INTERNATIONAL JOURNAL OF CONSERVATION SCIENCE
ISSN: 2067-533X Volume 17, Issue 2, 2026: 733-748 www.ijcs.ro
DOI: 10.36868/IJCS.2026.02.22

ENHANCING EARLY WARNING OF HIGH PARTICULATE MATTER EVENTS FOR ENVIRONMENTAL CONSERVATION IN MALAYSIA USING A RESAMPLING METHOD

Mahiran MUHAMMAD[1], Noor Fadhilah Ahmad RADI, Norazian MOHAMED NOOR[2], Ahmad Zia UL-SAUFIE[1,*]

[1] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
[2]Faculty of Civil Engineering & Technology, Universiti Malaysia Perlis, Jejawi, Arau 02600, Perlis, Malaysia.

Abstract

High particulate events or also known as extreme events of PM_{10} concentrations during pollution episodes have led to critical public health and environmental problems in Malaysia. The prediction of PM_{10} concentrations, particularly during high-impact events, remains a challenge due to the issue of imbalanced air pollution data when the model trains on more normal events and underestimates extreme PM_{10} concentrations. From the perspective of conservation science, this limitation will disrupt the role of early warning and risk mitigation. Therefore, a resampling method is crucial to improve model predictions towards these extreme events. This study proposed the Moving Block Bootstrapping with a ratio-based (MBB-Ratio) resampling technique to enhance predictive modeling during high PM_{10} episodes. The results show that the MBB-Ratio is able to mitigate data imbalance and boost predictive accuracy for extreme PM_{10} concentrations when the performance of XGBoost shows a notable improvement, where the RMSE, MAE, and MAPE dropped from 108.3010 to 18.9009 (82.5478%), 85.1041 to 13.6975 (83.905%), and 0.3634 to 0.0644 (82.2785%). This research contributes to the development and implementation of the MBB-Ratio resampling approach, specifically tailored to increase the representation of extreme events and enhance PM_{10} concentration prediction, particularly towards high-particulate events. Through this integration, the MBB-Ratio can be utilized to assist timely conservation efforts from related policies.

Keywords: *Particulate Matter; Extreme events; Imbalanced data; Moving Block Bootstrapping with Ratio-based (MBB-Ratio); Extreme Gradient Boosting (XGBoost)*

Introduction

Globally, haze events and rapid urbanization cause a crucial air pollution issue when these problems produce continuous pressure on the environment and ecosystem [1]. Nitrogen Dioxide (NO_2), Ozone (O_3), Sulphur Dioxide (SO_2), Carbon Monoxide (CO), and Particulate Matter (PM) are pollutants that highly affect most of the countries [2]. Malaysia has been exposed to transboundary haze events repeatedly, where the particulate matter, PM10, is highly concentrated, affecting human health and the environment [3]. PM_{10} is particulate matter less than 10 micrometers in diameter [3]. High concentrations of PM_{10} form haze that causes milky air with visibility restrictions, which affects human activities and the ecosystem [4]. These increased concentrations of these fine particles are often recognized as outliers, extreme, or high particulate air pollution events [5]. PM_{10} is one of the primary pollutants related to respiratory problems such

* Corresponding author: ahmadzia101@uitm.edu.my

as rhinitis, chronic obstructive pulmonary disease, upper respiratory infections, and asthma [6]. Therefore, PM_{10} concentration prediction during high PM_{10} events is vital for early steps in preserving health, conserving timely alerts, and mitigating public exposure.

Predicting extreme PM_{10} concentrations becomes difficult when air pollution is inherently imbalanced. The issue of an imbalanced air pollution dataset is the presence of normal observations that contain more than extreme air pollution events [7]. This imbalance can pose challenges in both classification and regression problems. The classification problems in an imbalanced dataset arise due to the presence of a minority class and a majority class [8]. Contrary to classification problems, regression problems are an issue when the target variable is continuous [9]. The main issue of the imbalanced problem in the air pollution dataset is underestimating the extreme events' predictions. This arises when the rare target values appear infrequently, causing the model training to follow normal observations and insufficiently capture these rare observations and, consequently, perform poorly on these extreme events [10], [11]. Therefore, a more comprehensive analysis of data distribution is crucial for enhancing predictive performance in imbalanced regression tasks, particularly to achieve a better prediction for high particulate events that produce substantial conservation management implications [12].

Strategies for imbalanced domain learning

As outlined by [8], three main strategies are categorized to address imbalanced regression problems: i) resampling, ii) evaluation metrics, and iii) regression models. Resampling is the most popular approach, which is typically used by research practitioners in dealing with imbalanced regression [10], [13-16]. The purpose of resampling is to change the data distribution before utilizing the model development [17]. The primary goal of this modification is to encourage the learning algorithm to pay greater attention to high particulate events [18]. These methods are widely adopted because resampling does not fix any learning algorithm, and the original data set distribution is allowed to be modified [7]. Nevertheless, [19] stressed that the success of that approach strongly depends on how the data was modified, which is always a challenge for researchers.

Former research has explored some resampling techniques to emphasize imbalanced regression problems. Several researchers overcome the imbalanced regression by applying methods that are based on the classification method called Synthetic Minority Oversampling Technique (SMOTE) [20]. Among these are SMOTE with Gaussian Noise (SMOGN) [19], Geometric SMOTE [21], SMOTE for regression (SMOTER) [22], SMOTE integrated with ensemble boosting (SMOTEBoost) [14], and others. SMOTER, SMOGN, and Geometric SMOTE produce synthetic observations through interpolation or noise-based techniques. Even though some of these approaches enhance the representation of high particulate events, they can also introduce noisy and unrealistic data points. As a result, the model may mislead patterns that reduce the generalization performance of machine learning models [23], [24]. Introducing synthetic data that does not represent real-data characteristics can diminish the integrity of the training set, possibly resulting in the model identifying patterns that lack generalizability to real observations. Together, the limitations of noisy data may decrease the model's capability to accurately predict over the full range of the target distribution, consequently reducing the model's overall prediction performance. Moreover, many of these techniques mainly aim to minimize overall prediction error, with limited emphasis on addressing extreme cases. In response to this limitation, [25] introduced the theory of imbalanced regression by investigating the correlation between the distribution and test error, where label distribution smoothing (LDS) and feature distribution smoothing (FDS) techniques were implemented. However, this technique still relies on interpolation to generate minority samples, which leads to overfitting.

Hence, to prevent the noisy and synthetic data and at the same time reduce the error prediction for extreme or high particulate events, Moving Block Bootstrapping (MBB) is introduced as a resampling method in dealing with the imbalanced air pollution data. MBB is one

of the variants that derives from the bootstrapping method, which involves resampling contiguous blocks of observed data [26]. This block-based approach improves data quality by maintaining the dependencies between consecutive observations, thereby offering a more realistic learning environment for predictive models [27]. [28] recommends the use of MBB as an alternative for improving the estimation of environmental datasets.

However, a study from [29] shows that the MBB resampling approach effectively enhances the model's performance in predicting normal events but fails to improve the prediction for high particulate events. Therefore, motivated by these constraints, Moving Block Bootstrapping (MBB) with a ratio-based resampling method (denoted by MBB-Ratio) is proposed for handling extreme values in the imbalanced air-pollution dataset. By intentionally adjusting the sampling to increase the representation of blocks containing rare PM_{10} observations, MBB-Ratio aims to create a more balanced dataset that allows the model to learn patterns associated with both normal and extreme air quality conditions more effectively. This enhancement is expected to improve the model's generalization, particularly for rare but impactful pollution events that are often underrepresented. The use of balanced input data significantly improves the accuracy of advanced models such as Extreme Gradient Boosting, which are particularly adept at learning from both normal and extreme pollution levels [30], [31], [32].

By applying the MBB-Ratio, this study contributed to robust resampling strategies that boost the detection and modeling of extreme pollution events. This analytical improvement strengthens the reliability of predictive models, helping conservation experts to understand the environmental problems associated with episodic air pollution, such as habitat degradation and disturbances to ecosystem balance [33]. Furthermore, improving air quality allows better awareness about the haze events, simultaneously decreasing their impact on vulnerable ecosystems. Timely and accurate predictions are crucial for enabling quick warnings, informing public health advisories, and shaping long-term sustainable environmental policies [34].

Experimental part

Research Flow

The research flow of this study is shown in Figure 1.

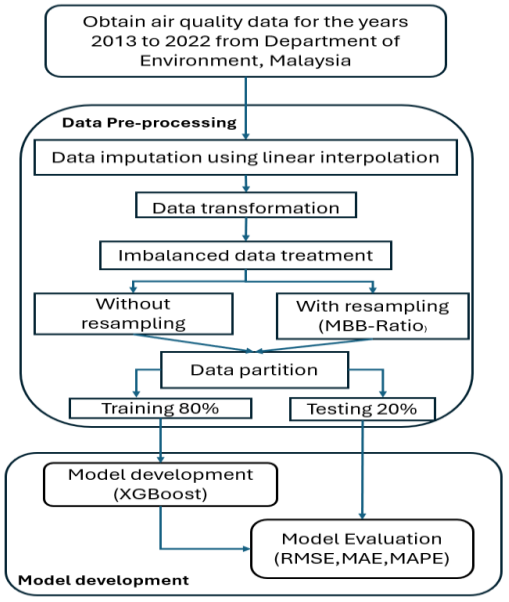


Fig. 1. Research Flowchart

The figure presents the development of a resampling approach for dealing with high particulate events in imbalanced air pollution using the MBB-Ratio strategy in Shah Alam, Malaysia. This study obtained the data from the Department of Environment (DOE), Malaysia. The air quality data covers from 2013 to 2022. The process starts with data retrieval from DOE. Then, before model development, the data goes through the data-processing stage, where the focus is on emphasizing the imbalanced data handling by using the MBB-Ratio resampling approach. Then, the Extreme Gradient Boosting (XGBoost) machine learning model is applied. This model is developed to compare the effectiveness of the XGBoost with the MBB-Ratio resampling approach against the original model without the resampling procedure. The model accuracy is evaluated based on the performance of Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Square Error (RMSE). Lastly, the MBB-Ratio approach is examined to determine whether it can serve as an effective resampling approach or vice versa.

Air Pollution Monitoring Dataset

The data consists of 83,431 hourly observations for 10 variables, including air pollutants and meteorological parameters. Table 1 shows the air pollutants and meteorology variables, where the PM_{10} concentrations for the next 24 hours (PM10, t+24 h) act as the dependent variable and the others as independent variables.

Table 1. The air pollutants and meteorological variables

Variable	Role of variable
PM_{10} concentrations for the next 24 hours ($PM_{10, t+24h}$)	Dependent
Particulate Matter with an aerodynamic diameter of less than or equal to 10 μg (PM_{10})	Independent
Sulphur Dioxide (SO_2)	Independent
Nitric Oxide and Nitrogen Dioxide (NO_x)	Independent
Nitrogen Dioxide (NO_2)	Independent
Ozone (O_3)	Independent
Carbon Monoxide (CO)	Independent
Wind Speed (WS)	Independent
Wind Direction (WD)	Independent
Relative Humidity (RH)	Independent
Temperature (T)	Independent

Data Pre-processing

In this study, data preprocessing involves data imputation, data transformation, imbalanced data treatment, and data partition. Missing value in a dataset is a problem typically encountered by researchers in environmental studies. The unavailability of any data restrains the capability to accurately conclude or interpret observations [35]. The missing data must be handled because full data is essential to perform statistical analysis. This study employed linear interpolation for dealing with missing data. According to [36], this linear interpolation technique estimates the missing air pollution data better than other methods.

Several data transformation steps were performed to ensure consistency in units. These include unit conversion of gas pollutant variables: SO_2, NO_2, NO_x, O_3, and CO that were originally recorded in parts per million (ppm), an impractically small unit. These variables were converted to parts per billion to maintain unit consistency across the dataset. Meanwhile, the wind direction (WD) variable, initially expressed in degrees, was transformed into a dimensionless wind direction index to support model interpretation [37].

For data partitioning in this study, a splitting method was implemented, where the model performance is effective when the dataset allocates 80% for model development (training set) and 20% for performance evaluation (testing set) [38].

Imbalance Data Treatment

Resampling strategies are the most common way to deal with an imbalanced dataset [39]. In this study, the Moving Block Bootstrapping (MBB) approach will be applied to solve the

extreme events in imbalanced regression problems. In case of a lack of experience in econometric model specification, block bootstrapping is proposed as one of the most widely used bootstrap methods in the domain of time series.

This study addresses the obstacles of imbalanced regression in predicting extreme PM_{10} concentrations. As outlined by [40], the Air Pollution Index (API), presented in Table 2, classifies air quality as good, moderate, unhealthy, very unhealthy, and hazardous, which can serve as a reference framework for interpreting air quality and guiding management decisions. API is recognized as a straightforward and comprehensive method for explaining air quality situations that are easily understood by the public [41]. The breakpoint concentration is provided for PM_{10} associated with each API category. An API value within the range 101-200 µg/m³ indicates that the air quality is unhealthy. This unhealthy API is parallel to the PM_{10} breakpoint concentration, which is 155 µg/m³. Therefore, in this study, a PM_{10} concentration that is greater than or equal to 155 µg/m³ is designated as the extreme threshold to represent high particulate pollution events or extreme events. Meanwhile, the PM_{10} breakpoint concentration that is below 155 µg/m³ is considered a normal event. Defining a clear extreme threshold is an important strategy that allows the MBB procedure to place greater emphasis on capturing these high particulate events.

Table 2. Breakpoint of PM_{10} concentration

API	Air Quality Status	Breakpoint of concentration	Relevance Threshold
0-50	Good	0<y<54	Normal events
51-100	Moderate	55≤y<155	PM_{10} <155 µg/m³
101-150	Unhealthy	155≤y<255	
151-200	Unhealthy	255≤y<355	
201-300	Very unhealthy	355≤y<454	Extreme events
301-400	Hazardous	455≤y<554	
401-500	Hazardous	555≤y<654	PM_{10}≥155 µg/m³

Moving Block Bootstrapping (MBB)

The MBB is a resampling technique to evaluate statistical estimates when faced with time series data [42]. The MBB conducts the resampling mechanism exclusively from rows of pre-formed overlapping blocks. Unlike the conventional bootstrap, MBB differs in that the data is resampled in consecutive blocks, rather than by individual values [43]. This approach ensures that the temporal structure of the dataset is preserved within each block [26].

According to [27], the MBB procedure involves the process presented in figure 2. At the preliminary stage, the original is segmented into overlapping blocks of fixed length. The total number of overlapping blocks is identified by the equation B= n-l+1, where n is the total number of observations in the dataset and l is the length of each block. By using the equation b=n/l, several blocks are randomly selected with replacement from the original set. All resampled blocks are combined to form a new resampled dataset.

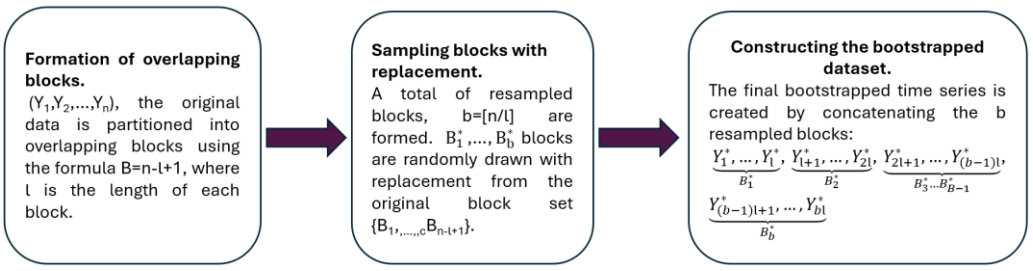


Fig. 2. The MBB procedure

MBB with Ratio-based resampling

This study applied the MBB-Ratio resampling to address the imbalance between extreme and normal PM_{10} concentration data. The MBB method involves dividing the original dataset into overlapping blocks of fixed length. This study employs a 24-hour target, as the prediction pertains to the subsequent 24 hours. Blocks were classified as extreme when they contained at least one PM_{10} concentration of 155 µg/m³ or higher, while those with values below this threshold were categorized as normal blocks. To develop a more balanced distribution for model training, a 1:1 block ratio was randomly selected for the extreme and normal blocks. This is inspired by [44], [45], where they resampled the data by using a 1:1 block ratio to rebalance the distribution. However, this research resampled the data within each respective block to preserve the data dependency within the block. All resampled blocks were combined to form a new resampled dataset. This approach ensures that the newly formed dataset supports more robust model learning and evaluation for imbalanced regression problems. Figure 3 below demonstrates the MBB-Ratio process, where a 1:1 block ratio was randomly selected for extreme and normal events. In this process, the blocks of equal length (for examples in figure 3, two for normal blocks and two for extreme blocks) are independently selected. By enforcing an equal number of extreme and normal blocks, the resampled dataset mitigates the imbalance issue and increases the observation of rare events during training. Unlike previous resampling that modifies data at each observation level, the block-based strategy preserves the temporal dependence of PM_{10} concentrations within each block. This enables the model to better learn patterns of both extreme and normal events in the air pollution dataset.

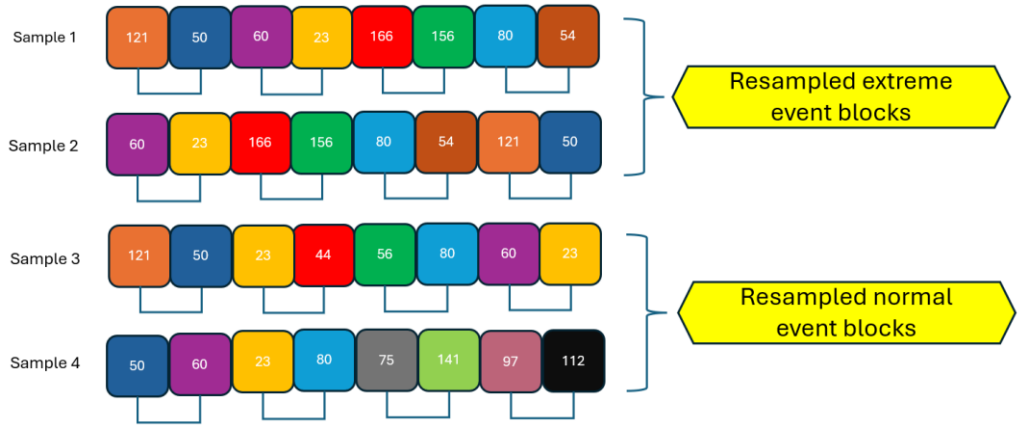


Fig. 3. MBB-Ratio Process

Model Development

Extreme Gradient Boosting (XGBoost)

XGBoost is a high-performing and popular machine learning model that is broadly used for classification and regression cases [46], [47]. It follows the concept of gradient boosting, where models are built sequentially. The model starts with an initial prediction. The first tree is constructed to correct the largest prediction error from this point, and each new tree aims to reduce the error of the previous one. In this way, each new tree focuses on reducing the predictions made by the previous tree. XGBoost excels because it is fast, accurate, and efficient. It includes advanced functions such as a regularization term to prevent overfitting problems [48]. Due to its efficiency and accuracy, XGBoost is frequently applied in real-world applications, specifically in

air pollution environmental modelling [31], [32], [49], [50], [51]. XGBoost aims to minimize the following regularized objective function [52]:

$$\text{Obj} = L(\text{ø}) = \sum_i L(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \quad (1)$$

where: L is the loss function that measures the difference between $\hat{y}_I$ and $y_{i;}$ $\hat{y}_I$ is the prediction value; y_i is the actual value; $\Omega(f_k)$ is a regularization term that controls the complexity of each tree f_k, thereby helping to prevent overfitting [53].

Model Performance

The models were compared based on the model's error by using accuracy indicators, namely Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Square Error (RMSE). The best model is determined when it has high accuracy, in which the performance indicator is closer to zero [54]. Equations (2)-(4) show the performance indicators' formulas used in this study:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} \left| \hat{Y}_i - Y_i \right| \quad (2)$$

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \frac{|Y_i - \hat{Y}_i|}{Y_I} \quad (3)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\hat{Y}_i - Y_i \right)^2} \quad (4)$$

where: n is the total number of observations; $\hat{Y}_i$ is the predicted value of $PM_{10,\,t+24h;}$ Y_i is the actual value of $PM_{10,t+24h;}$ $\overline{Y}_i$ is the mean of the actual value of $PM_{10,\,t+24h}$

Results and discussion

Descriptive Statistics Summary for PM_{10} Concentration with and without MBB-Ratio resampling approach

Table 3 presents the descriptive statistics for PM_{10} concentrations in Shah Alam from 2013 to 2022. The table compares the original dataset without resampling and the resampled dataset with the MBB-Ratio resampling approach. In the original dataset (without resampling), the number of PM_{10} concentration observations is N=82431. However, when the MBB-Ratio was applied, the number of PM_{10} concentration observations increased to ($N_{MBB\text{-}Ratio}$ =137280). The application of MBB-Ratio notably boosted both normal and extreme events, when the total number of normal events for PM_{10} concentration observations increased from 81499 to 115044, while the total number of extreme events increased from 932 to 22236. Additionally, the mean, median, and standard deviation of PM_{10} concentration increase after the application of the MBB-Ratio approach to the imbalanced air pollution data. The mean increased from 41.7078 µg/m³ to 88.92 µg/m³; the median rose from 35 µg/m³ to 61.56 µg/m³; and the standard deviation expanded from 31.7368 µg/m³ to 77.65 µg/m³. These increments show that the higher frequency of extreme PM_{10} values was introduced through the MBB-Ratio approach.

The PM_{10} concentration's skewness was evaluated to evaluate the distribution changes and the presence of extreme values before and after the application of the resampling approach, MBB-Ratio. Before resampling, the skewness of PM_{10} concentrations was 4.574, indicating a strongly right-skewed distribution. However, after the application of the MBB-Ratio, the skewness reduced to 1.91, indicating a more balanced distribution when extreme events representation is increased. A clearer distribution can be seen from figure 4, which illustrates the

distribution of PM_{10} without and after applying the MBB-Ratio resampling techniques. Figure 4 demonstrates the PM_{10} without resampling (blue line) that exhibits a sharp peak with lower concentrations and a thin right tail. This reflects a highly right-skewed distribution and relatively few extreme cases for the PM_{10} without resampling. However, after resampling with MBB-Ratio (red line), the distribution of PM_{10} remains right-skewed, but the peak flattens with a broader distribution, extending into higher concentrations of PM_{10} values up to 500 µg/m³. This visual change shows that the MBB-Ratio approach added more PM_{10} extreme concentrations, leading to a reduced skewness and more balanced distribution in the air pollution dataset. This pattern depicts that resampling with the MBB-Ratio improves the number of extreme air pollution events in the dataset, delivering a more balanced distribution for a better further analysis.

Table 3. Descriptive Statistics for PM_{10} concentrations in Shah Alam with and without the MBB-Ratio resampling approach

	Without resampling	**Resampling with MBB-Ratio**
Number of PM_{10} concentration observations in the air pollution dataset	82431	137280
Number of normal events of PM_{10} concentration observations	81499	115044
Number of extreme events of PM_{10} concentration observations	932	22236
Mean	41.7078	88.920
Median	35.000	61.560
Standard deviation	31.7368	77.650
Variance	1007.226	6029.960
Skewness	4.574	1.910
Minimum	0.630	0.630
Maximum	575.000	575.000

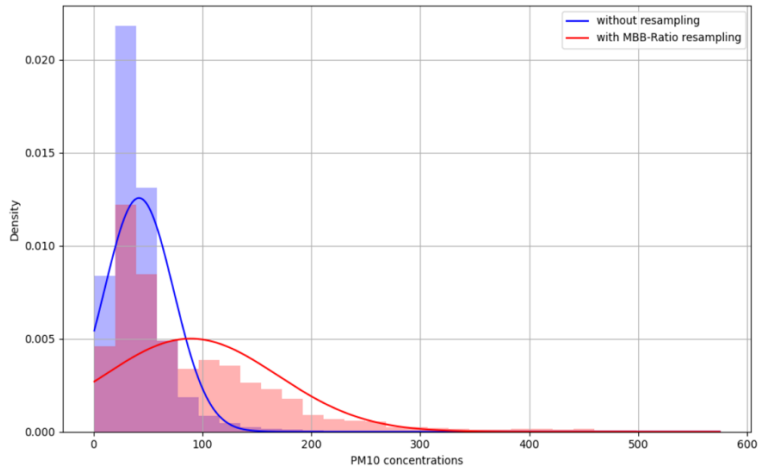


Fig. 4. Distribution plot with and without the MBB-RW resampling strategy

Figure 5 presents a correlation heatmap in Shah Alam. The coefficient of correlation represents the strength and direction of the relationship between the variables. The heatmap uses a color gradient ranging from blue (indicating a strong negative correlation (r = -1)) to red (indicating a strong positive correlation, +1), with lighter shades representing weaker correlations. It is found that most of the variables demonstrate positive correlation with PM_{10}, suggesting a positive relationship between the variables and PM_{10} concentrations. However, an exception is observed in the case of Relative Humidity, where the correlation coefficients show a negative relationship between RH and PM_{10} (r = -0.03).

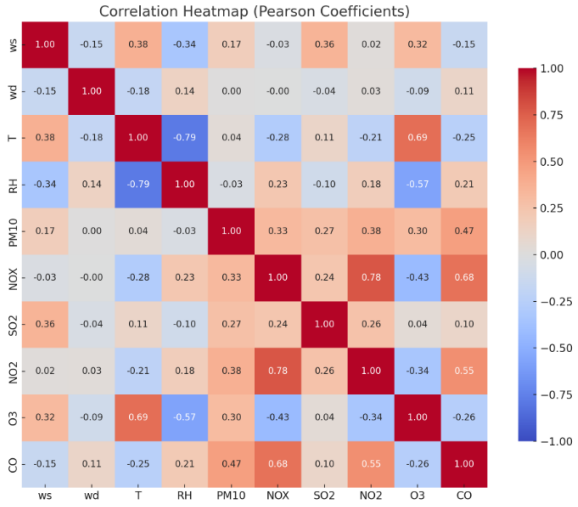


Fig. 5. Air pollution correlation heatmap in Shah Alam

The negative correlation implies that when the humidity level increases, the PM_{10} is prone to decrease slightly [1]. All the parameters have moderate to weak correlations with PM_{10} concentration. The correlation of PM_{10} with CO is r = 0.47, indicating the highest correlation, while WD does not correlate with PM_{10} (r = 0). In air pollution research, PM_{10} concentrations are influenced by multiple contributing factors. Consequently, the correlation between each variable and PM_{10} concentration is typically weak (references). It is important to understand these correlations in air quality analysis, especially when the motivation of the study is to develop a model and predict extreme pollution events.

The performance of the Extreme Gradient Boosting (XGBoost) Machine Learning (ML) model

Table 4 shows the performance comparison of the ML model, named Extreme Gradient Boosting (XGBoost), evaluated for the overall, normal, and extreme events with and without the MBB-Ratio resampling approach for hourly PM_{10} concentration predictions in the Shah Alam station from the period of 2013 to 2022.

Table 4. Performance comparison of XGBoost with and without MBB-Ratio resampling approach

		Without resampling	**With MBB-Ratio resampling approach**
Overall dataset	MAE	14.2396	12.7727
	MAPE	0.4815	0.3168
	RMSE	19.8421	17.5100
	N	16487	27446
Normal events	MAE	13.4048	12.5953
	MAPE	0.4830	0.3652
	RMSE	18.2636	17.2294
	N	16295	23037
Extreme events	MAE	85.1041	13.6975
	MAPE	0.3634	0.0644
	RMSE	108.3010	18.9009
	N	192	4419

The performance indicators, Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE), are provided for the overall, normal, and extreme events, together with the total number of observations (N) in each event's section.

The results present a clear disparity in modeling performance across different levels. For the overall event, the performance of the overall air pollution event without the resampling approach shows that the MAE, MAPE, and RMSE values are 14.2396, 0.4815, and 19.8421, respectively. When performance is split by normal and extreme event type, the model performs reasonably well during normal events, with a reduced MAE, MAPE, and RMSE of 13.4048, 0.4830, and 18.2636. However, when it comes to extreme pollution events, the performance drops to 108.3010, 85.1041, and 0.3634. These sharp error discrepancies show that the models can predict well for normal pollution events when the learning algorithm focuses on the most frequent cases but exhibit poor predictive accuracy on the extreme events [32].

The poor performance for extreme scenarios can be attributed to the highly imbalanced distribution. Only 192 observations from 16487 total observations belong to extreme events, while the normal events account for 16295 observations. This major imbalance in distribution causes the model to poorly predict extreme events, since the model will be biased toward the normal cases during the model's training [10], [55], [56]. The result discussed above indicates that the model outperforms in predicting normal events but fails to cater for the extreme pollution events. This shows that applying a suitable resampling strategy is important to improve model robustness and accuracy, especially for underrepresented extreme cases.

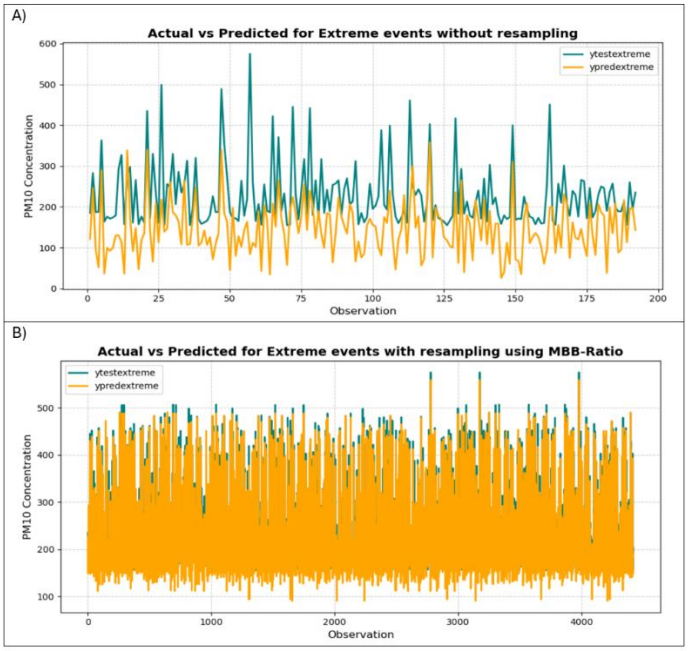
Based on the previous evaluation, it is important to apply the MBB with a 1:1 ratio-based approach in tackling the data imbalance issue, especially by increasing the representation of extreme PM_{10} observations. The XGBoost model's performance on the overall pollution dataset shows an improvement when the MAE, MAPE, and RMSE dropped from 14.2396 to 12.7277, 0.4815 to 0.3168, and 19.8421 to 17.51 after applying the MBB-Ratio approach. These results demonstrate that MBB-Ratio is effective for the overall air pollution dataset when it reflects a slight but stable improvement in the model's ability to predict common pollution levels after resampling. For normal events, the performance of MBB-Ratio is approximately similar to the overall dataset, and the model performed slightly better. The performance of normal events increases when the MAE decreases from 13.40 to 12.5953, the MAPE drops from 0.4830 to 0.3652, and the RMSE is reduced from 18.26 to 17.2294. This suggests a consistent prediction for normal pollution events.

The most notable enhancement was observed in the predictive performance of extreme events. After the utilization of the MBB-Ratio, the MAE decreased from 85.1041 to 13.6975, the MAPE reduced from 0.3634 to 0.0644, and the RMSE dropped dramatically from 108.3010 to 18.9009. These improvements can be attributed to the increased representations of extreme events in the dataset when resampled with the MBB-Ratio. The number of extreme cases increased from 192 to 4419, exposing more extreme events for the model's learning. Integrating the resampled data into the XGBoost mechanism helps boost the model's capability to learn from both normal and extreme pollution cases, as the MBB-Ratio method applies a balanced 1:1 ratio, eventually enhancing its performance in forecasting high PM_{10} episodes with greater accuracy.

The performance of extreme event prediction can also be visualized in Figure 6, which presents a performance comparison of the XGBoost model based on the actual vs. predicted values before and after MBB-Ratio implementation.

The figure illustrates the effect of resampling on the model's ability to predict extreme PM_{10} concentrations ($\geq$ 155 $\mu g/m^3$), which correspond to unhealthy air quality levels under the Malaysian Air Pollution Index (API). Figure 6 presents model A, which represents the actual vs. predicted values of extreme events without the resampling approach. Meanwhile, Model B presents actual vs. predicted for extreme events with the MBB-Ratio resampling approach. Referring to plot form Model A, the predicted values consistently underestimate the actual PM_{10}

concentrations for extreme events, reflecting the model's bias due to data imbalance. In contrast, the plot for Model B demonstrates a much closer alignment between actual and predicted values.



Note: Model A is a model without the MBB-Ratio resampling approach
Model B is a model with the MBB-Ratio resampling approach.

Fig. 6. Performance comparison of the XGBoost model based on the actual vs predicted values before and after MBB-Ratio implementation

This approach artificially balances the dataset by increasing the representations of extreme blocks. As a result, the model has the ability to produce the peaks by increasing the representations of the true PM_{10} concentrations, thus mitigating the imbalance problems, thereby improving the prediction accuracy for extreme pollution events, which leads to more high-impact air quality detection and conservation-oriented management during high particulate pollution episodes.

Conclusion

This research emphasizes attention to the natural difficulty in modeling, estimating, and predicting extreme air pollution events in imbalanced datasets. Preliminary results without the resampling approach showed that the model continuously underestimates the extreme PM_{10} concentrations. The model's performance increased after resampling using MBB with a ratio-based 1:1 by better representing the extreme events in the training data, when the error of RMSE, MAE, and MAPE shows a reduction from 108.3010 to 18.9009, 85.1041 to 13.6975, and 0.3634 to 0.0644. Visual comparisons demonstrated notable improvements in peak alignment, variability, and magnitude accuracy of predictions. These improvements in predictive performance suggest the capability of the MBB-Ratio in addressing data imbalance and enhancing model prediction towards extreme events. From a conservation science perspective, improved prediction of high PM_{10} episodes is crucial for predicting periods of highly polluted air that can affect the ecosystem, degrade the air quality, and contribute to environmental degradation. Overall, enhancing resampling methods in the air pollution field is not only

important for improving the air quality forecasting models but also contributes to sustainable environmental management by supporting early warning systems and public health decision-making. In turn, this encourages more proactive conservation actions, such as temporary industrial control, traffic control, or public health warnings.

Acknowledgments

This research was funded by the GIP Research Grant, Research Management Institute, Universiti Teknologi MARA, Malaysia (600-RMC/GIP 5/3 (032/2024)). We would like to express our sincere thanks to the Faculty of Computer and Mathematical Sciences, Universiti Teknologi Mara (UiTM), for their unwavering support throughout this study. Additionally, we also extend our heartfelt appreciation to the Department of Environment (DOE) Malaysia for supplying air quality data that made this research possible.

References

[1] N. A. A. A. Rahim, N. M. Noor, I. A. M. Jafri, N. Ramli, M. A. Kamaruddin, and G. Deák, "Predicting Particulate Matter (PM10) during High Particulate Event (HPE) using Quantile Regression in Klang Valley, Malaysia," in *IOP Conference Series: Earth and Environmental Science*, Institute of Physics, 2023. DOI: 10.1088/1755-1315/1216/1/012003.

[2] S. Gulati, A. Bansal, A. Pal, N. Mittal, A. Sharma, and F. Gared, "Estimating PM2.5 utilizing multiple linear regression and ANN techniques," *Science Reports*, vol. 13, no. 1, 2023. DOI: 10.1038/s41598-023-49717-7.

[3] S. N. Redzuan *et al.*, "Characteristics of PM10 Level during Haze Events in Malaysia Based on Quantile Regression Method," *Atmosphere (Basel)*, vol. 14, no. 2, 2023. DOI: 10.3390/atmos14020407.

[4] S. U. Park, J. H. Cho, and M. S. Park, "Identification of Visibility Reducing Weather Phenomena Due to Aerosols," *Environmental Management and Sustainable Development*, vol. 2, no. 1, pp. 126–142, 2013. DOI: 10.5296/emsd.v2i1.3628.

[5] I. A. Mohd Jafri, N. Mohamed Noor, N. A. A. A Rahim, A. Z. Ul Saufie, and G. D. Habil, "Prediction of Particulate Matter (PM10) during High Particulate Event in Peninsular Malaysia using Novel Hybrid Model," *E3S Web of Conferences, EDP Sciences*, 2023. DOI: 10.1051/e3sconf/202343701001.

[6] S. W. H. M. Zulkifli, H. B. Samsudin, and N. Majid, "Association between PM10 and respiratory diseases admission in Peninsular Malaysia during the haze," DOI:10.21203/rs.3.rs-3037064/v1.

[7] M. Muhammad, A. Z. Ul-Saufie, N. F. A. Radi, N. M. Noor, and A. Gusnanto, "Modified resampling strategy for extreme values in imbalanced air pollution data using moving block bootstrapping approach with relevance weighting (MBB-RW)," *Science Reports*, vol. 16, no. 1, 2026. DOI: 10.1038/s41598-025-28416-5.

[8] J. G. Avelino, G. D. C. Cavalcanti, and R. M. O. Cruz, "Resampling strategies for imbalanced regression: a survey and empirical analysis," *Artificial Intelligence Review*, vol. 57, no. 4, art. no. 82, 2024. DOI: 10.1007/s10462-024-10724-3

[9] S. B. Belhaouari, A. Islam, K. Kassoul, A. Al-Fuqaha, and A. Bouzerdoum, "Oversampling Techniques for Imbalanced Data in Regression," DOI: 10.2139/ssrn.4577876.

[10] R. P. Ribeiro and N. Moniz, “Imbalanced regression and extreme value prediction,” *Machine Learning (Springer Nature)*, vol. 109, pp. 1803–1835,2020. DOI: 10.1007/s10994-020-05900-9.

[11] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, “Review of Classification Methods on Unbalanced Data Sets,” *IEEE Access*, vol. 9, pp. 64606–64628, 2021. DOI: 10.1109/ACCESS.2021.3074243.

[12] X. Liu and H. Tian, “Research on Imbalanced Data Regression Based on Confrontation,” *Processes*, vol. 12, no. 2, 2024, DOI: 10.3390/pr12020375.

[13] P. Branco *et al.*, “REBAGG: Resampled Bagging for Imbalanced Regression,” *Proc. Mach. Learn. Res.*, vol. 94, pp. 67–81, 2018, Accessed: Apr. 27, 2026. [Online]. Available: https://www.researchgate.net/publication/328065440_REBAGG_REsampled_BAGGing_for_Imbalanced_Regression

[14] N. Moniz, R. Ribeiro, V. Cerqueira, and N. Chawla, “SMOTEBoost for regression: Improving the prediction of extreme values,” in *Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018*, Institute of Electrical and Electronics Engineers Inc., pp. 150–159, 2018. DOI: 10.1109/DSAA.2018.00025.

[15] A. Silva, R. P. Ribeiro, and N. Moniz, “Model Optimization in Imbalanced Regression,” in *International Conference on Discovery Science*, 2022. DOI: https://doi.org/10.48550/arXiv.2206.09991.

[16] L. Torgo and R. Ribeiro, "Utility-Based Regression," *Lecture Notes in Computer Science*, pp. 597-604, DOI: 10.1007/978-3-540-74976-9_63.

[17] E. A. Felix and S. P. Lee, "Systematic literature review of preprocessing techniques for imbalanced data," *IET Software*, vol. 13, no. 6, pp. 479-496, 2019. DOI: 10.1049/iet-sen.2018.5193.

[18] L. Torgo and R. Ribeiro, “Predicting rare extreme values,” in *Conference: Advances in Knowledge Discovery and Data Mining*, pp. 816–820, 2006. DOI: 10.1007/11731139_95.

[19] P. Branco, R. P. Ribeiro, L. Torgo, B. Krawczyk, and N. Moniz, “SMOGN: a Pre-processing Approach for Imbalanced Regression,” *Proc. Mach. Learn. Res.*, vol. 74, pp. 36–50, 2017. Accessed: Apr. 27, 2026. [Online]. Available: https://proceedings.mlr.press/v74/branco17a.html

[20] 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, 2002. DOI: https://doi.org/10.1613/jair.953.

[21] J. Fonseca and F. Bacao, "Geometric SMOTE for imbalanced datasets with nominal and continuous features," *Expert Systems with Applications*, vol. 234, Article Number: 121053, 2023. DOI: 10.1016/j.eswa.2023.121053.

[22] L. Torgo, R. P. Ribeiro, B. Pfahringer, and P. Branco, “SMOTE for regression,” *Portuguese Conference on Artificial Intelligence*, pp. 378–389, 2013. DOI: 10.1007/978-3-642-40669-0_33.

[23] H. He and E. A. Garcia, “Learning from imbalanced data,” *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, 2009, pp. 1263–1284. DOI:10.1109/TKDE.2008.239.

[24] M. Steininger, K. Kobs, P. Davidson, A. Krause, and A. Hotho, “Density-based weighting for imbalanced regression,” *Machine Learning (Springer Nature)*, vol. 110, no. 8, pp. 2187–2211, 2021. DOI:10.1007/s10994-021-06023-5.

[25] Y. Yang, K. Zha, Y.-C. Chen, H. Wang, and D. Katabi, “Delving into Deep Imbalanced Regression,” *International Conference on Machine Learning*, 2021. https://doi.org/10.48550/arXiv.2102.09554.

[26] B. Radovanov and A. Marcikic, “A comparison of four different block bootstrap methods,” *Croatian Operational Research Review*, vol. 5, no. 2, pp. 189–202, 2014. DOI:10.17535/crorr.2014.0007.

[27] T. A. Kuffner, S. M. S. Lee, and G. A. Young, “Block bootstrap optimality and empirical block selection for sample quantiles with dependent data,” *Biometrika*, vol. 103, no. 1, pp. 1–18, 2018. DOI: 10.1093/biomet/asaa075.

[28] S. N. Z. A. Burhanuddin, S. M. Deni, and N. Shaadan, "Controlled Sampling Approach in Improving Multiple Imputation for Missing Seasonal Rainfall Data," DOI: 10.21203/rs.3.rs-679692/v1.

[29] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, “Review of Classification Methods on Unbalanced Data Sets,” *IEEE Access*, vol. 9, pp. 64606–64628, 2021. DOI: 10.1109/ACCESS.2021.3074243.

[30] I. Ayus, N. Natarajan, and D. Gupta, “Comparison of machine learning and deep learning techniques for the prediction of air pollution: a case study from China,” *Asian Journal of Atmospheric Environment*, vol. 17, no. 1, Dec. 2023. DOI: 10.1007/s44273-023-00005-w.

[31] T. H. Dao *et al.*, “Analysis and Prediction for Air Quality Using Various Machine Learning Models,” *Proceedings of the Seventh International Conference on Research in Intelligent and Computing in Engineering*, PTI, pp. 89–94, 2023. DOI:10.15439/2022r03.

[32] M. Muhammad, A. Zia Ul-Saufie, and A. Radi, “Evaluating the Performance of Tree-Based Models in Predicting Haze Events in Malaysia,” *IJACSA International Journal of Advanced Computer Science and Applications*, vol. 16, no. 4, pp. 1127–1137, 2025. DOI: https://dx.doi.org/10.14569/IJACSA.2025.01604107.

[33] N. Amila *et al.*, “Modeling of Particulate Matter (PM10) during high particulate event (HPE) in Klang Valley, Malaysia,” *International Journal of Conservation Science*, vol. 13, pp. 1065–107, 2022. DOI: 10.24818/ijcs.2022.3.14.

[34] Z. A. Rais, N. Ramli, N. M. Noor, H. A. Hamid, A. Z. Ul-Saufie, And M. K. N. Mahmad, "Analysis of Air Pollution in Malaysia: Implications for Environmental Conservation using Granger Causality and Pearson Correlation," *International Journal of Conservation Science*, vol. 16, no. 1, pp. 149-164, 2025. DOI: 10.36868/ijcs.2025.01.09.

[35] Z. Libasin, A. Zia Ul-Saufie, and A. Hasfazilah, “Identifying missing data mechanisms among incomplete air pollution datasets in Malaysia,” *Conference for Environmental Integration (EMCEI-3)*, pp. 77–79, 2024. https://doi.org/10.1007/978-3-031-43922-3_18.

[36] N. M. Noor, M. M. Al Bakri Abdullah, A. S. Yahaya, and N. A. Ramli, "Comparison of Linear Interpolation Method and Mean Method to Replace the Missing Values in Environmental Data Set," *Materials Science Forum*, vol. 803, pp. 278-281, 2014. DOI: 10.4028/www.scientific.net/msf.803.278.

[37] A. Kumar and P. Goyal, "Forecasting of air quality in Delhi using principal component regression technique," *Atmospheric Pollution Research*, vol. 2, no. 4, pp. 436-444, 2011. DOI: 10.5094/apr.2011.050.

[38] 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," *Journal of Environmental and Public Health*, vol. 2023, pp. 1-26, 2023. DOI: 10.1155/2023/4916267.

[39] N. Moniz, P. Branco, L. Torgo, and B. Krawczyk, “Evaluation of Ensemble Methods in Imbalanced Regression Tasks,” *Proc. Mach. Learn. Res.*, vol. 74, pp. 129–140, 2017. [Accessed: Apr. 27, 2026]. [Online]. Available: https://proceedings.mlr.press/v74/moniz17a.html

[40] DOE, “Air Pollutant Index (API) Calculation,” 2024. [Accessed: Feb. 21, 2026]. [Online]. Available: https://www.doe.gov.my/wp-content/uploads/2021/09/API_Calculation.pdf

[41] N. L. Abd Rani, A. Azid, S. I. Khalit, H. Juahir, and M. S. Samsudin, "Air Pollution Index Trend Analysis in Malaysia, 2010-15," *Polish Journal of Environmental Studies*, vol. 27, no. 2, pp. 801-807, 2018. DOI: 10.15244/pjoes/75964.

[42] S. Mignani and R. Rosa, "The moving block bootstrap to assess the accuracy of statistical estimates in Ising model simulations," *Computer Physics Communications*, vol. 92, no. 2-3, pp. 203-213, 1995. DOI:10.1016/0010-4655(95)00114-7.

[43] Ł. Sroka, “Applying block bootstrap methods in silver prices forecasting,” *Econometrics*, vol. 26, no. 2, pp. 15–29, 2022. DOI: 10.15611/eada.2022.2.02.

[44] K. Omari, C. Taoussi, A. Oukhatar, U. Sultan Moulay Slimane, and B. Mellal, “Comparative Analysis of Undersampling, Oversampling, and SMOTE Techniques for Addressing Class Imbalance in Phishing Website Detection,” *IJACSA International Journal of Advanced Computer Science and Applications*, vol. 16, no. 2, pp. 751–757, 2025. https://dx.doi.org/10.14569/IJACSA.2025.0160276.

[45] P. Lee, "Resampling Methods Improve the Predictive Power of Modeling in Class-Imbalanced Datasets," *International Journal of Environmental Research and Public Health*, vol. 11, no. 9, pp. 9776-9789, 2014. DOI:10.3390/ijerph110909776.

[46] Z. Arif Ali, Z. H. Abduljabbar, H. A. Tahir, A. Bibo Sallow, and S. M. Almufti, “Extreme Gradient Boosting Algorithm with Machine Learning: A Review,” *Academic Journal of Nawroz University*, vol. 12, no. 2, pp. 320–334, 2023. DOI: 10.25007/ajnu.v12n2a1612.

[47] N. M. Shahani, X. Zheng, C. Liu, F. U. Hassan, and P. Li, "Developing an XGBoost Regression Model for Predicting Young’s Modulus of Intact Sedimentary Rocks for the Stability of Surface and Subsurface Structures," *Frontiers in Earth Science*, vol. 9, Article Number: 761990, 2021. DOI:10.3389/feart.2021.761990.

[48] G. Martínez-Munoz, C. Bentejac, and A. Csorg O B Gonzalo Martínez-Munoz, “A Comparative Analysis of XGBoost,” 2019. https://doi.org/10.48550/arXiv.1911.01914.

[49] H. Jing and Y. Wang, "Research on Urban Air Quality Prediction Based on Ensemble Learning of XGBoost," *E3S Web of Conferences*, vol. 165, Article Number: 02014, 2020. DOI: 10.1051/e3sconf/202016502014.

[50] K. Kumar and B. P. Pande, “Air pollution prediction with machine learning: a case study of Indian cities,” *International Journal of Environmental Science and Technology*, vol. 20, no. 5, pp. 5333–5348, 2023. DOI: 10.1007/s13762-022-04241-5.

[51] T. Nguyen, S. El Outayek, S. H. Lim, and V. Nguyen, "A systematic approach to selecting the best probability models for annual maximum rainfalls – A case study using data in Ontario (Canada)," *Journal of Hydrology*, vol. 553, 2017, pp. 49-58. doi: 10.1016/j.jhydrol.2017.07.052.

[52] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” pp. 785–794, 2016. DOI: 10.1145/2939672.2939785.

[53] B. Pan, “Application of XGBoost algorithm in hourly PM2.5 concentration prediction,” in *IOP Conference Series: Earth and Environmental Science*, Institute of Physics Publishing, 2018. DOI: 10.1088/1755-1315/113/1/012127.

[54] S. Y. Fong, S. Abdullah, and M. Ismail, “Forecasting of Particulate Matter (PM10) concentration based on gaseous pollutants and meteorological factors for different monsoons of urban coastal areas in Terengganu,” *Journal of Sustainability Science and Management Special Issue Number*, no. 5, pp. 3–17, 2018.

[55] P. Branco, L. Torgo, and R. P. Ribeiro, "Pre-processing approaches for imbalanced distributions in regression," *Neurocomputing*, vol. 343, pp. 76–99, 2019. DOI: 10.1016/j.neucom.2018.11.100.

[56] J. Ren, M. Zhang, C. Yu, and Z. Liu, "Balanced MSE for Imbalanced Visual Regression," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, pp. 7916–7925, 2022. DOI: 10.1109/CVPR52688.2022.00777.

Received: October 02, 2025
Accepted: April 10, 2026