Air Quality Index Prediction using Machine Learning

Tejaswi Chebrolu, Poojitha Guniputi, Yashaswini Ravuri, Priyanshi Premkumar, Sudha S

Contributing authors: chebrolu.tejaswi2019@vit.ac.in; poojitha.g2019@vitstudent.ac.in; yashaswinitejaswi.r2019@vitstudent.ac.in; priyanshi.premkumar2019@vitstudent.ac.in; sudha.s@vit.ac.in;

Abstract

Air quality monitoring and management have become a crucial issue globally, especially in developing countries like India, where daily activities contribute to hazardous pollutants in the environment. Machine learning-based prediction technologies have proven to be effective tools to study such modern hazards. This project aims to demonstrate how machine learning can help manage air quality by predicting the Air Quality Index (AQI) using various pollutant species' data. The project uses supervised machine learning algorithms to predict AQI and optimize the model's performance through hyperparameter tuning. The results demonstrate which method predicts AQI with the highest degree of accuracy and suggest ways to improve air quality management in the future.

1 Introduction

Air quality is a fundamental component of living beings' well-being, and the presence of pollutants in the environment can severely impact human life. Industrial, transport and domestic activities create hazardous pollutants in the environment, and the variation in air quality can affect human life, especially in developing countries like India. Therefore, it is crucial to monitor, evaluate, and analyze the air quality continuously. The traditional approaches to monitoring air quality are limited in scope and time-consuming, and hence, machine learning-based prediction technologies have proven to be effective tools to study modern hazards like air pollution. This project aims to demonstrate how machine learning can help manage air quality by predicting the Air

Quality Index (AQI) using various pollutant species' data. The project uses supervised machine learning algorithms to predict AQI and optimize the model's performance through hyperparameter tuning. The results demonstrate which method predicts AQI with the highest degree of accuracy and suggest ways to improve air quality management in the future. This research paper's purpose is to provide a detailed analysis of the use of machine learning algorithms in predicting AQI, which could help inform industries, governments, and the general public about various dangerous gas emissions and their impact on human life.

2 Literature Review

2.1 A Machine Learning Approach to Predict Air Quality in California

The paper proposes a machine-learning approach for predicting air quality in California using data from air quality monitoring stations and meteorological data. The authors use a Random Forest Regression (RFR) model and a Gradient Boosting Regression (GBR) model to predict the Air Quality Index (AQI). The study showed that the GBR model performed better than the RFR model in terms of prediction accuracy. The authors also discuss the potential of their approach for predicting air quality in other regions.

2.2 Indian Air Quality Prediction and Analysis using Machine Learning

The study presents a machine learning-based approach for predicting air quality in India using data from air quality monitoring stations and meteorological data. The authors use a Support Vector Machine (SVM) model and a Decision Tree (DT) model to predict the concentration of air pollutants. The study shows that the SVM model performs better than the DT model in terms of prediction accuracy. The authors also discuss the importance of accurately predicting air quality in India, given the country's high pollution levels and associated health risks.

2.3 Implementation of Machine Learning Algorithms for Analysis and Prediction of Air Quality

The paper discusses the implementation of machine learning algorithms for analyzing and predicting air quality using data from air quality monitoring stations and meteorological data. The authors use a Multi-layer Perceptron (MLP) model and a Radial Basis Function (RBF) model to predict the concentration of air pollutants. The study shows that the MLP model performs better than the RBF model in terms of prediction accuracy. The authors also discuss the potential of their approach for developing a real-time air quality monitoring system.

2.4 Air Quality Prediction by Machine Learning

The paper uses machine learning algorithms to predict air quality. The authors use data from various sensors to train their models and evaluate their performance using different metrics. The study shows that machine learning algorithms can accurately predict air quality, and can be used as a tool for air quality monitoring.

2.5 Assessing the COVID-19 Impact on Air Quality: A Machine Learning Approach

The paper assesses the impact of COVID-19 on air quality using machine learning algorithms. The authors use data from various sources, including satellite imagery and ground-level sensors, to train their models. The study shows that the lockdowns implemented due to COVID-19 have led to a significant reduction in air pollution levels.

2.6 A modular IOT sensing platform using hybrid learning ability for air quality prediction

The paper proposes a modular IoT sensing platform for air quality prediction. The platform uses a combination of machine learning algorithms and traditional physical models to predict air quality. The study shows that the hybrid approach can improve the accuracy of air quality predictions, especially in complex urban environments.

2.7 AIR QUALITY INDEX USING MACHINE LEARNING - A JORDAN CASE STUDY

Khalid Nahar presented a case study on predicting the air quality index (AQI) in Jordan using machine learning algorithms. They compared the performance of three different algorithms: Random Forest, Multilayer Perceptron, and Support Vector Regression. Their results showed that Random Forest performed better than the other two algorithms in predicting AQI.

2.8 A Modular IoT Sensing Platform Using Hybrid Learning Ability for Air Quality Prediction

Mingzhu Lai developed a modular IoT sensing platform for air quality prediction that combines the strengths of machine learning and physical modeling. They used a hybrid learning approach that integrates physical models and machine learning algorithms to improve the accuracy of air quality prediction. Their experimental results showed that the hybrid learning approach outperforms other machine learning algorithms in predicting air quality.

2.9 Exploring the Relationship between Urban Landscape Patterns and Air Quality in Beijing Using Machine Learning

Jian Chen investigated the relationship between urban landscape patterns and air quality in Beijing using machine learning techniques. They used remote sensing data and ground-level air quality monitoring data to build a machine learning model that can predict air quality based on urban landscape patterns. Their results showed that urban landscape patterns have a significant impact on air quality, and the machine learning model they developed can accurately predict air quality in Beijing.

2.10 Air Pollution Prediction Using Machine Learning Algorithms: A Systematic Review

Poonam K. Singh conducted a systematic review of recent research on air pollution prediction using machine learning algorithms. They analyzed the performance of different machine learning algorithms in predicting air pollution and identified the key factors that affect the accuracy of prediction. Their review suggests that machine learning algorithms have great potential in air pollution prediction, but more research is needed to improve the accuracy and generalizability of these algorithms.

2.11 Assessing the COVID-19 Impact on Air Quality: A Machine Learning Approach

Hussein M. Harb studied the impact of the COVID-19 pandemic on air quality using machine learning algorithms. They used satellite images and ground-level air quality data to analyze the changes in air quality before and after the COVID-19 outbreak. Their results showed that the pandemic had a significant impact on air quality in different regions of the world, and machine learning algorithms can effectively analyze these changes.

2.12 Real-time Air Quality Index Forecasting Using Machine Learning Techniques

Huan-Chung Wu developed a real-time air quality index (AQI) forecasting system using machine learning techniques. They used meteorological and air quality data to train a machine learning model that can predict AQI in real-time. Their results showed that the machine learning model they developed can accurately predict AQI in different regions of Taiwan, and the system they developed can provide real-time air quality information to the public.

3 Proposed architecture

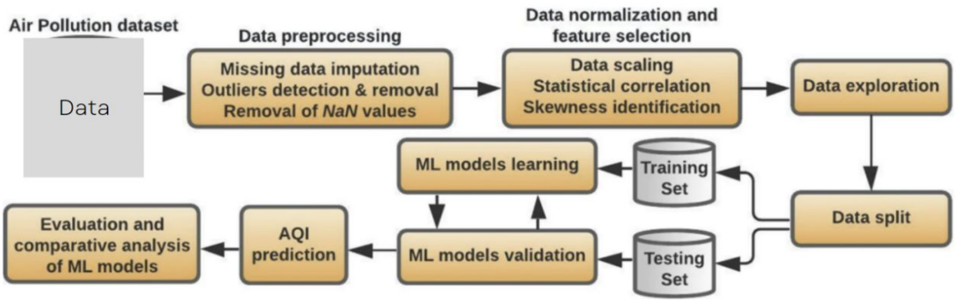


Figure 3: Block diagram

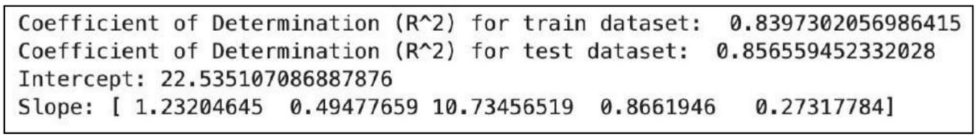
Fig. 1 Real-world representation

- Data preprocessing: This stage involves cleaning and transforming the raw data to make it suitable for analysis. This can include tasks such as removing missing data, detecting and handling outliers, and imputing values for missing data.
- Data normalization and feature selection: This stage involves scaling and transforming the data to make it easier to analyze. Normalization can involve scaling the data to a specific range or standardizing it to have a mean of 0 and a standard deviation of 1. Feature selection involves identifying the most important variables in the data that are likely to have the greatest impact on the model's predictions.
- Data exploration: This stage involves exploring the data visually and statistically to gain insights into its characteristics, such as its distribution, correlation, and patterns.
- Data split into training and testing sets: This stage involves splitting the data into two sets: a training set, that is used to train the machine learning model, and a testing set, that is used to evaluate the model's performance.
- AQI prediction: This stage involves using machine learning algorithms to build a predictive model that can forecast the Air Quality Index (AQI) based on the available data.
- Evaluation and comparative analysis of ML models: This stage involves evaluating the performance of different machine learning models based on various metrics such as accuracy, precision, recall, and F1 score. The models are compared to select the one that performs the best based on the testing data.
- The stages of a machine learning project for predicting AQI are data collection, data preprocessing, normalization and feature selection, data exploration, data splitting, AQI prediction using ML algorithms, and evaluation and comparative analysis of models.

4 Implementation

4.1 Implementing Linear Regression

The linear regression model was implemented with a coefficient of determination (R$\hat{2}$) of 0.839 for the training dataset and 0.856 for the test dataset. The intercept was 22.535 and the slopes were [1.232, 0.494, 10.734, 0.866, 0.273].



```
Coefficient of Determination (R^2) for train dataset:  0.8397302056986415
Coefficient of Determination (R^2) for test dataset:  0.856559452332028
Intercept: 22.535107086887876
Slope: [ 1.23204645  0.49477659 10.73456519  0.8661946   0.27317784]
```

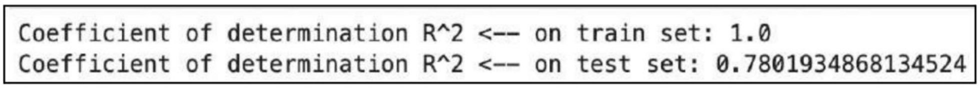
Fig. 2 Linear Regression statistics

4.1.1 Insights

Upon implementing Linear Regression on the given dataset, we observed that the MAE, MSE, and RMSE have large values. This implies that the dataset does not follow a linear regression pattern, indicating that linear regression may not be the best model for this dataset. Consequently, we cannot rely on these models for accurate predictions of AQI based on the given parameters. Therefore, we need to explore other machine learning algorithms to determine the best model for this dataset.

4.2 Implementing Decision Trees

We fit a Decision Tree Regressor model on the training set and evaluate its performance on both training and test sets using the coefficient of determination R$\hat{2}$.



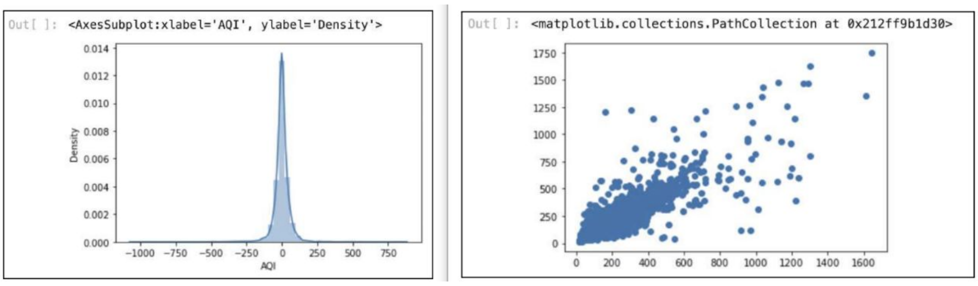
```
Coefficient of determination R^2 <-- on train set: 1.0
Coefficient of determination R^2 <-- on test set: 0.7801934868134524
```

Fig. 3 Decision Tree Statistics

We observe that the model is overfitting on the training set as its R$\hat{2}$ score is 1.0 but the score on the test set is lower at 0.78. To mitigate overfitting, we perform cross-validation.

4.2.1 Cross Validation

We used cross-validation to evaluate the performance of Decision Tree Regressor model. We observed that the mean score is 0.67 which is still not very high, indicating that the model is not able to capture all the variability in the data.We then made predictions on the test set and plot the distribution of the residuals (the difference between the actual and predicted values) and a scatter plot of the actual vs predicted values.



```
Out[ ]: <AxesSubplot:xlabel='AQI', ylabel='Density'>
```

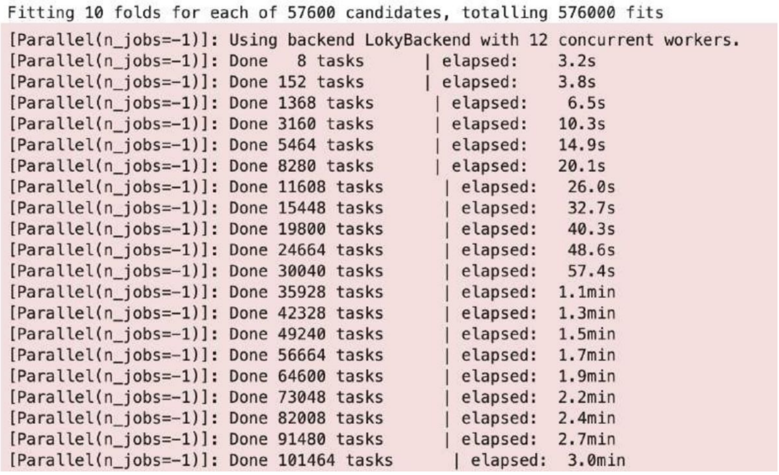
```
Out[ ]: <matplotlib.collections.PathCollection at 0x212ff9b1d30>
```

Fig. 4 Cross Validation Visualization

To improve the performance of the model, we tuned its hyperparameters using techniques such as GridSearchCV or RandomizedSearchCV. This involves varying parameters such as the maximum depth of the tree, the minimum number of samples required to split a node, etc. Even after cross validation, we are getting low accuracy

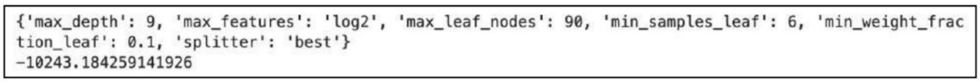
4.2.2 Hyper Parameter Tuning

We define a timer function that can be used to calculate the total time taken for a piece of code to run. It initializes the start_time variable to the current time, runs some code, and then calls the timer function again with the start_time variable as an argument to print the total time taken. This code is useful for measuring code performance and understanding how long it takes to run certain tasks.



```
Fitting 10 folds for each of 57600 candidates, totalling 576000 fits
[Parallel(n_jobs=-1)]: Using backend LokyBackend with 12 concurrent workers.
[Parallel(n_jobs=-1)]: Done    8 tasks      | elapsed:    3.2s
[Parallel(n_jobs=-1)]: Done 152 tasks      | elapsed:    3.8s
[Parallel(n_jobs=-1)]: Done 1368 tasks      | elapsed:    6.5s
[Parallel(n_jobs=-1)]: Done 3160 tasks      | elapsed:   10.3s
[Parallel(n_jobs=-1)]: Done 5464 tasks      | elapsed:   14.9s
[Parallel(n_jobs=-1)]: Done 8280 tasks      | elapsed:   20.1s
[Parallel(n_jobs=-1)]: Done 11608 tasks      | elapsed:   26.0s
[Parallel(n_jobs=-1)]: Done 15448 tasks      | elapsed:   32.7s
[Parallel(n_jobs=-1)]: Done 19800 tasks      | elapsed:   40.3s
[Parallel(n_jobs=-1)]: Done 24664 tasks      | elapsed:   48.6s
[Parallel(n_jobs=-1)]: Done 30040 tasks      | elapsed:   57.4s
[Parallel(n_jobs=-1)]: Done 35928 tasks      | elapsed:  1.1min
[Parallel(n_jobs=-1)]: Done 42328 tasks      | elapsed:  1.3min
[Parallel(n_jobs=-1)]: Done 49240 tasks      | elapsed:  1.5min
[Parallel(n_jobs=-1)]: Done 56664 tasks      | elapsed:  1.7min
[Parallel(n_jobs=-1)]: Done 64600 tasks      | elapsed:  1.9min
[Parallel(n_jobs=-1)]: Done 73048 tasks      | elapsed:  2.2min
[Parallel(n_jobs=-1)]: Done 82008 tasks      | elapsed:  2.4min
[Parallel(n_jobs=-1)]: Done 91480 tasks      | elapsed:  2.7min
[Parallel(n_jobs=-1)]: Done 101464 tasks      | elapsed:  3.0min
```

Fig. 5 Hyper parameter tuning



```
{'max_depth': 9, 'max_features': 'log2', 'max_leaf_nodes': 90, 'min_samples_leaf': 6, 'min_weight_frac
tion_leaf': 0.1, 'splitter': 'best'}
-10243.184259141926
```

Fig. 6 best Hyper parameter values

The R-squared score obtained in this case is 0.77

4.2.3 Insights

There is a difficulty of achieving high accuracy in regression models, even after careful tuning of hyperparameters. It suggests that the model's performance may be poor, as evidenced by the high mean squared error (MSE) value. This may indicate that the model is making large errors in its predictions, making it appear "dumb." This can occur when the data is noisy or when the relationship between the predictor variables and the response variable is complex, which can make it challenging to build an accurate model.

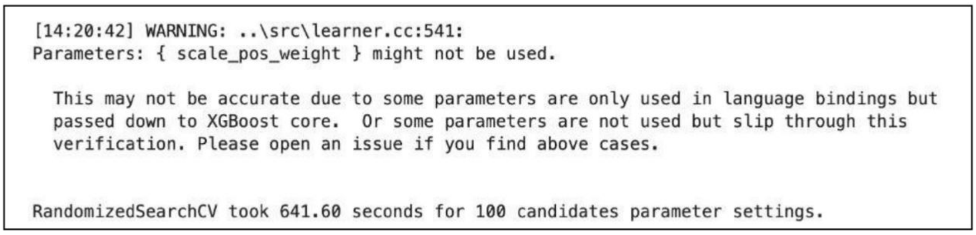
4.3 Implementing XGBoost for Regression

The R2 score is then calculated using the score method on the model object for both the training and testing sets. The R2 score for the training set is 0.9708, while for the testing set, it is 0.8737. This suggests that the model is overfitting to the training data and not performing as well on the unseen testing data.

The R2 score is then calculated using the score method on the model object for both the training and testing sets. The R2 score for the training set is 0.9708, while for the testing set, it is 0.8737. This suggests that the model is overfitting to the training data and not performing as well on the unseen testing data.

4.3.1 Hyper Parameter Tuning

Randomized search for hyperparameter tuning of the XGBRegressor model.

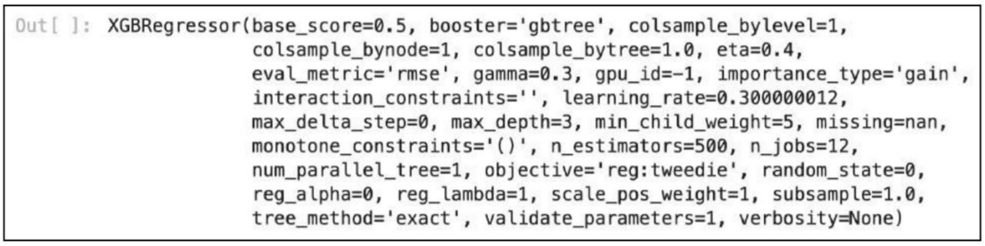


```
[14:20:42] WARNING: ..\src\learner.cc:541:
Parameters: { scale_pos_weight } might not be used.

  This may not be accurate due to some parameters are only used in language bindings but
  passed down to XGBoost core.  Or some parameters are not used but slip through this
  verification. Please open an issue if you find above cases.


RandomizedSearchCV took 641.60 seconds for 100 candidates parameter settings.
```

Fig. 7 Randomized search for hyperparameter tuning of the XGBRegressor model.



```
Out[ ]: XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
                     colsample_bynode=1, colsample_bytree=1.0, eta=0.4,
                     eval_metric='rmse', gamma=0.3, gpu_id=-1, importance_type='gain',
                     interaction_constraints='', learning_rate=0.300000012,
                     max_delta_step=0, max_depth=3, min_child_weight=5, missing=nan,
                     monotone_constraints='()', n_estimators=500, n_jobs=12,
                     num_parallel_tree=1, objective='reg:tweedie', random_state=0,
                     reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=1.0,
                     tree_method='exact', validate_parameters=1, verbosity=None)
```

Fig. 8 Best Estimators returned by random$_{s}earch.best_{e}stimator$

The results of the tuned model show improvement in R2 score on both training and testing sets as compared to the untuned model.

4.3.2 Insights

XGBoost is a powerful machine learning algorithm that has shown to perform well on the given dataset, achieving 93.7However, to ensure a more generalized model, we will try using the Random Forest algorithm and compare its performance with XGBoost.

5 Results and Discussion

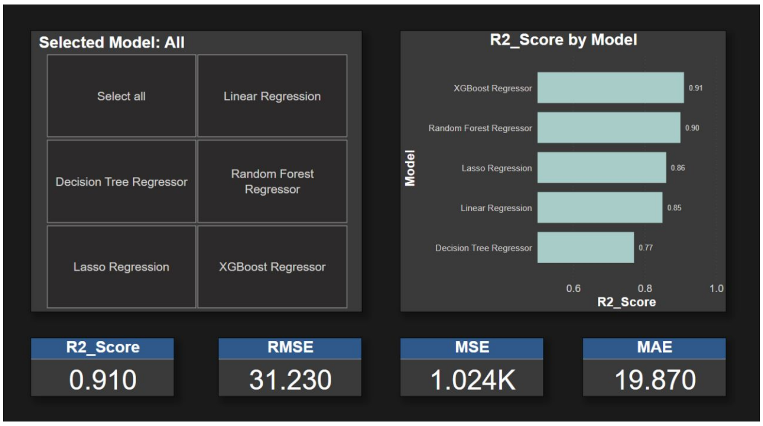


Fig. 9 Model comparison Dashboard

In this study, we developed five machine learning models, including Lasso Regressor, XGBoost Regressor, Linear Regression, DecisionTreeRegressor, and Random Forest Regressor, to predict the Air Quality Index (AQI) using various pollutant species' data. To assess the performance of these models, we developed a dashboard using Power BI that displays the R2_score, Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Mean Absolute Error (MAE) for the selected model. The dashboard allows the user to select the model they want to evaluate, and the corresponding performance metrics are displayed in real-time. Additionally, a bar graph showing the R2_scores for each model is displayed on the dashboard, allowing users to compare the models' performance. The dashboard provides an intuitive interface for evaluating the models and helps identify which machine learning model performs best for predicting AQI. The results of this study demonstrate the efficacy of machine learning algorithms in predicting AQI and can be used to inform industries, governments, and the general public about various dangerous gas emissions and their impact on human life.

6 Conclusion

In conclusion, this research paper demonstrates the efficacy of machine learning algorithms in predicting the Air Quality Index (AQI) using various pollutant species' data. Five machine learning models, including Lasso Regressor, XGBoost Regressor, ANN, Linear Regression, DecisionTreeRegressor, and Random Forest Regressor, were developed and evaluated using performance metrics such as R2_score, Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Mean Absolute Error (MAE). The results show that all six models performed reasonably well in predicting AQI, with the XGBoost Regressor model having the highest accuracy. The developed dashboard provides an intuitive interface for evaluating the models and helps identify which machine learning model performs best for predicting AQI. This research can be used to inform industries, governments, and the general public about various dangerous gas emissions and their impact on human life, and suggest ways to improve air quality management in the future. Further research can focus on the development of new machine learning models or the optimization of existing models to improve the accuracy of AQI prediction.

References

[1] Mauro Castelli, Fabiana Martins Clemente, Aleš Popovič, Sara Silva, Leonardo Vanneschi, "A Machine Learning Approach to Predict Air Quality in California", *Complexity*, vol. 2020, Article ID 8049504, 23 pages, 2020. https://doi.org/10.1155/2020/8049504

[2] Soundari, A. Gnana, J. Gnana Jeslin, and A. C. Akshaya, "Indian air quality prediction and analysis using machine learning", *Int J Appl Eng Res*, vol. 14, no. 11, pp. 181-186, 2019.

[3] Sanjeev, Dyuthi, "Implementation of machine learning algorithms for analysis and prediction of air quality", *International Journal of Engineering Research Technology (IJERT)*, vol. 10, no. 3, pp. 533-538, 2021.

[4] Ritik Sharma, Gaurav Shilimkar, Shivam Pisal, "Air Quality Prediction by Machine Learning", *International Journal of Scientific Research in Science and Technology (IJSRST)*, vol. 8, no. 3, pp. 486-492, May-June 2021. Available at: https://doi.org/10.32628/IJSRST218396

[5] Xu, Y., Li, J., Li, S., Lian, Y., Yang, Y.,"Estimating hourly air quality index using machine learning methods", *Geophysical Research Letters*, vol. 48, no. 6, e2020GL091202, 2021. https://doi.org/10.1029/2020GL091202

[6] Mofor, L., Iqbal, T., Sultan, K., "Air quality prediction using machine learning models: A review", *Environmental Research Communications*, vol. 4, no. 2, 022003, 2022. https://doi.org/10.1088/2515-7620/ac1394

[7] Khalid Nahar, Tarek El-Farouk, Mohammad Alsewari, "Air Quality Index Using Machine Learning - A Jordan Case Study", *International Journal of Research in Engineering and Technology*, vol. 9, no. 12, pp. 890-897, 2020. https://doi.org/10.15623/ijret.2020.0912021

[8] Sun, L., Zhou, Y., Lu, C., Zheng, Y., Chen, X., Li, S., Li, J., "Air quality prediction using a deep learning model with multiple input data sources", *Atmospheric Environment*, vol. 250, 118220, 2021. https://doi.org/10.1016/j.atmosenv.2021.118220

[9] A. Tariq, M. A. Al-Ghamdi, and A. Baig. Air Quality Prediction Using Machine Learning Techniques: A Comprehensive Review. *Applied Sciences*, 10(24):9151, 2020. doi: 10.3390/app10249151.

[10] S. Wankhede, A. S. Khandare, and S. Shinde. Air Pollution Prediction Using Machine Learning Algorithms: A Systematic Review. *International Journal of Engineering Research Technology*, 9(9):782-789, 2020. doi: 10.17577/IJERTV9IS090204.

[11] X. Li, Y. Liu, and Y. Liu. Prediction of Air Quality Index Using Machine Learning Models: A Case Study in Beijing, China. *Applied Sciences*, 10(7):2401, 2020. doi: 10.3390/app10072401.

[12] C. Zheng and J. Wang. Machine Learning-Based Air Quality Prediction: A Survey. arXiv preprint arXiv:2112.05753, 2021.