# Harnessing the XG-Boost Ensemble for Intelligent Prediction and Identification of Factors with a High Impact on Air Quality: A Case Study of Urban Areas in Jakarta Province, Indonesia

Wahyu Wibowo[1] and Harun Al Azies[2], Susi A. Wilujeng[3]

[1] Department of Business Statistics, Faculty of Vocational Studies,
Institut Teknologi Sepuluh Nopember, 60111, Surabaya, Indonesia
[2] Study Program in Informatics Engineering, Faculty of Computer Science,
Universitas Dian Nuswantoro, 50131, Semarang, Indonesia
[3]Department of Environmental Engineering, Faculty of Civil, Planning, and Geo Engineering,
Institut Teknologi Sepuluh Nopember, 60111, Surabaya, Indonesia
wahyu_w@statistika.its.ac.id

**Abstract.** This article aims to develop an accurate air quality prediction model to handle Jakarta's air pollution challenges. In this study, data from air quality monitoring stations' conventional air pollution indexes were employed. In the research phase, data is explored, SMOTE is used to manage imbalances, and XG-Boost is used to develop a model with the best parameters. The evaluation stage shows the model's ability to predict air quality. With an accuracy rate of 99.516%, an F1 score of 99.528%, and a recall rate of 99.509%, the results were very astounding. These performance indicators show the model's exceptional ability to classify and predict air quality levels. Furthermore, this study investigates the significance of various variables in predicting air quality. A thorough evaluation of measures such as weight, gain, total gain, and cover indicators reveals the significance of numerous aspects. Even while SO2 helps predict air quality, the prevalence of PM2.5 on several measures reveals a significant influence. This study contributes to a better understanding of the complicated dynamics of air quality prediction by employing advanced analytical approaches and accurate models. This knowledge is useful in developing targeted solutions to address air pollution issues and promote healthier urban environments.

**Keywords:** Air Quality Prediction, XGBoost Algorithm, Jakarta Air Pollution, Predictive modeling

## 1 Introduction

The consequences of air pollution are deeply concerning. Air pollution has had major implications, including increasing dangers to human health [1]. According to the World Health Organization (WHO), the number of deaths from air pollution reached an alarming 4.2 million in 2019 [2]. The situation is compounded further by the growing number

of vehicles in Indonesia, which emit hazardous petrol pollutants [3]. Carbon monoxide (CO) concentrations are also a factor in global warming and temperature fluctuations on Earth [4]. The amount of contaminants in the air has significantly increased recently due to Indonesia's rapid economic growth and social development [5]. The main concern is the assessment of environmental air quality and the control of air pollution, which is a serious issue with effects on many facets of daily life, the environment, and the ecosystem as a whole, particularly in urban areas like Indonesia's DKI Jakarta Province. Urban air quality is under threat and becoming a source of concern, notwithstanding the exciting economic expansion and the rapid rate of urbanization [6].

Government departments use the Air Quality Index (AQI) to notify the public about the present or predicted levels of air pollution and to monitor air quality [7]. The higher the AQI score, the more of the population is likely to suffer from an adverse health impact, which can be severe. The AQI is a measurement technique that indicates the five primary characteristics of air pollution that are the subject of observation, namely carbon monoxide (CO), sulfur dioxide (SO2), nitrogen dioxide (NO2), surface ozone (O3), and dust particles with a size of 10 micrometers or less (PM10) [7], [8]. As a result, timely air quality prediction is critical for the government to understand the pattern of changes in air quality and can be used to implement efficient air pollution control and management strategies. The World Air Quality Index (WAQI), which updates its data in real time, gives Jakarta an AQI score of 167. Jakarta is now ranked second on the list of the world's most polluted cities, trailing only Dubai City in the United Arab Emirates, which holds the top spot with an AQI score of 176. This graph demonstrates the significant problems these two cities are facing with air pollution, which has a detrimental effect on both the general public's health and the environment. Jakarta is listed as the city with the worst air quality, particularly in the Southeast Asia (ASEAN) area. The issue of air pollution has significant effects on sustainability, the environment, and human health.

However, it is becoming obvious that simply monitoring air quality is insufficient to address the underlying causes of air pollution. In recent years, several approaches for predicting air pollution have been developed. Although the Gaussian dispersion model is commonly employed in many research on air pollution [9], [10], other statistical techniques can be used to forecast pollutant concentrations [11]–[15]. Although these models are based on physical notions, detailed information regarding pollution sources and other factors is not always available. To get past this limitation, a more comprehensive technique is required. Machine learning algorithms are one way that seeks to deliver a more detailed response [16]–[18]. As a result, the focus of this research will be on establishing an evaluation framework to forecast AQI levels using machine learning approaches. In this context, the XGBoost algorithm is critical in overcoming the shortcomings of prior techniques. This study's innovative framework incorporates not just machine learning algorithms but also processes for dealing with imbalanced data and parameter tuning, which can increase prediction quality. The use of the XGBoost algorithm, imbalanced data processing techniques, and parameter adjustment results in a more comprehensive approach to predicting air quality and addressing complicated air pollution concerns, particularly in urban regions like DKI Jakarta Province, Indonesia. The paper is structured as follows: Section II presents related work on air quality

prediction research. Section III presents the research framework. Section IV presents the numerical results obtained. Finally, Section V provides conclusions and directions for further development.

## 2 Related Works

The provided literature review gives a general overview of the many methodologies used to predict air quality using machine learning techniques. Ma et al. (2019) studied air quality in Guangdong, China, using a two-way short-term memory transfer model (TL-BLSTM). Transfer learning is used in this technique to optimize predictions by exploiting information with lesser temporal resolution [19]. Ameer et al. (2019) compared various machine-learning approaches for predicting air quality in smart cities. They evaluated model performance using Apache Spark and evaluation criteria such as mean absolute error (MAE) and root mean square error (RMSE) [20]. Madan et al. (2020) provide a study that predicts air quality using several machine learning algorithms such as linear regression, decision trees, random forests, artificial neural networks, and support vector machines. This study makes use of Kaggle datasets, which are divided into training and test data [21]. Harishkumar et al. (2020) projected particulate matter concentrations in the air using Taiwan Air Quality Monitoring data. In terms of predictive ability, the suggested model outperforms the standard model when statistical metrics such as RMSE, MAE, MSE, and the Coefficient of Determination ($R^2$) are used [22]. Meanwhile, Du et al. (2019) created a hybrid deep learning system to predict PM2.5 air pollution using one-dimensional CNN and Bi-LSTM. This method produces accurate estimations of air pollution [23].

Castelli et al. (2020) predicted pollutant and particle levels, as well as the air quality index (AQI), in California using the Support Vector Regression (SVR) approach using a radial basis function (RBF) kernel. They were able to estimate hourly pollutant concentrations with great accuracy, including the AQI classification into six US Environmental Protection Agency categories [24]. Navares and Aznarte (2020) used a recurrent neural network (LSTM) to solve the challenge of predicting air quality in Madrid. They discovered that comprehensive models that combine many components outperform individual models and can provide helpful estimates in managerial and clinical settings [25]. Zhou et al. (2019) proposed using a Deep Multilayer Long Short-Term Memory (DM-LSTM) model to forecast air quality in Taipei City, Taiwan, several stages in advance. This approach improved the consistency and accuracy of regional air quality forecasts [26]. Al-Janabi et al. (2021) developed an intelligent computing-based technique for predicting air pollution. They divided the data into training and testing sections using cross-validation. To calculate the accuracy of air quality prediction, the symmetric mean absolute perception error (SMAPE) is used [27].

This study differs significantly from the previously cited literature studies. Although some prior studies have used various machine learning approaches and artificial neural networks to forecast air quality, this study focuses specifically on the Jakarta air quality scenario. Unlike previous research conducted in California, Madrid, or Taipei, this study obtained data from air quality monitoring stations in DKI Jakarta Province,

Indonesia. Furthermore, the Synthetic Minority Oversampling Technique (SMOTE) was used in this work to address the issue of an imbalance in the number of samples in different classes. By doing this, it is made possible for the developed model to deal with data inconsistencies and provide predictions of Jakarta's air quality that are more accurate. Additionally, the XGBoost method and the Randomized Search approach are employed for parameter tuning throughout the model construction phase. To find the ideal parameter that yields the most accurate predictions, the model is then randomly run through a collection of potential parameter combinations. With the help of this method, the model performs better and is better able to forecast Jakarta's air quality.

## 3 Materials and Methods

This study employs a research framework composed of four primary steps to analyze air quality in the DKI Jakarta Province area. This methodological approach is deliberately created to comprehend the components that influence air quality and to create accurate predictive models. Each stage of the methodology has a specific goal that adds to the overall success of the research.

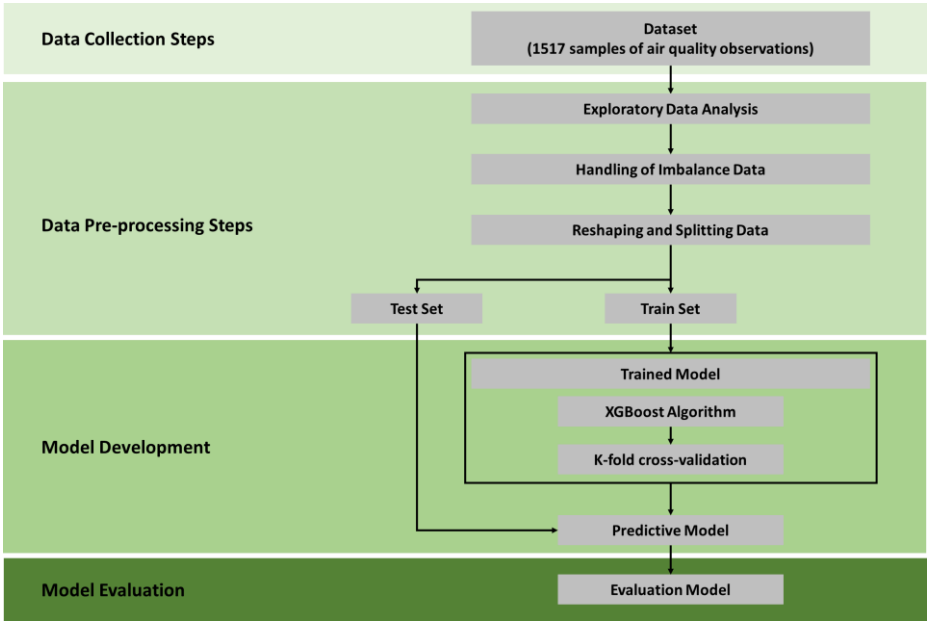


**Fig. 1.** Research Framework for Air Quality Prediction Models in Jakarta, Indonesia.

The first stage is data collection, which outlines how data is obtained from current sources, according to Figure 1, which serves as the foundation for this study. The following phase, known as data processing, explains the actions performed to prepare data before using it to train the model. The next section of this stage's model development describes the process of developing the model or algorithm for predicting air quality.

Analyzing a model's performance, the aim of this stage is to evaluate the performance of the model that is currently being created.

### 3.1 Data Collection Steps

This research dataset examines the Air Quality Index (AQI) as recorded by five air quality monitoring sites located throughout DKI Jakarta Province in 2021. This information comes from a reliable source, Jakarta Open Data, which can be viewed through the official website: https://data.jakarta.go.id/. The AQI is an important metric for assessing air quality in a given area. Based on the degree of air pollution by various pollutants such as PM10, PM25, SO2, CO, NO2, and O3, this index provides an overview of how excellent or terrible air quality is at a given time. This dataset contains variables that characterize several air quality metrics measured in DKI Jakarta Province. Every attribute and target variable has additional details below. This dataset's attribute variables describe several aspects of air quality metrics measured in the DKI Jakarta Province area. Each attribute variable comprises data about the concentration of a certain substance in the air, which is utilized as a feature to predict air quality.

This study's attribute variables include PM10, which refers to a material particle having a diameter of fewer than 10 micrometers (PM10) [8]. These particles are inhalable and have the potential to harm human health, particularly the respiratory tract. A material particle (PM2.5) having a diameter of fewer than 2.5 micrometers is referred to by the PM25 property. These particles are smaller than PM10 and can enter the lungs of people more easily, having more detrimental effects on their health [8]. The air contains sulfur dioxide, or SO2, which is measured as SO2. A gas called SO2 is produced by human activities like the burning of fossil fuels and is bad for both the environment and human health [8]. The symbol CO represents the concentration of carbon monoxide (CO). Numerous things, including motor vehicles, can emit CO, which can be hazardous if inhaled in large quantities. O3 denotes the amount of ozone (O3) in the atmosphere. Ozone is a dangerous contaminant on the Earth's surface, but it also serves as a natural barrier against ultraviolet light in the upper atmosphere. The amount of nitrogen dioxide (NO2) in the atmosphere is referred to as NO2 [8]. NO2, which is mostly created by human activities such as fuel combustion, may affect the human respiratory system. Meanwhile, the predicted category of air pollution standard index (PSI) is the target variable. This is the outcome of the Air Pollution Standard Index (PSI) calculation. PSI is an air quality metric that describes how safe or dangerous the air we breathe is. To offer an indication of the level of air pollution, this category covers many levels of "good," "moderate," and "unhealthy" air quality.

### 3.2 Data Pre-processing Steps

The data acquired in the previous stage will go through a pre-processing process at this level. This procedure ensures that the data is of good quality and ready for use in model building. This stage is divided into three major sub-stages. Exploratory Data Analysis (EDA): At this step, the data is examined to discover patterns and other information that can aid in a deeper understanding of the data's features [26], [28]. In this study, the correlation matrix was also used to investigate the association. This EDA stage will

enable researchers to have a deeper understanding of the data being used. This will help with the selection of relevant features for model construction, as well as the detection of any errors or irregularities in the data that must be rectified. Furthermore, correlation analysis can provide preliminary insight into how these parameters relate to one another, which can benefit later model interpretation.

Imbalance Data Handling Stage: There is frequently an imbalance in the number of samples in each target class in the dataset used for air quality prediction. For example, the number of samples in the "good" or "moderate" category could be significantly more than the number of samples in the "unhealthy" category. Because of this imbalance, the model may perform better when forecasting the majority class, while the minority class may perform poorly. The Synthetic Minority Over-Sampling Technique (SMOTE) is an oversampling technique that focuses on minority classes by developing synthetic samples that look like existing minority samples [29], [30]. SMOTE collects samples from the minority class and then generates a new synthetic sample by connecting existing points [31]. This increases the number of samples in the minority class without having to repeat the present sample. The number of samples in the minority class becomes more balanced with the majority class when SMOTE is used [32]. This will help the model learn more effectively and avoid bias toward the dominant class. In other words, the model will be better at predicting the minority class, which is usually more essential in situations like these, where poor air quality predictions have a big impact on health and the environment.

Data splitting: After processing, the data is divided into training and test data in a predetermined proportion. Training data is used to train the model, whereas test data is used to evaluate the model's performance. The proportion of training data to test data in this study is 70:30 [33], [34], with the majority of the data used to train the model (70%), and just a small percentage (30%) used for testing. This ratio ensures that the model has enough opportunities to learn from the training data and that it also performs well on the test data.

### 3.3 Model Development

The model development stage is critical in this study since it will build an air quality prediction system using the XG-Boost algorithm [35]. The XG-Boost method was selected for a variety of reasons, including its ability to handle complex data and give superior results in classification challenges. The XG-Boost algorithm will be used to develop an air quality prediction model at this step. The XG-Boost algorithm was chosen because of its strong reputation in the data science and artificial intelligence communities [36]. XG-Boost (Extreme Gradient Boosting) is an ensemble learning algorithm that combines several simple predictive models (weak learner) into a stronger model (strong learner) [37], [38]. This is done by using reinforcement techniques that can reduce the bias and variance that exists in the model. Before training the model, it is necessary to determine the optimal parameters for the XG-Boost algorithm. This step was performed using a random search method [39]. This method will try various combinations of different parameters and measure their performance using cross-validation techniques [40]. The optimal parameters will be selected based on the best performance

results during cross-validation. Once the optimal parameters are found through a random search, the XG-Boost model will be trained using the training data and these parameters using the K-Fold validation technique [41]. In this technique, the training data will be split into several folds and the model will be trained and evaluated at each fold. This ensures that the developed model can generalize well to a wide variety of data. By using the XG-Boost algorithm and optimized parameters, as well as involving K-Fold Validation in the evaluation, this model development stage aims to produce an air quality prediction model that has high performance and can generalize to data that has never been seen before [42]. This model will be ready to be tested at the evaluation stage to see how well it performs in predicting air quality based on the attributes in the dataset.

### 3.4 Model Evaluation

The final stage is model evaluation, in which the predictive model's performance is tested using important metrics such as accuracy, F1 score, and recall [43], [44]. Accuracy is a metric that measures how successfully a classification model predicts each class in a dataset. It involves dividing the total number of predictions made by the model by the number of accurate forecasts across all classes. Accuracy provides a summary of the model's overall performance [45]. Recall is a metric that assesses how well a model can detect real positive cases among all positive cases in a dataset. In multiclass classification, recall is calculated independently for each class. In instances when avoiding false negatives (not recognizing true positive cases) is critical, recall is key [46]. The F1-score is a metric that combines recall and accuracy, or the model's ability to recognize positives correctly. The F1-score provides the harmonic average of these two values. In multi-class classification, the F1 score can be calculated independently for each class, and then the average of all F1 scores for that class is determined[45].

## 4 Results and Discussion

Before going on to the next step of modeling to predict air quality based on the standard air pollution index in DKI Jakarta Province, this stage is an exploratory data analysis (EDA), which attempts to understand the features of the data. To comprehend the distribution of the values for each attribute and how frequently each value occurs, the results of the EDA analysis using this histogram indicate how the data is spread out. This is particularly helpful in finding trends and traits of the attributes that would influence the outcomes of predictions for air quality. In accordance with Figure 2, various attribute variables are essential in determining air quality, including PM10, PM25, SO2, CO, O3, and NO2. In this scenario, the three attributes, namely PM10, PM25, and SO2, have a distribution close to the bell curve (normal distribution), implying that the majority of air quality measurements on these attributes are centered around the average value with relatively low variation. The distribution, which is comparable to the bell curve on the PM10, PM25, and SO2 characteristics, reveals that particle and sulfur dioxide concentrations tend to be close to the mean or median values most of the time. This shows that the concentrations of PM10, PM2.5, and sulfur dioxide particles are often steady. Because air quality typically has steady pollution levels, these

characteristics may be significant factors in gauging air quality generally in the prediction context. The distribution of the other attributes, particularly CO, O3, and NO2, is positive skew. The distribution tail of the positive skew distribution extends to the right or beyond. This shows that measures of air quality typically have low values, but that certain measurements are higher or more extreme.

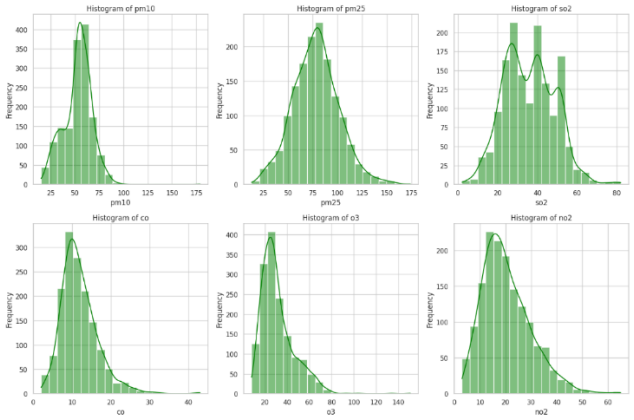


**Fig. 2.** The distribution pattern of each attribute variable.

According to the positive skew distribution, the majority of attribute values are in the low to moderate range, with greater extreme values occurring less frequently. This shows that there have been times when the levels of nitrogen dioxide, ozone, and carbon monoxide have drastically increased, which can have a considerable impact on air quality. These characteristics can be used by predictors to pinpoint times when air pollution is at its highest.

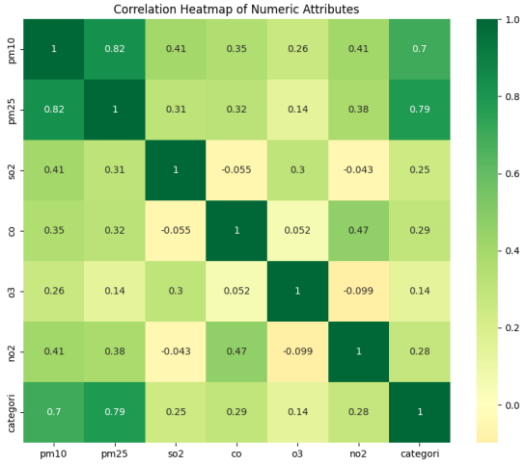


**Fig. 3.** Correlation Heatmap of Numeric Attributes: Each color's darker contrast denotes a strong link, either positive or negative. Brighter contrast, on the other hand, denotes a weak or absent link.

According to Figure 3, a high correlation between PM10 and PM25 particle concentrations shows that changes in PM10 typically follow changes in PM25. The modestly positive connection between PM10 and SO2 and PM10 and NO2 particles, however, suggests that gaseous pollutants like SO2 and NO2 can coexist with big airborne particles. The fact that there is little association between the pollution parameters (CO, O3, SO2, and NO2) and air quality suggests that the quality of the air is also influenced by the weather, pollution sources, and other intricate interactions. In some low negative relationships between pollution parameters, such as those between SO2 and NO2 and between O3 and NO2, one pollutant can become more concentrated while another pollutant becomes less concentrated. The next stage is to verify the number of classes in the air quality target variable. The goal of this stage is to determine whether or not there is an imbalance in the distribution of data among air quality categories. If the distribution of data between air quality classes is highly unequal, the performance of the model to be created may suffer, especially if the model tends to favor the majority class while ignoring the minority class.

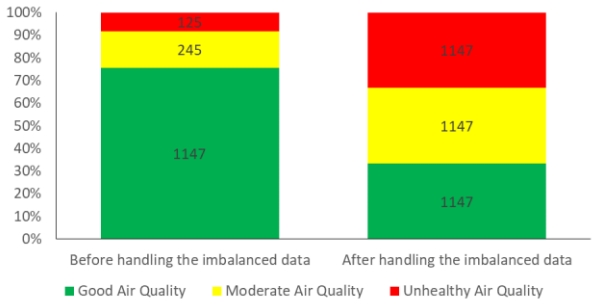


**Fig. 4.** Comparison of the distribution of air quality data before and after handling the imbalance.

According to preliminary examination, the class distribution of air quality parameters appears to be unbalanced. This experiment uses the SMOTE (Synthetic Minority Oversampling Technique) approach to manage uneven data to address this imbalance. Using the features of existing minority samples, this method generates synthetic data for the minority class. As seen in Figure 4, the data distribution was adequately balanced after applying the SMOTE technique. After handling the data imbalance using the SMOTE approach, the next step in this research is to divide the data into training data and test data with a ratio of 70:30, resulting in a composition of 2408 samples of training data size and 1033 samples of test data size. The objective of this division is to ensure that, in addition to generalizing to training data, the developed model can perform well on data that has never been seen before. After the data has been exchanged, the next step is to build a model using the XG-Boost technique. In classification tasks, one of the most efficient algorithms is widely used. The XG-Boost algorithm will have some parameters adjusted in this study to determine the ideal configuration. These are the parameters that were used:

a. **Subsample**: This setting regulates the number of samples utilized in each iteration of the tree-building process. Since only a small portion of the data is used

in each iteration, choosing lower values for the subsample can aid in preventing overfitting [47]. Configurations 0.6, 0.7, and 0.8 are used in this investigation.

b. **n estimators**: which are parameters specifying the number of trees to be built in the algorithm [48]. A larger number of trees generally increases the model's ability to capture complex patterns in the data. A total of 100, 200 are configurations of n estimators in this study.

c. **max depth**: 3, 4, and 5 different variations are used. This parameter controls the maximum depth of each tree in the ensemble [49]. The right setting for max depth is necessary to avoid overfitting [50]. If the model is given too much depth, it could be able to detect noise in the training data.

d. **Learning rate**: How many steps are changed with each iteration of the tree weights depends on the learning rate [51]. Choosing the right learning rate can have an impact on model convergence and overall performance. If the learning rate is too high, the model could fail to detect the global minimum. This arrangement makes use of 0.01 and 0.02.

e. **Gamma:** Gamma configurations 0, 0.1, and 0.2 were applied in this experiment. When a tree is formed, the gamma value controls when the nodes split [52]. The right gamma value can help prevent overfitting by controlling the tree's complexity.

f. **Colsample by tree:** This option regulates the proportion of features used to build each tree. The model can perform better in generalization and be more resilient to overfitting by selecting a feature subset [53]. The settings are 0.4, 0.5, and 0.6.

This experiment employs parameter settings to test various parameter value combinations to discover the best combination that delivers optimal model performance for the Jakarta air quality forecast case. After tuning, ideal parameters will be acquired, which may be utilized to construct the best-performing XG-Boost model.

**Table 1.** Parameter tuning results using the randomized search approach.

| Experiment | Parameters | | | | | | Mean Test Score | Standard Test Score |
|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | | |
| 1 | 0.8 | 100 | 3 | 0.02 | 0 | 0.6 | 0.997093 | 0.002116 |
| 3 | 0.8 | 200 | 5 | 0.01 | 0 | 0.6 | 0.997094 | 0.002490 |
| 7 | 0.7 | 200 | 5 | 0.02 | 0.1 | 0.5 | 0.996678 | 0.001017 |
| 11 | 0.7 | 200 | 3 | 0.02 | 0.2 | 0.6 | 0.996678 | 0.002116 |
| 14 | 0.8 | 100 | 4 | 0.02 | 0 | 0.5 | 0.996679 | 0.002815 |
| 18 | 0.8 | 200 | 5 | 0.01 | 0 | 0.6 | 0.996679 | 0.002815 |
| 19 | 0.7 | 200 | 3 | 0.02 | 0 | 0.5 | 0.997509 | 0.002421 |
| 22***** | 0.7* | 200* | 5* | 0.01* | 0.1* | 0.6* | 0.997509 | 0.002034 |
| 28 | 0.6 | 200 | 4 | 0.02 | 0.2 | 0.5 | 0.996678 | 0.002116 |
| 29 | 0.8 | 200 | 4 | 0.02 | 0.1 | 0.5 | 0.997094 | 0.002490 |

Note: A = Subsample, B = n estimators, C = max depth, D = Learning rate, E = Gamma, F = Colsample by tree, *) Selected Parameter Combinations

The experimental results of the parameters in Table 1 (only the 10 best combinations are shown) show the various parameter combinations tested, as well as the average value (mean test score) and standard deviation (standard test score) of the XG-Boost model's performance on the SMOTE dataset. The optimal parameter combination will be chosen to construct the final model. Table 1 shows many parameter combinations with high mean test scores, showing that models with these parameters can produce good results in predicting air quality in Jakarta. The parameters are searched in this case for combinations with the highest possible mean test score and the lowest possible standard test score. The parameter combinations with the greatest mean test score in the 22nd trial, according to Table 1, are: subsample: 0.7, n estimators: 200, max depth: 5, learning rate: 0.01, gamma: 0.1, and sample by tree: 0.6. The mean test score for this combination is 0.997509, while the standard test score is 0.002034. Because it has a high mean test score, which suggests good performance, and a low standard test score, which indicates model consistency on varied test data, this experiment is regarded as a good choice for constructing an accurate and stable model for predicting air quality in Jakarta. This parameter combination is used to train the model on all training data.

**Table 2.** K-Fold validation results.

| Performance Metric | Accuracy | F1-score | Recall |
|---|---|---|---|
| Fold 1 | 0.9938 | 0.9937 | 0.9938 |
| Fold 2 | 0.9979 | 0.9979 | 0.9979 |
| Fold 3 | 0.9979 | 0.9979 | 0.9978 |
| Fold 4 | 0.9979 | 0.9978 | 0.9979 |
| Fold 5* | 0.9979 | 0.9979 | 0.9980 |

Note: *) Selected Folds

The model's performance in each fold is shown by the results of 5-Fold Cross Validation (CV) [54]. Table 2 shows that the model's performance at each fold is quite good. For each fold, the accuracy, F1-score, and average recall are in the 0.9937 to 0.9980 range. This shows that the computer can predict Jakarta's air quality accurately. Overall, Fold 5 has the highest F1 score and recall, indicating that the model performs best when categorizing data in this fold. Based on the performance evaluations in each fold, the 5th fold is the best choice to represent the overall performance of the generated XG-Boost model. Whenever the performance results for each fold are good and consistent, the best model is chosen to create predictions on data that has never been seen before (test data). In this case, the best model can be selected based on previous experimental discoveries of the best parameters and folds. The best model may be employed to predict air quality in Jakarta based on new data.

Figure 5 shows the confusion matrix, which reflects the model's predicted results on the test data. This confusion matrix illustrates how well the model classifies each class. The algorithm correctly identified 324 samples as "good," whereas 0 samples were incorrectly labeled as "good." The model correctly classified 369 samples in the "Medium" category, whereas no samples were incorrectly classified as "Medium." The

program correctly identifies 335 samples as "unhealthy," but 5 samples are incorrectly labeled as such. Overall, the model does an excellent job of classifying air into "good" and "medium" categories with flawless accuracy and precision. However, there are some forecast inaccuracies in the "unhealthy" air class. This may be an area that requires additional effort to improve model performance in classifying underrepresented classes.

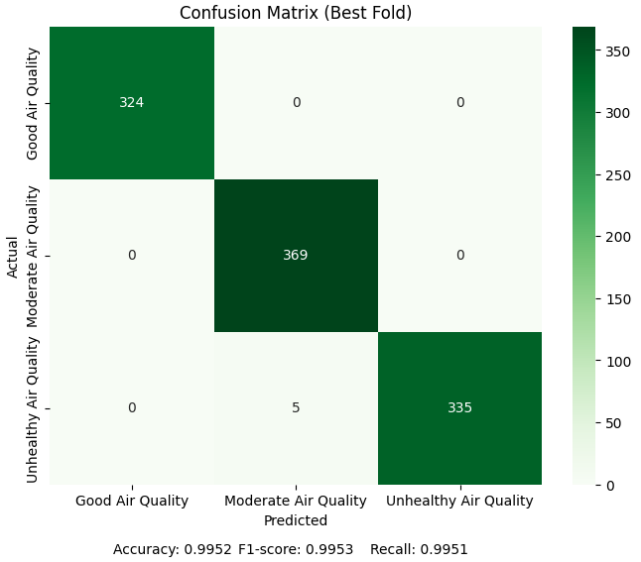


**Fig. 5.** Confusion Matrix based on Modeling Results with Testing Datasets.

The confusion matrix shows the results of the model's performance on test data (testing), which received an accuracy value of 0.995160. The accuracy of the model is measured by how well it classifies the data as a whole. With an accuracy value of roughly 0.995, it indicates that the model successfully predicted 99.5% of all test data. Furthermore, the harmonic mean of precision and recall is 0.995288 in the F1-Score. The F1-Score value of roughly 0.995 implies a balance between accuracy and recall, indicating that the model performs well in measuring precision and recognizing true positives. The last metric is the recall value, which compares the model's ability to recognize true positives to the total number of actual positives. A recall value of roughly 0.995 suggests that the model is quite good at classifying positive classifications.

After determining the optimal model, a more in-depth examination of the significance of variables is required to comprehend the role and influence of each feature on Jakarta air quality predictions. The importance of each variable on model predictions is measured by the variable's significance. According to this theory, high-importance features have a larger impact on prediction choices than low-importance features. The significance of factors in predicting Jakarta's air quality can shed light on the characteristics that are most crucial in determining air pollution levels. This information can help

anyone understand the factors that affect air quality, including environmental scientists, policymakers, and the general public.

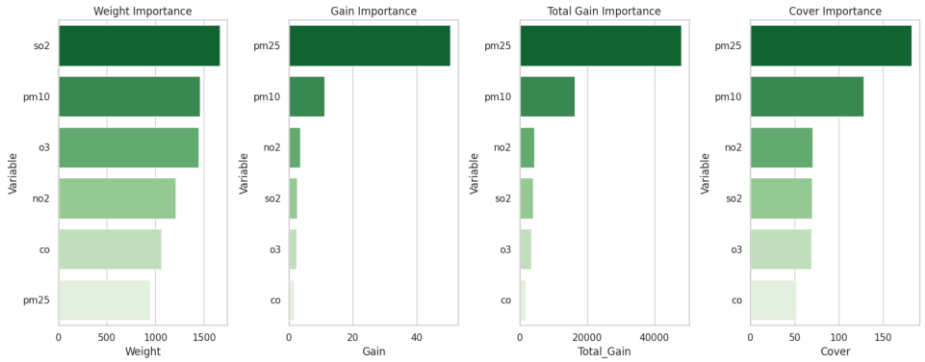


**Fig. 6.** Comparison of Importance Variable Indicators.

The variable importance measuring approach in this study makes use of several indicators, including weight, gain, total gain, and cover as shown in Figure 6. Each metric offers a unique perspective on the significance of a feature in terms of affecting the model's predictions. The interpretation of each variable importance metric employed is as follows:

a. **Weight:** A feature's relative value based on weight indicates how frequently a node in a decision tree is divided by it. The weight of a feature increases with the frequency of use. The SO2 (Sulfur Dioxide) feature has a high contribution to the Weight indicator. This shows that SO2 also has a significant influence on model predictions, although it may not be as strong as PM2.5 on other indicators.

b. **Gain:** When used to separate nodes in a decision tree, gain denotes a feature's contribution to an improvement in impurity reduction (typically using the Gini or Entropy technique) [55]. The contribution of a feature to increasing the model's accurate separation is inversely proportional to its gain. In this experiment, the PM25 feature performs remarkably well in Gain, indicating its essential contribution to improving the model's separation quality.

c. **Total Gains:** Total Gain is the sum of the gains realized by a feature and all of its derivative features. It describes the overall contribution of features and their derivatives to lowering impurity in the decision tree [55]. The PM25 feature has the largest overall gain, which denotes that it has contributed most to improving the model's separation abilities over time.

d. **Covers:** In a decision tree, a feature's impact on the amount of training data is quantified. The cover value increases as more data are impacted by the characteristic. The PM25 feature in this study has a high cover value, indicating that it affects a significant portion of the choices made by the model's trees.

The PM25 feature highlights three indicators: gain, total gain, and cover. This reveals that PM25 has a major impact on air quality estimates. Because PM2.5 is a fine particulate emitted by a range of sources, including autos, industry, and combustion, its dominance in three indices indicates that PM2.5 levels have a substantial impact on Jakarta's air pollution. Controlling PM2.5 emissions could be the primary focus of efforts to enhance air quality. Despite having low relevance values in two indicators (gain and cover), SO2 has a considerable impact on air quality predicting, as evidenced by its high weight value. SO2 levels in the air have a significant role in affecting air quality. The high importance of SO2 in the "Weight" indicator may indicate that fluctuations in SO2 levels in the air can have a large impact on overall air quality. Even though it dominates by weight indicator, it is critical to continue monitoring and controlling SO2 levels to maintain good air quality. CO has a low significance value for all indicators, indicating that CO has no meaningful impact on predicting air quality in the context of this study. However, it should be highlighted that CO levels should be regularly monitored because this gas can come from sources such as motor vehicles and combustion and can have major health consequences.

## 5 Conclusion

According to the findings of this study, the XGBoost parameters improved using random search have a considerable impact on model performance. The optimal parameter combination, with subsample values of 0.7, n estimators of 200, max depth of 5, learning rate of 0.01, gamma of 0.1, and sample by tree of 0.6, may offer reliable and consistent predicts of air quality. The 5-fold cross-validation technique yields a model with consistent performance across all folds. The accuracy, F1-score, and recall values in each fold range from 0.9937 to 0.9980. This demonstrates that the algorithm can accurately estimate air quality in Jakarta and is constant across different conditions. Furthermore, the test data evaluation results are excellent. The model has an accuracy of 0.995160, an F1-score of 0.995288, and a recall of 0.995098. This demonstrates that the model is highly good at categorizing air quality.

PM2.5 and SO2 are the key variables influencing air quality predictions in Jakarta, according to an examination of the importance of variables. This model can be used to understand and address the air pollution issues in Jakarta since PM2.5 has a substantial impact on three critical indicators, while SO2 has a strong impact on the "Weight" indicator, demonstrating its significant significance in air quality prediction. In their efforts to preserve better air quality, environmental scientists, legislators, and the general public can benefit greatly from this knowledge.

## References

1. J. L. Domingo and J. Rovira, “Effects of air pollutants on the transmission and severity of respiratory viral infections,” *Environ Res*, vol. 187, p. 109650, Aug. 2020, doi: 10.1016/J.ENVRES.2020.109650.

2. M. Liu *et al.*, “Population susceptibility differences and effects of air pollution on cardiovascular mortality: epidemiological evidence from a time-series study,” *Environmental Science and Pollution Research*, vol. 26, no. 16, pp. 15943–15952, Jun. 2019, doi: 10.1007/S11356-019-04960-2/FIGURES/1.
3. P. Lestari, M. K. Arrohman, S. Damayanti, and Z. Klimont, “Emissions and spatial distribution of air pollutants from anthropogenic sources in Jakarta,” *Atmos Pollut Res*, vol. 13, no. 9, p. 101521, Sep. 2022, doi: 10.1016/J.APR.2022.101521.
4. I. Mehmood *et al.*, “Carbon Cycle in Response to Global Warming,” *Environment, Climate, Plant and Vegetation Growth*, pp. 1–15, Jan. 2020, doi: 10.1007/978-3-030-49732-3_1/COVER.
5. A. Raihan, D. A. Muhtasim, M. I. Pavel, O. Faruk, and M. Rahman, “An econometric analysis of the potential emission reduction components in Indonesia,” *Cleaner Production Letters*, vol. 3, p. 100008, Dec. 2022, doi: 10.1016/J.CLPL.2022.100008.
6. G. McGranahan, J. Songsore, and M. Kjellén, “Sustainability, Poverty and Urban Environmental Transitions,” *The Earthscan Reader in Sustainable Cities*, pp. 107–133, Dec. 2021, doi: 10.4324/9781315800462-8.
7. F. Abulude, I. Abulude, S. Oluwagbayide, S. Afolayan, and D. Ishaku, “Air Quality Index: Case of One-Day Monitoring of 253 Urban and Suburban Towns in Nigeria,” *Environmental Sciences Proceedings 2021, Vol. 8, Page 4*, vol. 8, no. 1, p. 4, Jun. 2021, doi: 10.3390/ECAS2021-10342.
8. A. Mishra, Z. M. Jalaluddin, and C. V. Mahamuni, “Air Quality Analysis and Smog Detection in Smart Cities for Safer Transport using Machine Learning (ML) Regression Models,” *Proceedings - 2022 IEEE 11th International Conference on Communication Systems and Network Technologies, CSNT 2022*, pp. 200–206, 2022, doi: 10.1109/CSNT54456.2022.9787618.
9. A. Tiwari *et al.*, “Considerations for evaluating green infrastructure impacts in microscale and macroscale air pollution dispersion models,” *Science of The Total Environment*, vol. 672, pp. 410–426, Jul. 2019, doi: 10.1016/J.SCITOTENV.2019.03.350.
10. A. Masih, “Machine learning algorithms in air quality modeling,” *Global Journal of Environmental Science and Management*, vol. 5, no. 4, pp. 515–534, Oct. 2019, doi: 10.22034/GJESM.2019.04.10.
11. M. T. Lei, J. Monjardino, L. Mendes, D. Gonçalves, and F. Ferreira, “Macao air quality forecast using statistical methods,” *Air Qual Atmos Health*, vol. 12, no. 9, pp. 1049–1057, Sep. 2019, doi: 10.1007/S11869-019-00721-9/TABLES/4.
12. S. Chen, J. qiang Wang, and H. yu Zhang, “A hybrid PSO-SVM model based on clustering algorithm for short-term atmospheric pollutant concentration forecasting,” *Technol Forecast Soc Change*, vol. 146, pp. 41–54, Sep. 2019, doi: 10.1016/J.TECHFORE.2019.05.015.
13. H. Wang, Q. Yilihamu, M. Yuan, H. Bai, H. Xu, and J. Wu, “Prediction models of soil heavy metal(loid)s concentration for agricultural land in Dongli: A comparison of regression and random forest,” *Ecol Indic*, vol. 119, p. 106801, Dec. 2020, doi: 10.1016/J.ECOLIND.2020.106801.

14. S. Abdullah, M. Ismail, A. N. Ahmed, and A. M. Abdullah, “Forecasting Particulate Matter Concentration Using Linear and Non-Linear Approaches for Air Quality Decision Support,” *Atmosphere 2019, Vol. 10, Page 667*, vol. 10, no. 11, p. 667, Oct. 2019, doi: 10.3390/ATMOS10110667.
15. X. Su, J. An, Y. Zhang, P. Zhu, and B. Zhu, “Prediction of ozone hourly concentrations by support vector machine and kernel extreme learning machine using wavelet transformation and partial least squares methods,” *Atmos Pollut Res*, vol. 11, no. 6, pp. 51–60, Jun. 2020, doi: 10.1016/J.APR.2020.02.024.
16. M. Ali, A. Dewan, A. K. Sahu, and M. M. Taye, “Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions,” *Computers 2023, Vol. 12, Page 91*, vol. 12, no. 5, p. 91, Apr. 2023, doi: 10.3390/COMPUTERS12050091.
17. Y. Xu *et al.*, “Artificial intelligence: A powerful paradigm for scientific research,” *The Innovation*, vol. 2, no. 4, p. 100179, Nov. 2021, doi: 10.1016/J.XINN.2021.100179.
18. B. W. Otok, A. Suharsono, Purhadi, R. E. Standsyah, and H. Al Azies, “Partitional Clustering of Underdeveloped Area Infrastructure with Unsupervised Learning Approach: A Case Study in the Island of Java, Indonesia,” *Journal of Regional and City Planning*, vol. 33, no. 2, pp. 177–196, Aug. 2022, doi: 10.5614/JPWK.2022.33.2.3.
19. J. Ma, J. C. P. Cheng, C. Lin, Y. Tan, and J. Zhang, “Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques,” *Atmos Environ*, vol. 214, p. 116885, Oct. 2019, doi: 10.1016/J.ATMOSENV.2019.116885.
20. S. Ameer *et al.*, “Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities,” *IEEE Access*, vol. 7, pp. 128325–128338, 2019, doi: 10.1109/ACCESS.2019.2925082.
21. T. Madan, S. Sagar, and D. Virmani, “Air Quality Prediction using Machine Learning Algorithms-A Review,” *Proceedings - IEEE 2020 2nd International Conference on Advances in Computing, Communication Control and Networking, ICACCCN 2020*, pp. 140–145, Dec. 2020, doi: 10.1109/ICACCCN51052.2020.9362912.
22. Doreswamy, K. S. Harishkumar, Y. Km, and I. Gad, “Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models,” *Procedia Comput Sci*, vol. 171, pp. 2057–2066, Jan. 2020, doi: 10.1016/J.PROCS.2020.04.221.
23. S. Du, T. Li, Y. Yang, and S. J. Horng, “Deep Air Quality Forecasting Using Hybrid Deep Learning Framework,” *IEEE Trans Knowl Data Eng*, vol. 33, no. 6, pp. 2412–2424, Jun. 2021, doi: 10.1109/TKDE.2019.2954510.
24. M. Castelli, F. M. Clemente, A. Popovič, S. Silva, and L. Vanneschi, “A Machine Learning Approach to Predict Air Quality in California,” *Complexity*, vol. 2020, 2020, doi: 10.1155/2020/8049504.
25. R. Navares and J. L. Aznarte, “Predicting air quality with deep learning LSTM: Towards comprehensive models,” *Ecol Inform*, vol. 55, p. 101019, Jan. 2020, doi: 10.1016/J.ECOINF.2019.101019.

26. Y. Zhou, F. J. Chang, L. C. Chang, I. F. Kao, and Y. S. Wang, "Explore a deep learning multi-output neural network for regional multi-step-ahead air quality forecasts," *J Clean Prod*, vol. 209, pp. 134–145, Feb. 2019, doi: 10.1016/J.JCLEPRO.2018.10.243.
27. S. Al-Janabi, M. Mohammad, and A. Al-Sultan, "A new method for prediction of air pollution based on intelligent computation," *Soft comput*, vol. 24, no. 1, pp. 661–680, Jan. 2020, doi: 10.1007/S00500-019-04495-1/TABLES/15.
28. R. Indrakumari, T. Poongodi, and S. R. Jena, "Heart Disease Prediction using Exploratory Data Analysis," *Procedia Comput Sci*, vol. 173, pp. 130–139, Jan. 2020, doi: 10.1016/J.PROCS.2020.06.017.
29. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/JAIR.953.
30. A. D. Amirruddin, F. M. Muharam, M. H. Ismail, N. P. Tan, and M. F. Ismail, "Synthetic Minority Over-sampling TEchnique (SMOTE) and Logistic Model Tree (LMT)-Adaptive Boosting algorithms for classifying imbalanced datasets of nutrient and chlorophyll sufficiency levels of oil palm (Elaeis guineensis) using spectroradiometers and unmanned aerial vehicles," *Comput Electron Agric*, vol. 193, p. 106646, Feb. 2022, doi: 10.1016/J.COMPAG.2021.106646.
31. W. Wibowo and I. Dewi Ratih, "Classification of Non-Performing Financing Using Logistic Regression and Synthetic Minority Over-sampling Technique-Nominal Continuous (SMOTE-NC)," *Int. J. Advance Soft Compu. Appl*, vol. 13, no. 3, 2021, doi: 10.15849/IJASCA.211128.09.
32. M. F. Ijaz, G. Alfian, M. Syafrudin, and J. Rhee, "Hybrid Prediction Model for Type 2 Diabetes and Hypertension Using DBSCAN-Based Outlier Detection, Synthetic Minority Over Sampling Technique (SMOTE), and Random Forest," *Applied Sciences 2018, Vol. 8, Page 1325*, vol. 8, no. 8, p. 1325, Aug. 2018, doi: 10.3390/APP8081325.
33. A. Tella and A. L. Balogun, "GIS-based air quality modelling: spatial prediction of PM10 for Selangor State, Malaysia using machine learning algorithms," *Environmental Science and Pollution Research*, vol. 29, no. 57, pp. 86109–86125, Dec. 2022, doi: 10.1007/S11356-021-16150-0/TABLES/5.
34. B. Choubin *et al.*, "Spatial hazard assessment of the PM10 using machine learning models in Barcelona, Spain," *Science of The Total Environment*, vol. 701, p. 134474, Jan. 2020, doi: 10.1016/J.SCITOTENV.2019.134474.
35. L. Yao, T. M. T. Lei, S. C. W. Ng, and S. W. I. Siu, "Application of ANN, XGBoost, and Other ML Methods to Forecast Air Quality in Macau," *Sustainability 2023, Vol. 15, Page 5341*, vol. 15, no. 6, p. 5341, Mar. 2023, doi: 10.3390/SU15065341.
36. R. P. Sheridan, W. M. Wang, A. Liaw, J. Ma, and E. M. Gifford, "Extreme Gradient Boosting as a Method for Quantitative Structure-Activity Relationships," *J Chem Inf Model*, vol. 56, no. 12, pp. 2353–2360, Dec. 2016, doi: 10.1021/ACS.JCIM.6B00591/SUPPL_FILE/CI6B00591_SI_033.ZIP.
37. T. Chen and T. He, "xgboost: eXtreme Gradient Boosting," 2023.

38. J. Ma, J. C. P. Cheng, Z. Xu, K. Chen, C. Lin, and F. Jiang, “Identification of the most influential areas for air pollution control using XGBoost and Grid Importance Rank,” *J Clean Prod*, vol. 274, p. 122835, Nov. 2020, doi: 10.1016/J.JCLEPRO.2020.122835.
39. A. Nugroho and H. Suhartanto, “Hyper-Parameter Tuning based on Random Search for DenseNet Optimization,” *7th International Conference on Information Technology, Computer, and Electrical Engineering, ICITACEE 2020 - Proceedings*, pp. 96–99, Sep. 2020, doi: 10.1109/ICITACEE50144.2020.9239164.
40. L. Sun, “Application and improvement of Xgboost algorithm based on multiple parameter optimization strategy,” *Proceedings - 2020 5th International Conference on Mechanical, Control and Computer Engineering, ICMCCE 2020*, pp. 1822–1825, Dec. 2020, doi: 10.1109/ICMCCE51767.2020.00400.
41. J. Yang, P. Jiang, R. U. D. Nassar, S. A. Suhail, M. Sufian, and A. F. Deifalla, “Experimental investigation and AI prediction modelling of ceramic waste powder concrete – An approach towards sustainable construction,” *Journal of Materials Research and Technology*, vol. 23, pp. 3676–3696, Mar. 2023, doi: 10.1016/J.JMRT.2023.02.024.
42. R. Ghorbani and R. Ghousi, “Comparing Different Resampling Methods in Predicting Students’ Performance Using Machine Learning Techniques,” *IEEE Access*, vol. 8, pp. 67899–67911, 2020, doi: 10.1109/ACCESS.2020.2986809.
43. R. Choudhary, T. Gopalakrishnan, D. Ruby, A. Gayathri, V. S. Murthy, and R. Shekhar, “An Efficient Model for Predicting Liver Disease Using Machine Learning,” *Data Analytics in Bioinformatics: A Machine Learning Perspective*, pp. 443–457, Jan. 2021, doi: 10.1002/9781119785620.CH18.
44. W. Wibowo, R. Amelia, F. A. Octavia, and R. N. Wilantari, “Classification using nonparametric logistic regression for predicting working status,” *AIP Conf Proc*, vol. 2329, no. 1, Feb. 2021, doi: 10.1063/5.0043598/962507.
45. Muljono, P. N. Andono, S. A. Wulandari, H. Al Azies, and M. Naufal, “Tempo Recognition of Kendhang Instruments Using Hybrid Feature Extraction,” *Journal of Applied Science and Engineering*, vol. 27, no. 3, pp. 3177–2190, 2023, doi: 10.6180/JASE.202403_27(3).0004.
46. P. Ghosh *et al.*, “Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques,” *IEEE Access*, vol. 9, pp. 19304–19326, 2021, doi: 10.1109/ACCESS.2021.3053759.
47. Y. Dai, Q. Zhou, M. Leng, X. Yang, and Y. Wang, “Improving the Bi-LSTM model with XGBoost and attention mechanism: A combined approach for short-term power load prediction,” *Appl Soft Comput*, vol. 130, p. 109632, Nov. 2022, doi: 10.1016/J.ASOC.2022.109632.
48. M. Ahmad *et al.*, “Extreme Gradient Boosting Algorithm for Predicting Shear Strengths of Rockfill Materials,” *Complexity*, vol. 2022, 2022, doi: 10.1155/2022/9415863.
49. R. Guo, Z. Zhao, T. Wang, G. Liu, J. Zhao, and D. Gao, “Degradation State Recognition of Piston Pump Based on ICEEMDAN and XGBoost,” *Applied Sciences 2020, Vol. 10, Page 6593*, vol. 10, no. 18, p. 6593, Sep. 2020, doi: 10.3390/APP10186593.

50. X. Ren, H. Guo, S. Li, S. Wang, and J. Li, “A novel image classification method with CNN-XGBoost model,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10431 LNCS, pp. 378–390, 2017, doi: 10.1007/978-3-319-64185-0_28/COVER.
51. J. Chen, F. Zhao, Y. Sun, and Y. Yin, “Improved XGBoost model based on genetic algorithm,” *International Journal of Computer Applications in Technology*, vol. 62, no. 3, pp. 240–245, 2020, doi: 10.1504/IJCAT.2020.106571.
52. Y. Liang *et al.*, “Product marketing prediction based on XGboost and LightGBM algorithm,” *ACM International Conference Proceeding Series*, pp. 150–153, Aug. 2019, doi: 10.1145/3357254.3357290.
53. M. Parsa, “A data augmentation approach to XGboost-based mineral potential mapping: An example of carbonate-hosted ZnPb mineral systems of Western Iran,” *J Geochem Explor*, vol. 228, p. 106811, Sep. 2021, doi: 10.1016/J.GEXPLO.2021.106811.
54. J. P. Haumahu, S. D. H. Permana, and Y. Yaddarabullah, “Fake news classification for Indonesian news using Extreme Gradient Boosting (XGBoost),” *IOP Conf Ser Mater Sci Eng*, vol. 1098, no. 5, p. 052081, Mar. 2021, doi: 10.1088/1757-899X/1098/5/052081.
55. S. Deng, Y. Zhu, S. Duan, Z. Fu, and Z. Liu, “Stock Price Crash Warning in the Chinese Security Market Using a Machine Learning-Based Method and Financial Indicators,” *Systems 2022, Vol. 10, Page 108*, vol. 10, no. 4, p. 108, Jul. 2022, doi: 10.3390/SYSTEMS10040108.