**Journal of Applied Informatics and Computing (JAIC)**
Vol.8, No.2, December 2024, pp. 384~389
e-ISSN: 2548-6861 

# Prediction of Air Quality Index Using Ensemble Models

**Theresia Herlina Rochadiani [1]***
* Informatics, Pradita University
theresia.herlina@pradita.ac.id [1]

**Article Info**

***Article history:***

Received …
Revised …
Accepted …

***Keyword:***

*Air Quality Index,*
*Ensemble models,*
*Prediction,*
*Regression.*

**ABSTRACT**

The impact of air pollution on health is measured by the Air Quality Index (AQI). Accurate AQI prediction is essential for pollution reduction and public health recommendations. Traditional methods of monitoring air quality are inaccurate and time-consuming. This study uses IoT-based air quality data from Kampung Kalipaten, Tangerang to build an AQI prediction model with machine learning, specifically an ensemble model. Ensemble techniques such as bagging and boosting, which increase the reliability of predictions by reducing model bias and inconsistency, improve AQI prediction. Four ensemble models used in this study, they are Random Forest Regressor, Gradient Boosting Regressor, Adaboosting Regressor, and Bagging Regressor. As the evaluation, RMSE, MAE, and $R^2$ metrics used. Gradient Boosting Regressor perform the best prediction with RMSE value of 0.6087, MAE value of 0.4659, and $R^2$ value of 0.6442, although no significant differences of RMSE, MAE, and $R^2$ value of the rest models.



## I. INTRODUCTION

To address environmental issues, make appropriate policies, and reduce health risks, identification of severe air pollution areas, implementation of real-time monitoring, and investigation of effective prediction methods have become important steps [1]. Due to its significant impact on public health and the environment, air quality is a major concern in cities and industries around the world. The Air Quality Index (AQI) is a common tool that indicates current and projected levels of air pollution.

The AQI is an index that measures the impact of air pollution on a person's health over a short period of time [2]. Accurate AQI predictions are essential for providing public health advice and taking action to reduce pollution. There are six categories of AQI as shown in Figure 1. The air quality in good category if an AQI score between 0 and 50, meaning that there is little or no health risk. AQI in moderate category if AQI ranges from 51-100. Air quality is acceptable, but a small number of people may experience moderate health problems due to ozone or particle pollution. While the AQI value is between 101 and 150, it is in unhealthy for sensitive groups category, which sensitive groups may experience health impacts. However, the general public may not be affected. When the AQI value is between 151 and 200, unhealthy category, everyone may experience health impacts. Members of sensitive groups may experience more severe health impacts. An AQI in very unhealthy category if its value between 201 and 300. It triggers a health alert, meaning that people may experience more serious health effects. Hazardous category, the last category, if AQI value above 300. It indicates an emergency health warning. Everyone is at greater risk of serious health impacts.

| Air Quality Index (AQI) Values | Levels of Health Concern | Colors |
|---|---|---|
| *When the AQI is in this range:* | *...air quality conditions are:* | *...as symbolized by this color:* |
| 0 to 50 | Good | Green |
| 51 to 100 | Moderate | Yellow |
| 101 to 150 | Unhealthy for Sensitive Groups | Orange |
| 151 to 200 | Unhealthy | Red |
| 201 to 300 | Very Unhealthy | Purple |
| 301 to 500 | Hazardous | Maroon |

Figure 1. Categories of AQI [3]

As the settlement, Kampung Kalipaten in West Pakulonan Village concern about the air quality because it affects the quality of life of its residents. Using the IoT based-air quality monitoring system using LoRaWAN, it monitors the air quality in its open spaces that usually used for social activities [4].

Traditional methods for monitoring air quality typically involve manual data collection and analysis, which can be time-consuming and inaccurate, while it is challenging for statistical models to handle data time series with non-linear properties [5]. Machine learning improves the prediction and monitoring of air quality indices. Ensemble models, which combine multiple learning algorithms to improve predictive performance [6]–[8] have emerged as a powerful alternative. The accuracy and robustness of these models are improved through the use of multiple algorithms. Like study as did by Liang et al. [9], several prediction models have been built using data collected over eleven years by the Taiwan Environmental Protection Administration (EPA). For air quality index (AQI) level prediction, machine learning methods such as adaptive boosting (AdaBoost), artificial neural networks (ANNs), random forests, stacked ensembles, and support vector machines (SVMs) have yielded promising results. A series of experiments, using datasets for three different regions, found that stacked ensembles performed notably better than random forests and artificial neural networks (ANNs). Chenchen Li et al. [10] used Linear Regression, Lasso Regression, Ridge Regression, Decision Tree Regression, K Nearest Neighbour Regression, Multi-Layer Perceptron Regression, Random Forest Regression, Gradient Boosting Regression, AdaBoost Regression to predict air quality of Henan Province. The best prediction resulted by Random Forest Regression and Gradient Boosting Regression algorithm.

This study looks at the use of ensemble methods to predict AQI based on air quality data captured by IoT-based Air Quality Monitoring in Kampung Kalipaten from since September until October 2022. We use ensemble techniques such as bagging and boosting to combine predictions from baseline models. We use this approach to improve the reliability of AQI forecasts by reducing model bias and inconsistency.

## II. METHODOLOGY

In this study, there are four ensemble models applied for AQI prediction, i.e. Random Forest Regressor, Gradient Boosting Regressor, Ada Boost Regressor, and Bagging Regressor.

### A. Data Collection

Dataset used in this study is collected from the IoT-based Air Quality Monitoring in Kampung Kalipaten, Tangerang. The IoT system measured the air quality index by using RAK1906 WisBlock Environmental Sensor Module that consists of the Bosch® BME680 module-based sensors, i.e., humidity, gas, temperature, and pressure sensors [4]. The data gathered from September until October 2022. The raw data consist of 7 columns that contain information about the condition of air, the datetime of data captured, temperature, pressure, humidity, iaq, and device_id. The total data is 6803 rows.

### B. Data Preprocessing

For enhancing data reliability, it needs do preprocessing. In this study, the missing values, outliers, skewness are checked and managed.

1) *Missing values :* The number of missing values in this dataset is displayed in Figure 2. Temperature, pressure, humidity, and iaq are numerical data. To handle missing values in these attributes, the mean of its value imputed.

```
data.isnull().sum()

Condition        0
Created At     495
temperature    515
pressure       511
humidity       548
iaq            576
device_id      495
dtype: int64
```

Figure 2. The number of missing values

2) *Outliers :* Outliers are deviating from the rest of dataset, and they may make the model inaccurate. Using boxplot, the outliers are checked. As shown in Figure 4, no outliers found in the dataset. So no need specific handling this dataset from outliers.

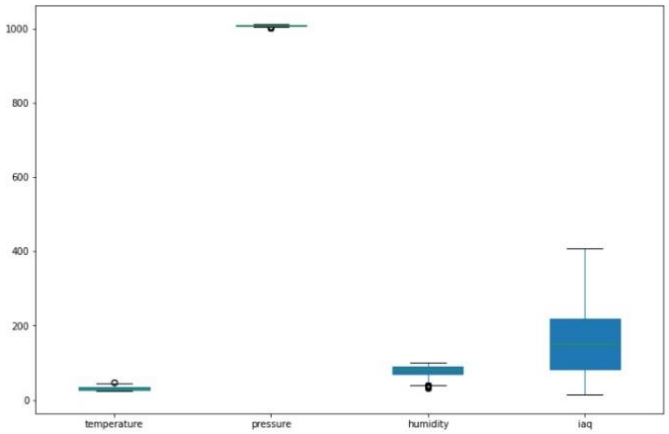


Figure 3. Boxplot of attributes in dataset

3) *Skewness :* The degree to which a random variable's probability distribution deviates from the normal distribution is measured by its skewness. Right-skewed probability distributions have their tails on the right side, while left-skewed probability distributions have their tails on the left.

Using histogram, the skewness of dataset is illustrated. As shown in Figure 4, after exponential transformation applied,

the left-skewed skewness of humidity distribution becomes symmetric.

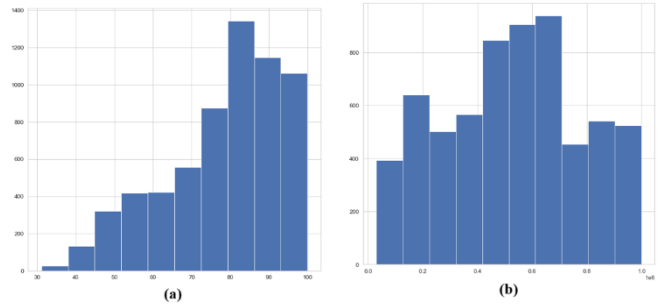


Figure 4. Distribution of humidity (a) Before transformation (b) After transformation.

### C. Feature Selection

For identifying the crucial features, the correlation matrix used. As illustrated in Figure 5, the humidity has high positive correlation and the temperature has high negative correlation with the target attribute, iaq. As a good approach, only attributes have strong correlation with the target are selected as features. So, only humidity and temperature used as features to build the model.

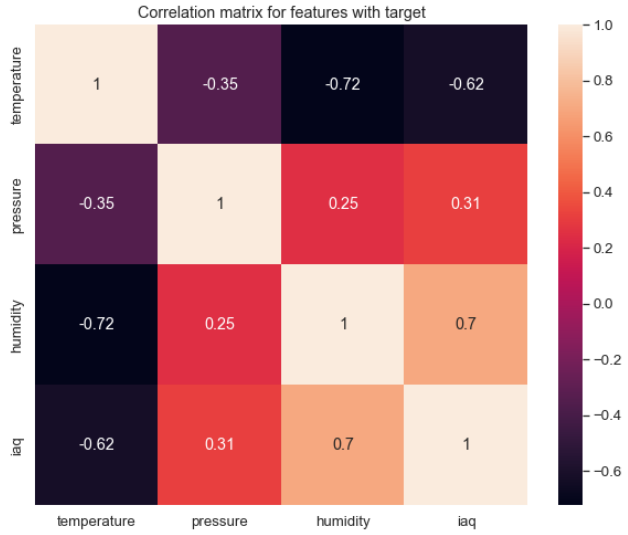


Figure 5. Correlation matrix for features with target.

### D. Feature Engineering

Features measured on different scales can disproportionately influence the model. Standardization ensures that all features contribute equally by transforming them to a common scale. By applying the feature standardization, the values of each feature in the dataset have zero-mean and unit variance.

### E. Modelling

As mentioned before, there are four ensemble models used in this study.

1) *Random Forest:* A nonlinear model that integrates multiple decision trees into a forest is known as a random forest algorithm. Majority selection and random sampling are two key concepts to consider in understanding random forests. For each decision tree, in particular, a training set is randomly selected from the entire sample set [11].
2) *Gradient Boosting Regressor :* Boosting machine learning algorithms that combine weak learners or decision trees to create a strong learner. In addition, to improve performance and reduce error rates, regression models continuously add new decision trees to the previous model each iteration [12].
3) *Adaboost Regressor:* The Adaboost algorithm focuses on creating weak learners and then combining them to create a strong learner. The whole process is done by determining the number of estimators, learning rate, and stumps, assigning weights to them, and at each iteration, the most incorrectly predicted data is selected and fed into the new model [13].
4) *Bagging Regressor:* Bagging is periodic sampling method, with replacement according to a uniform probability distribution, is used to fit the model to the collected data. To produce more accurate predictions, this method uses the results of majority voting or aggregate predictions. This method improves the performance of high variance models when applied to models with overfitting problems. Bagging can reduce the uncertainty of model predictions but still maintain accuracy [14].

### F. Evaluation

Based on review study done by Isakndaryan et al., Root mean squared error (RMSE) and Mean Absolute Error (MAE) are the most metrics used in machine learning-based prediction of air quality [9]. In this study, RMSE, MAE, and $R^2$ are used as metrics for evaluating the performance of the prediction model. RMSE is the square root of average of the squared differences between predicted and actual values [15]. Mathematically, it is expressed as shown in Equation (1).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - A_i)^2} \quad (1)$$

The mean absolute error (MAE) between true and actual values is quantified. The lowest possible value is required for this [16]. Equation (2) shows the formula of MAE.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|P_i - A_i| \quad (2)$$

$R^2$ or coefficient of determination is a statistical measure used in regression analysis to assess how well a model explains the variability of the dependent variable. Its formula is shown in Equation (3).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(P_i - A_i)^2}{\sum_{i=1}^{n}A_i^2} \quad (3)$$

where n= number of data, $P_i$= Predicted value, and $A_i$=Actual value

## III. Result and Discussion

For building the model and evaluating the performance, data has been split, 80% for training and 20% for testing. To reduce overfitting, cross validation is used and to improve the performance of models, GridSearchCV is applied to find the best parameters for each model. As shown in Table I, each model has own hyperparameters. By using GridSearchCV the best combination of hyperparameters is found.

### A. Random Forest Regressor

To optimize the performance of the model, the hyperparameter tuning is needed. In this study, the hyperparameters of Random Forest Regressor are:

1) *Max_depth:* Limiting the depth of tree can prevent overfitting. If set to None, nodes are expanded until all leaves are pure or contain fewer than the minimum samples required to split. Here the best value of this parameter is 10.
2) *Max_features:* For looking the best split, the number of features is needed to be considered. The hyperparameter tuning found that square root of the number of feature (sqrt) is the best value.
3) *Min_samples_leaf:* The minimum number of samples required to be at a leaf node, to prevent creating nodes with a few samples. The optimal value for this parameter is 2.
4) *Min_samples_split*: The minimum number of samples required to split an internal node. Ten is the right value for this parameter.
5) *N_estimators:* More trees can improve performance but also increase computation time. N_estimators is a parameter to set the number of trees in the forest. The best value for this parameter is 100.

### B. Gradient Boosting Regressor

Like hyperparameters of Random Forest Regressor, Gradient Boosting Regressor has hyperparameters max_depth, min_samples_leaf, min_samples_split, and n_estimators with values 4, 4, 10, 100, respectively. Parameter learning_rate is also tuned to control the contribution of each tree to the final model. Here, the best value for learning rate of this model is 0.05.

### C. Adaboost Regressor

In Adaboost Regressor, parameter n_estimators denote the number of weak learners to use in boosting process. The optimal value of this parameter is 100. While the base_estimator_max_depth represent the maximum depth of individual tree from which the boosted ensemble built. Its best value is 5. Another parameter is loss. It specifies the loss function to use when updating the weights after each boosting iteration. The linear function is the right function for this model. Last, same with Gradient Boosting Regression, it has learning_rate parameter that control the contribution of each weak learner. The optimal value found using GridSearchCV for this parameter is 0.01.

### D. Bagging Regressor

The hyperparameters are fine-tuned to improve this model consist of :

1) *Base_estimator_max_depth:* The base estimator to fit on random subsets of the dataset. The result of tuning for this parameter is 5.
2) *Bootstrap:* Whether the samples are drawn with replacement. This value is set to True.
3) *Bootstrap_features:*Whether the features are drawn with the replacement. This value is set to False.
4) *Max_features:* The number of features to draw from the dataset to train each base estimator. The best value for this parameter is 1.
5) *Max_samples:* The number of samples to draw from the dataset to train each base estimator. The best value for this parameter is 0.5. It means a half of dataset.
6) *N-estimators:*The number of base estimators in the ensemble. The result from GridSearchCV this parameter's value is set to 500.

In this section, we will discuss the performance of four ensemble models used to predict the air quality index (AQI). These models are Random Forest Regressor, Gradient Boosting Regressor, AdaBoost Regressor, and Bagging Regressor. Each model is evaluated based on three main metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and coefficient of determination ($R^2$), as shown by Table I.

Random Forest Regressor is an ensemble model that uses the bagging method to combine predictions from multiple decision trees. The max_depth parameter set to 10 limits the maximum depth of the tree, which helps reduce overfitting. Using max_features with a value of 'sqrt' means that each tree only considers the square root of the total available features, which also helps in reducing correlation between trees and improving model generalization. The RMSE value of 0.6054 indicates that this model has a relatively low prediction error. The MAE of 0.4866 indicates that the mean absolute error of the model's predictions is also low. The coefficient of determination ($R^2$) of 0.6266 indicates that this model is able to explain about 62.66% of the variability in the target data. Overall, Random Forest Regressor shows good performance with a balance between accuracy and generalization.

Gradient Boosting Regressor is an ensemble model that builds decision trees incrementally, where each new tree is built to correct the errors of the previous tree. The learning_rate parameter of 0.05 controls the contribution of each new tree, which helps in reducing overfitting. The maximum tree depth (max_depth) is set to 4, which is shallower compared to Random Forest, but sufficient to

capture complex patterns in the data. The RMSE value of 0.6087 is slightly higher than Random Forest, but the MAE of 0.4659 is lower, indicating that this model is better at predicting values closer to the true values. The coefficient of determination ($R^2$) of 0.6442 is the highest among all models, indicating that Gradient Boosting Regressor is able to explain about 64.42% of the variability in the target data. This shows that this model is very effective in capturing complex and non-linear patterns in the data.

AdaBoost Regressor is an ensemble model that uses the boosting method to combine predictions from several base estimators, in this case a decision tree with a maximum depth of 5. The learning_rate parameter of 0.01 controls the contribution of each new estimator, which helps in reducing overfitting. The use of a loss with a value of 'linear' indicates that this model uses a linear loss function to measure the error. The RMSE value of 0.6189 indicates that this model has a slightly higher prediction error compared to Random Forest and Gradient Boosting. The MAE of 0.4753 indicates that the average absolute error of the prediction of this model is also slightly higher compared to Gradient Boosting. The coefficient of determination ($R^2$) of 0.6393 indicates that this model is able to explain about 63.93% of the variability in the target data. Although its performance is slightly below Gradient Boosting, AdaBoost Regressor still shows competitive performance.

Bagging Regressor is an ensemble model that uses the bagging method to combine predictions from several base estimators, in this case decision trees with a maximum depth of 5. The bootstrap parameter is set to True, which means each estimator is trained on a subset of the data taken with replacement. Using max_features with a value of 1.0 means that each estimator considers all available features, while max_samples with a value of 0.5 means that each estimator is trained on half of the total samples. The RMSE value of 0.6142 indicates that this model has a slightly higher prediction error compared to Random Forest and Gradient Boosting. The MAE of 0.4701 indicates that the average absolute error of the prediction of this model is also slightly higher compared to Gradient Boosting. The coefficient of determination ($R^2$) of 0.6443 is the highest along with Gradient Boosting, indicating that Bagging Regressor is able to explain about 64.43% of the variability in the target data. This shows that this model is very effective in capturing complex and non-linear patterns in the data.

From the analysis of the results in Table I, it can be concluded that Gradient Boosting Regressor and Bagging Regressor show the best performance in terms of $R^2$, indicating good ability to capture data variability. Both models can explain more than 64% of the variability in the target data, which is a strong indicator of their predictive ability. Meanwhile, when viewed from the MAE value, the lowest MAE value is owned by the Gradient Boosting Regressor model, indicating that this model is better at predicting values that are closer to the actual value. This shows that Gradient Boosting Regressor is very effective in capturing complex and non-linear patterns in data.

TABLE I
PERFORMANCE OF MODELS

| Models | Best parameters | Metrics evaluation | | |
|---|---|---|---|---|
| | | RMSE | MAE | $R^2$ |
| Random Forest Regressor | 'max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 10, 'n_estimators': 100 | 0.6054 | 0.4866 | 0.6266 |
| Gradient Boosting Regressor | 'learning_rate': 0.05, 'max_depth': 4, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 100 | **0.6087** | **0.4659** | **0.6442** |
| Adaboost Regressor | 'base_estimator__max_depth': 5, 'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 100 | 0.6189 | 0.4753 | 0.6393 |
| Bagging Regressor | 'base_estimator__max_depth': 5, 'bootstrap': True, 'bootstrap_features': False, 'max_features': 1.0, 'max_samples': 0.5, 'n_estimators': 300 | 0.6142 | 0.4701 | 0.6443 |

Meanwhile, the Random Forest Regressor model shows good overall performance with low RMSE and MAE and a fairly high $R^2$. This model is quite stable and robust against overfitting because it uses many decision trees.

The AdaBoost Regressor model also shows competitive performance, although slightly less accurate than other models. This model is still able to explain about 63.93% of the variability in the target data, which shows that this model is quite effective in capturing patterns in the data.

As we can see in the Table I, there is no significant difference in performance between models. This is because the parameters of each model have been well optimized, namely through parameter tuning such as n_estimators, max_depth, and learning_rate. In addition, the similarity in performance between models can also be caused by model evaluation carried out using consistent cross-validation

techniques, so that the results show similar performance because each model is tested on the same subset of data. With parameter tuning and cross validation, in addition to producing similar performance between models, these two things also reduce overfitting in ensemble models.

The Gradient Boosting Regressor model is the best prediction model compared to the other three models. This is because the Gradient Boosting Regressor is very effective in handling data that has complex non-linear relationships. Each tree in the ensemble is built to correct the errors of the previous tree, so the model can capture more complex patterns. In addition, Gradient Boosting iteratively reduces bias by adding new trees that focus on the errors of the previous tree. This makes it very good at capturing details and patterns that other models might miss. Another thing that makes the Gradient Boosting model the best is that the Gradient Boosting Regressor builds the model gradually and each step aims to correct the errors of the previous step, this model can adapt better to the training data, especially if the data has high variation.

However, it is important to note that the performance of the model is highly dependent on the characteristics of the dataset used. In some cases, other models such as Random Forest or AdaBoost may be superior if the data has different characteristics.

## IV. CONCLUSION

This study concludes that employing IoT-based air quality data, ensemble models, particularly the Gradient Boosting Regressor, can accurately predict AQI. Reducing bias and inconsistency in the model through the use of ensemble approaches like bagging and boosting results in predictions that are more trustworthy. The results imply that machine learning models, especially ensemble models, can greatly enhance the prediction of air quality. The limitations of this study are the dataset used is gathered not measure the air pollutant such as PM2.5, PM10, carbon monoxide, etc., and time duration to gather data is not more than two months, which can not represent all conditions of weather. So, for the next study, this prediction model can be improved by using dataset that collected in a year and consists of data of air pollutants. This prediction model of air quality index is essential for environmental and public health policy-making, such as building green area in the pollutant area, giving early warning for residences to wear mask when doing outdoor activities, making policies for using public transportation to minimize the air pollutant, etc.

## REFERENCES


[1] T. Madan, S. Sagar, and D. Virmani, "Air Quality Prediction using Machine Learning Algorithms –A Review," in *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, Dec. 2020, vol. 840, pp. 140–145, doi: 10.1109/ICACCCN51052.2020.9362912.

[2] N. S. Gupta, Y. Mohta, K. Heda, R. Armaan, B. Valarmathi, and G. Arulkumaran, "Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis," *J. Environ. Public Health*, vol. 2023, pp. 1–26, Jan. 2023, doi: 10.1155/2023/4916267.

[3] U. S. E. P. Agency and I. Division, "Air Quality Index (AQI)," in *Encyclopedia of Quality of Life and Well-Being Research*, no. February, Dordrecht: Springer Netherlands, 2014, pp. 120–120.

[4] T. H. Rochadiani *et al.*, "Design of Air Quality Monitoring Using LoRaWAN In Human Settlement," in *2022 IEEE 14th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, Dec. 2022, pp. 1–6, doi: 10.1109/HNICEM57413.2022.10109605.

[5] C. Wu, J. Wang, X. Chen, P. Du, and W. Yang, "A novel hybrid system based on multi-objective optimization for wind speed forecasting," *Renew. Energy*, vol. 146, pp. 149–165, Feb. 2020, doi: 10.1016/j.renene.2019.04.157.

[6] A. Mosavi, F. Sajedi Hosseini, B. Choubin, M. Goodarzi, A. A. Dineva, and E. Rafiei Sardooi, "Ensemble Boosting and Bagging Based Machine Learning Models for Groundwater Potential Prediction," *Water Resour. Manag.*, vol. 35, no. 1, pp. 23–37, Jan. 2021, doi: 10.1007/s11269-020-02704-3.

[7] Y. Li and Y. Pan, "A novel ensemble deep learning model for stock prediction based on stock prices and news," *Int. J. Data Sci. Anal.*, vol. 13, no. 2, pp. 139–149, Mar. 2022, doi: 10.1007/s41060-021-00279-9.

[8] B. Zhang, M. Duan, Y. Sun, Y. Lyu, Y. Hou, and T. Tan, "Air Quality Index Prediction in Six Major Chinese Urban Agglomerations: A Comparative Study of Single Machine Learning Model, Ensemble Model, and Hybrid Model," *Atmosphere (Basel).*, vol. 14, no. 10, 2023, doi: 10.3390/atmos14101478.

[9] Y.-C. Liang, Y. Maimury, A. H.-L. Chen, and J. R. C. Juarez, "Machine Learning-Based Prediction of Air Quality," *Appl. Sci.*, vol. 10, no. 24, p. 9151, Dec. 2020, doi: 10.3390/app10249151.

[10] C. Li, Y. Li, and Y. Bao, "Research on Air Quality Prediction Based on Machine Learning," in *2021 2nd International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI)*, Nov. 2021, pp. 77–81, doi: 10.1109/ICHCI54629.2021.00022.

[11] J. Zheng, D. Xin, Q. Cheng, M. Tian, and L. Yang, "The Random Forest Model for analyzing and Forecasting the US Stock Market under the background of smart finance," 2024, pp. 82–90.

[12] S. Park, S. Jung, J. Lee, and J. Hur, "A Short-Term Forecasting of Wind Power Outputs Based on Gradient Boosting Regression Tree Algorithms," *Energies*, vol. 16, no. 3, p. 1132, Jan. 2023, doi: 10.3390/en16031132.

[13] K. Mukherjee, S. S. Ahmed, M. Aasif, S. Kundu, and S. Ghosh, "House Rent Prediction Using Ensemble-Based Regression With Real-Time Data," in *Novel Research and Development Approaches in Heterogeneous Systems and Algorithms*, no. March 2023, 2023, pp. 258–271.

[14] J. T R, N. S. Reddy, and U. D. Acharya, "Modeling Daily Reference Evapotranspiration from Climate Variables: Assessment of Bagging and Boosting Regression Approaches," *Water Resour. Manag.*, vol. 37, no. 3, pp. 1013–1032, Feb. 2023, doi: 10.1007/s11269-022-03399-4.

[15] D. Iskandaryan, F. Ramos, and S. Trilles, "Air Quality Prediction in Smart Cities Using Machine Learning Technologies Based on Sensor Data: A Review," *Appl. Sci.*, vol. 10, no. 7, p. 2401, Apr. 2020, doi: 10.3390/app10072401.

[16] M. M. Londhe, "Data Mining and Machine Learning Approach for Air Quality Index Prediction," *Int. J. Eng. Appl. Phys.*, vol. 1, no. 2, pp. 136–153, 2021, [Online]. Available: https://ijeap.org/.