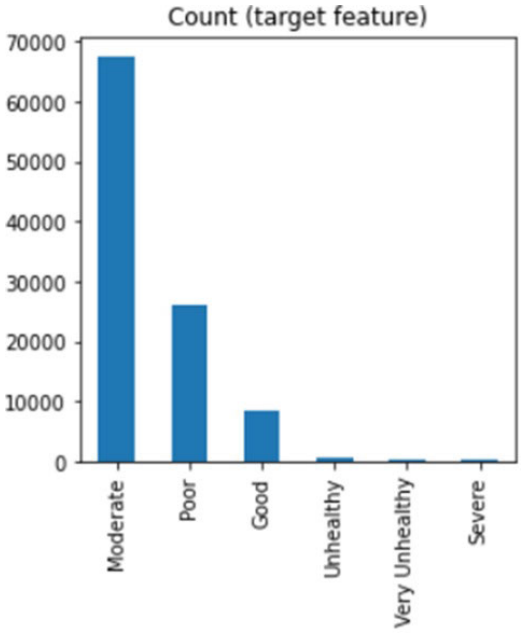


FIGURE 3. AQI classified categories.

The SMOTER technique is used to generate a synthetic minority and under-sampling the majority class, which helps to get a balanced dataset and ensures a more equitable representation of different AQI values. By generating more synthetic samples, the minority class can create a more balanced distribution of data points. Gaussian noise was also added into these synthetic samples to introduce variations and prevent overfitting. By utilizing balancing the dataset and the SMOTER technique, the regression models are trained using more representative and several sets of data points. In this stage, the model's ability to capture patterns and relationships across different AQI values of air pollutants will be enhanced, leading to improved model performance and more accurate predictions.

G. REGRESSION MODELS CONSTRUCTION

The final step in the air quality prediction approach is constructing a regression model to predict air quality. For this task, we train models using the following regression techniques.

1) **Decision Tree regression**: is a supervised machine learning algorithm commonly used to model non-linear relationships between output variables and input features. The algorithm partitions the data into subsets based on specific rules or criteria in this regression approach. These rules are selected to minimize the difference in space between the predicted and the actual values. By considering several input factors and training the model using historical air pollution and AQI data, Decision Tree regression can be applied to predict the air quality. The model analyzes the relationships between the input factors and AQI to accurately predict upcoming periods.
2) **Linear regression**: is a commonly employed statistical method in several approaches for prediction and forecasting air pollution [20]. It is used for examining the relations between pollutant concentrations and the AQI. Linear regression can make reliable predictions about future air pollution levels by analyzing historical data and discerning trends and patterns. Furthermore, Linear regression aids in identifying the primary factors contributing to air pollution. By assessing the regression coefficients, it becomes possible to determine how much the variable influences the AQI. This information can be crucial in formulating effective control measures to mitigate pollution and enhance air quality.
3) **Random Forest regression**: is a supervised learning technique that combines multiple Decision Trees and can be used for regression problems. The input data goes through multiple Decision Trees, and the average of each tree is used as the model's output in the training process [1].

H. EVALUATION MEASURES

In our evaluation stage, we aim to offer a comprehensive analysis of different regression approaches and performance metrics to provide the flexibility to select the classifier whose accuracy specifications are most relevant to users. Therefore, this part implies the most popular error rate metrics used in machine learning and information retrieval domains. We list these measures and explain each one next.

- Mean Absolute Error (MAE): is a metric used to calculate the mean value for the differences between the actual and predicted values observed from the model. It indicates the average of the model errors, as shown in the equation below:

$$MAE = \frac{1}{n}\sum_{j=1}^{n}(y_j - y'_j) \tag{2}$$

- Root Mean Square Error (RMSE): is a widely used for evaluating regression models. It is used to calculate the average deviation between predicting and actual model values. A lower RMSE value highlighted that the model achieved better performance. It can be calculated using the following formula:

$$RMSE = \sqrt{\frac{1}{n}(\sum_{j=1}^{n}(y_j - y'_j)^2)} \tag{3}$$

- R^2 Score: is used to find the variance of target variables in the model. It ranges from 0 to 1, with a higher value representing that the proposed model fits the dataset in a good way. It is calculated using the following formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y'_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{4}$$

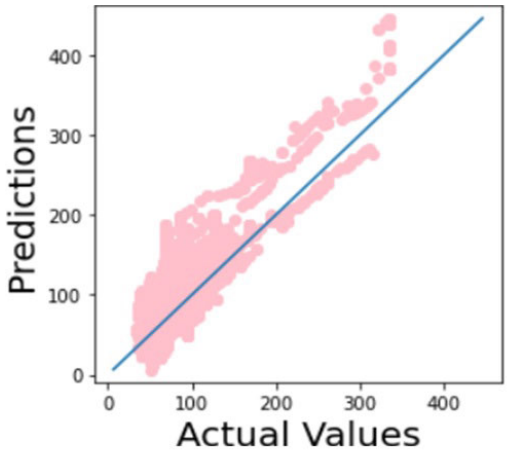


FIGURE 4. Actual vs predicted for linear regression.

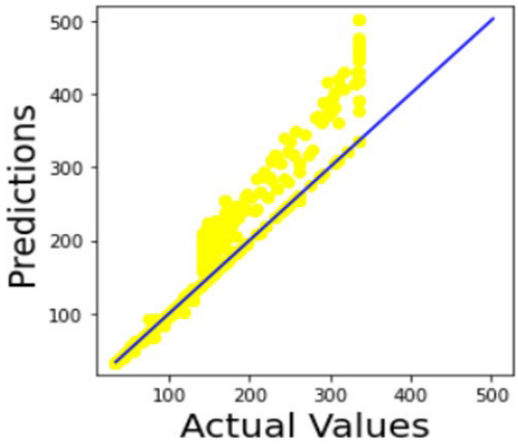


FIGURE 5. Actual vs predicted for decision tree regression.

IV. EXPERIMENTAL RESULTS

To validate reliability and effectiveness of air quality prediction methodologies, we present all experimental findings and compare them from different perspectives. Our initial evaluation focuses on comparing the actual and predicted values of each approach to provide a reliable indicator of approaches accuracy.

In addition, we compare the effectiveness of regression approaches in predicting air quality across two execution configurations including (a personal laptop and a cloud based platforms). Moreover, we emphasized measuring the execution times for each regression technique on both selected platforms. This analysis provides valuable insights into the computational efficiency and speed of the models. In the following sections, we discuss the detailed results, providing a comprehensive understanding of the performance of the regression techniques.

A. COMPARISON OF ACTUAL AND PREDICTED DATA

Our initial set of evaluation findings showcases the performance of our approaches in predicting air quality by comparing the actual values with the prediction values generated by models. By visually comparing these two sets of values, we can quickly assess the degree of proximity between them, offering valuable insights into the accuracy of each model. Figure 4 presents the actual values plotted versus the predicted values, focusing specifically on linear regression results. The blue line represents the ideal regression line, and the model's accuracy depends on the degree of alignment between the data points and this line. Upon examining the linear regression results, it becomes evident that the data points are clustered at the bottom of the graph and are not closely aligned with the regression line. This observation suggests that linear regression may not be the most suitable model for air quality prediction in this study.

Continuing with the evaluation of regression models, Figure 5 compares the actual and the prediction values of the Decision Tree regression model. Analyzing the results, we observe that the data points are more evenly distributed throughout the graph and closer to the regression line than the Linear regression case. This indicates that our study's Decision Tree regression model performs better in predicting air quality.

The improved distribution and proximity of data points to the regression line in the case of Decision Tree regression signify a higher level of accuracy and reliability in predicting air quality compared to linear regression. This suggests that the Decision Tree regression model may provide more precise predictions based on the dataset.

Concluding the evaluation of regression models, Figure 6 illustrates the comparison values of the Random Forest regression model. Upon analysis, we observe that the data points are distributed and closer to the regression line, and this graph looks similar to the Decision Tree model graph. While the Random Forest model may offer advantages in handling complex relationships and reducing overfitting, the Decision Tree model's simplicity and interpretability make it a compelling option for understanding the factors influencing air quality. The Decision Tree model can provide valuable insight into the variables representing the most significant impact on air quality, aiding decision-making processes.

B. PERFORMANCE EVALUATION USING DIFFERENT CONFIGURATIONS

This section represents the second set of evaluation results showing our approach's performance by applying regression models in two configurations: personal laptop and cloud platforms. Assessing the models' performance in different platforms is crucial to ensure the reliability and suitability of models for real-world applications. Additionally, it helps assess the effectiveness of computational resources on the model's performance. The following sub-sections provide the evaluation results for each configuration.

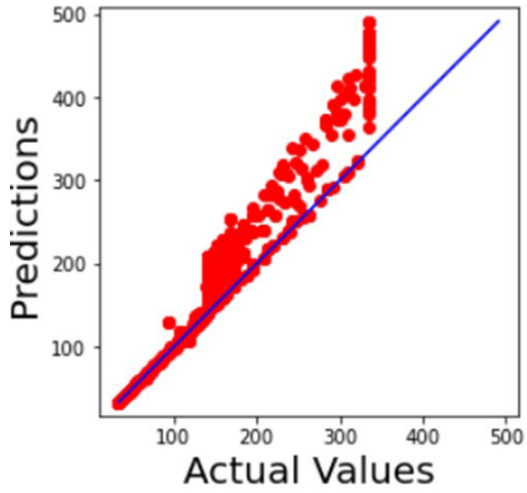


FIGURE 6. Actual vs predicted for random forest regression.

TABLE 2. Evaluation results of training dataset using laptop configuration.

Model	MAE	RMSE	R^2
Random Forest Regression	2.46	10.39	98.80
Decision Tree Regression	2.02	10.14	98.86
Linear Regression	32.19	42.70	79.73

TABLE 3. Evaluation results of testing dataset using laptop configuration.

Model	MAE	RMSE	R^2
Random Forest Regression	0.795	7.141	93.08
Decision Tree Regression	0.738	7.073	93.21
Linear Regression	20.137	25.562	11.36

1) PERFORMANCE EVALUATION IN FIRST CONFIGURATION

In the first configuration, Tables 2 and 3 present the results of error evaluation metrics, specifically RMSE and MAE, for the regression models when executed on a personal laptop platform. The findings show that the Decision Tree model outperforms other models. It achieved 2.02% of MAE and 10.14% of an RMSE, indicating its ability to make predictions with minimal average error and variability. On the other hand, Linear regression achieved low performance, with a relatively high value for MAE of 32.19% and RMSE of 42.70%, concluding that Linear regression may not be the suitable model for accurately predicting air quality.

2) PERFORMANCE EVALUATION IN SECOND CONFIGURATION

In the second configuration, the performance of the regression models was evaluated using a cloud platform. The evaluation outcomes for both the training and testing datasets are detailed in Tables 4 and 5, respectively.

Upon analyzing the evaluation metrics, it observes that the performance of the linear regression model remains relatively stable compared to the first configuration. The MAE and RMSE values are exhibit minimal variation, implying that the change in configuration does not significantly affect the model's performance. On the other hand, there is a slight improvement in the performance of the Decision Tree regression model when running on the cloud platform. The MAE and RMSE values for the training dataset show a marginal decrease, with an MAE of 1.97% and a RMSE of 9.94%. This improvement shows a slightly improved ability to predict air quality compared to the first configuration. Table 5 shows that the Random Forest performance is comparable to the Decision Tree model. Both models represent similar MAE and RMSE values with similar predictive capabilities. However, it is worth noting that the Random Forest model tends to have a longer execution time, which may limit its suitability and efficiency for certain real-world applications where time is crucial.

TABLE 4. Evaluation results of training dataset using cloud configuration.

Model	MAE	RMSE	R^2
Random Forest Regression	2.39	10.19	98.82
Decision Tree Regression	1.97	9.94	98.88
Linear Regression	32.56	43.08	78.89

TABLE 5. Evaluation results of testing dataset using cloud configuration.

Model	MAE	RMSE	R^2
Random Forest Regression	0.82	7.28	93.09
Decision Tree Regression	0.77	7.23	93.19
Linear Regression	20.51	26.16	10.89

C. EXECUTION TIME COMPARISON

This study compared the execution time for three regression models with the SMOTER technique on two different platforms: a personal laptop and a cloud. The goal was to evaluate the impact of cloud computing technology on the efficiency and speed-up of these models. The results presented in Table 6 demonstrate a significant reduction in execution time when the models run using cloud platform compared to the personal laptop. The reduction execution time of regression models highlights the advantages of utilizing cloud computing technology for machine learning tasks. For example, the execution time for SMOTER was reduced from 1292.89 seconds on the personal laptop to 464.22 seconds on the cloud, resulting in a reduction of approximately 64%. Similarly, the execution time of Decision Tree decreased from 0.46 seconds on the personal laptop to 0.28 seconds on the cloud, representing a significant reduction.

Moreover, for the Random Forest model the execution time was reduced from 39.40 seconds using the personal laptop to 17.27 seconds using cloud platform, indicating a reduction of approximately 56%. On the other hand, the execution time for Linear Regression model was already relatively low on

TABLE 6. Model execution times in seconds.

Execution Time	Personal Computer	Cloud
SMOTER	1292.89	464.22
Random Forest Regression	39.40	17.27
Decision Tree Regression	0.46	0.28
Linear Regression	0.07	0.02

the personal laptop, with only 0.07 seconds, and it further decreased to 0.02 seconds on the cloud.

These finding results demonstrated the benefits of utilizing cloud computing technology in reducing the execution time of regression models. Reducing the execution time of models help to achieve more efficient machine learning models. Particularly for larger and more complex datasets, cloud computing frameworks enable of distributing processing data and model training, providing a solution to avoid computational challenges and expedite the machine learning workflow.

V. CONCLUSION

This study provides a comprehensive comparative analysis of different regression models for predicting air quality in smart cities. Notably, the Decision Tree regression model demonstrated a high performance compared to other regression models. Incorporating Exploratory Data Analysis and the SMOTER technique played a pivotal role in enhancing model accuracy by addressing data imbalances and optimizing feature selection. Moreover, the study emphasized the advantages of utilizing cloud computing in regression modeling. Utilizing cloud resources led to reduced model execution time, resulting in enhanced efficiency and scalability. This accelerated experimentation, training, and deployment of the models, enhancing their practical applicability in real-world applications.

For future work recommendations, we explore diverse machine-learning approaches for predicting air quality and air pollution in smart cities. Additionally, investigating the effect of meteorological data, including temperature, pressure, humidity, and wind speed, further enhances AQI and air pollution prediction accuracy. This endeavor provides valuable insight into identifying air quality levels and contributes to more effective air quality management approaches.

REFERENCES

[1] S. Ameer, M. A. Shah, A. Khan, H. Song, C. Maple, S. U. Islam, and M. N. Asghar, "Comparative analysis of machine learning techniques for predicting air quality in smart cities," *IEEE Access*, vol. 7, pp. 128325–128338, 2019.

[2] M. Batty, K. W. Axhausen, F. Giannotti, A. Pozdnoukhov, A. Bazzani, M. Wachowicz, G. Ouzounis, and Y. Portugali, "Smart cities of the future," *Eur. Phys. J. Special Topics*, vol. 214, no. 1, pp. 481–518, Nov. 2012.

[3] I. Bougoudis, K. Demertzis, and L. Iliadis, "HISYCOL a hybrid computational intelligence system for combined machine learning: The case of air pollution modeling in Athens," *Neural Comput. Appl.*, vol. 27, no. 5, pp. 1191–1206, Jul. 2016.

[4] D. Ganeshkumar, "Air and sound pollution monitoring system using cloud computing," *Int. J. Eng. Res.*, vol. V9, no. 6, Jun. 2020.

[5] R. W. Gore and D. S. Deshpande, "An approach for classification of health risks based on air quality levels," in *Proc. 1st Int. Conf. Intell. Syst. Inf. Manage. (ICISIM)*, Oct. 2017, pp. 58–61.

[6] B.-J. He, L. Ding, and D. Prasad, "Enhancing urban ventilation performance through the development of precinct ventilation zones: A case study based on the greater sydney, Australia," *Sustain. Cities Soc.*, vol. 47, May 2019, Art. no. 101472.

[7] G. R. Kingsy, R. Manimegalai, D. M. S. Geetha, S. Rajathi, K. Usha, and B. N. Raabiathul, "Air pollution analysis using enhanced K-means clustering algorithm for real time sensor data," in *Proc. IEEE Region 10 Conf. (TENCON)*, Nov. 2016, pp. 1945–1949.

[8] C. G. Kirwan and F. Zhiyong, *Smart Cities and Artificial Intelligence: Convergent Systems for Planning, Design, and Operations*. Amsterdam, The Netherlands: Elsevier, 2020.

[9] Z. Lv, D. Chen, R. Lou, and Q. Wang, "Intelligent edge computing based on machine learning for smart city," *Future Gener. Comput. Syst.*, vol. 115, pp. 90–99, Feb. 2021.

[10] U. Mahalingam, K. Elangovan, H. Dobhal, C. Valliappa, S. Shrestha, and G. Kedam, "A machine learning model for air quality prediction for smart cities," in *Proc. Int. Conf. Wireless Commun. Signal Process. Netw. (WiSPNET)*, Mar. 2019, pp. 452–457.

[11] S. Mahanta, T. Ramakrishnudu, R. R. Jha, and N. Tailor, "Urban air quality prediction using regression analysis," in *Proc. IEEE Region Conf. (TENCON)*, Oct. 2019, pp. 1118–1123.

[12] H. Maleki, A. Sorooshian, G. Goudarzi, Z. Baboli, Y. Tahmasebi Birgani, and M. Rahmati, "Air pollution prediction by using an artificial neural network model," *Clean Technol. Environ. Policy*, vol. 21, no. 6, pp. 1341–1352, Aug. 2019.

[13] K. Nandini and G. Fathima, "Urban air quality analysis and prediction using machine learning," in *Proc. 1st Int. Conf. Adv. Technol. Intell. Control, Environ., Comput. Commun. Eng. (ICATIECE)*, Mar. 2019, pp. 98–102.

[14] P. J. Navarathna and V. P. Malagi, "Artificial intelligence in smart city analysis," in *Proc. Int. Conf. Smart Syst. Inventive Technol. (ICSSIT)*, Dec. 2018, pp. 44–47.

[15] G. Oliveri Conti, B. Heibati, I. Kloog, M. Fiore, and M. Ferrante, "A review of AirQ models and their applications for forecasting the air pollution health outcomes," *Environ. Sci. Pollut. Res.*, vol. 24, no. 7, pp. 6426–6445, Mar. 2017.

[16] J. W. Park, C. H. Yun, H. S. Jung, and Y. W. Lee, "Visualization of urban air pollution with cloud computing," in *Proc. IEEE World Congr. Services*, Jul. 2011, pp. 578–583.

[17] V. R. Pasupuleti, Uhasri, P. Kalyan, Srikanth, and H. K. Reddy, "Air quality prediction of data log by machine learning," in *Proc. 6th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, Mar. 2020, pp. 1395–1399.

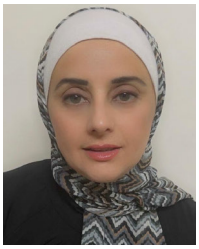
[18] R. M. Patil, D. H. T. Dinde, and S. K. Powar, "A literature review on prediction of air quality index and forecasting ambient air pollutants using machine learning algorithms," *Int. J. Innov. Sci. Res. Technol.*, vol. 5, no. 8, pp. 1148–1152, Sep. 2020.

[19] H. Peng, A. R. Lima, A. Teakles, J. Jin, A. J. Cannon, and W. W. Hsieh, "Evaluating hourly air quality forecasting in Canada with nonlinear updatable machine learning methods," *Air Qual., Atmos. Health*, vol. 10, no. 2, pp. 195–211, Mar. 2017.

[20] R. Sharma, G. Shilimkar, and S. Pisal, "Air quality prediction by machine learning," *Int. J. Sci. Res. Sci. Technol.*, vol. 8, pp. 486–492, 2021.

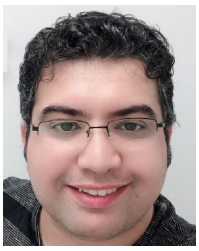
[21] S. Simu, V. Turkar, R. Martires, V. Asolkar, S. Monteiro, V. Fernandes, and V. Salgaoncary, "Air pollution prediction using machine learning," in *Proc. IEEE Bombay Sect. Signature Conf. (IBSSC)*, Dec. 2020, pp. 231–236.

[22] Z. Zhang, H. Chen, and X. Huang, "Prediction of air quality combining wavelet transform, DCCA correlation analysis and LSTM model," *Appl. Sci.*, vol. 13, no. 5, p. 2796, Feb. 2023.



SHOROUQ AL-EIDI received the M.S. degree in computer science from the Jordan University of Science and Technology, Jordan, and the Ph.D. degree in computer science from the Memorial University of Newfoundland, Canada. She is an Assistant Professor with the Computer Science Department, Tafila Technical University. Her research interests include cyber security, machine learning, networks, and big data analysis.

FATHI AMSAAD received the bachelor's degree in computer science from the University of Benghazi, Libya, in 2002, and the dual master's degrees in computer science and computer engineering from the University of Bridgeport, CT, USA, in 2011 and 2012, respectively, and the Ph.D. degree in engineering with an emphasis on computer science and engineering from the University of Toledo, OH, USA, in 2017. He is an Assistant Professor of Computer Science and Engineering at Wright State University, Dayton, OH, USA. He has supervised over ten graduate students, including Niveshitha Niveshitha. He has established the Semiconductor Microelectronics Assurance, Resilience, and Trust (SMART) Cybersecurity Research Lab at 490 Joshi Research Center, Computer Science and Engineering Department, Wright State University. At the SMART Cybersecurity Research Laboratory, He leads a research team comprising several graduate students (master's and Ph.D.) a Post-Doctoral Researcher, and a Research Assistant Professor. His research interests include Assured and Trusted Digital Microelectronics, Secure Heterogeneous Integration and Advanced Packaging, Blockchain-enabled Federated Learning, IoT Hardware Security, Machine/Deep Learning for Cybersecurity, AI Distributed Cloud Computing, Secure AI Hardware Accelerators, and Resilient Circuit Design (Memory/Microprocessor/ASICs/FPGAs). Both government and industry fund his research including AFRL, AFOSR, Intel, NSA, and the Ohio Department of Education. He has participated in several collaborative research proposals that have led to a cumulative sum of about $33 Million, including all partners along with Wright State University. He has served as an Organizer, Program Chair, Technical Program Committee member, Guest Editor, and on the Reviewer Board for several international conferences and journals. In addition to his research activities, he has established teaching experience in hardware security, IoT and embedded systems security, distributed computing, digital systems, and network administration and security curriculum.



OMAR DARWISH (Senior Member, IEEE) received the M.S. degree from the Jordan University of Science and Technology, Jordan, and the Ph.D. degree in computer science from Western Michigan University, USA. He is an Assistant Professor with the Information Security and Applied Computing Department, Game Above College of Engineering and Technology, Eastern Michigan University. He was an Assistant Professor, a Program Coordinator of computer information systems, and the Director of the IoT and Cybersecurity Laboratory, Ferrum College; a Visiting Assistant Professor with the Institute of Technology, West Virginia University; a Software Engineer with MathWorks; and a Programmer with Nuqul Group. His research interests include cyber security, the IoT, machine learning, networks, big data analysis, cloud computing, artificial intelligence, data mining, and information retrieval.

YAHYA TASHTOUSH received the B.Sc. and M.Sc. degrees in electrical engineering from the Jordan University of Science and Technology (JUST), Irbid, Jordan, in 1995 and 1999, respectively, and the Ph.D. degree (joint degree) in computer engineering from The University of Alabama in Huntsville, AL, USA, and the University of Alabama at Birmingham, AL, in 2006. He is a Full Professor with the College of Computer and Information Technology, JUST. His current research interests are the IoT, deep/machine learning, wireless networks, robotics, and fuzzy systems.

ALI ALQAHTANI received the Ph.D. degree in computer engineering from Oakland University, Rochester Hills, MI, USA, in 2020. He is currently an Assistant Professor with Najran University (NU). His research interests include machine learning in general and deep learning in image and signal processing, wireless vehicular networks (VANETs), wireless sensor networks, and cyber-physical systems.

NIVESHITHA NIVESHITHA (Graduate Student Member, IEEE) is a Graduate Student with Wright State University. His research interests include artificial intelligence, machine learning, and cloud computing.