scientific reports

Check for updates

OPEN

Modified resampling strategy for extreme values in imbalanced air pollution data using moving block bootstrapping approach with relevance weighting (MBB-RW)

Mahiran Muhammad[1], Ahmad Zia Ul-Saufie[1]✉, Noor Fadhilah Ahmad Radi[1], Norazian Mohamed Noor[2] & Arief Gusnanto[3]

Air pollution datasets typically exhibit a right-skewed distribution. These conditions are caused by the presence of extreme events leading to imbalanced data distribution. This imbalanced regression poses a notable challenge in predictive modeling, as the models tend to be biased towards frequent normal events while underperforming extreme events. Therefore, to address this issue, the resampling approach is crucial in handling these extreme events to improve model performance. In this study, a modified resampling strategy, called Moving Block Bootstrapping with Relevance Weighting (MBB-RW), is proposed to address imbalanced regression problems. By integrating MBB with relevance weighting, the time-series dependence of air pollution data is preserved while placing greater emphasis on extreme events. The findings demonstrate that MBB-RW can mitigate data imbalance and enhance model prediction accuracy for extreme events. These enhancements are evident in the performance of the Extreme Gradient Boosting (XGBoost) model in predicting PM_{10}, where the evaluation metrics showed significant reductions after applying MBB-RW: RMSE dropped from 108.3010 to 39.1846 (63.8188%) and MAE from 85.1041 to 27.1082 (68.14700%). The key contribution of this study is the development of MBB-RW resampling strategy designed to improve extreme values in an imbalanced regression dataset, with our focus on the air pollution dataset. Simultaneously, this method can be implemented to enhance the accuracy of PM_{10} concentration predictions, specifically during extreme air pollution events.

Keywords Particulate matter 10 (PM_{10}), Extreme events, Imbalanced data, Moving block bootstrapping (MBB), Moving block bootstrapping with relevance weighting (MBB-RW), Extreme gradient boosting (XGBoost)

Air pollution has become a severe environmental and public health issue, particularly in urban and industrial regions. Major causes include vehicle emissions, industrial activity, open burning, and transboundary haze, which can sometimes be triggered by forest fires in neighbouring countries[1]. Primary pollutants that impact most countries include Particulate Matter (PM), Nitrogen Dioxide (NO2), Carbon Monoxide (CO), Ozone (O3), and Sulphur Dioxide[2]. However, Particulate Matter 10 (PM_{10}) is a major pollutant in the air, and it has a larger impact on humans than other pollutants. In Malaysia, PM_{10} concentration is always higher than any other pollutant[3]. Therefore, the urge to predict the PM_{10} concentration, especially in extreme events, has become a crucial and tough task with growing motor and industrial developments[4,5].

Air pollution poses a challenge in making accurate predictions, requiring robust modeling techniques that can capture both common and extreme events. One of the major challenges in modeling air pollution data is the imbalance in the target distribution[6]. An abnormally high PM_{10} concentration level can result in the existence

[1]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam 40450, Selangor, Malaysia. [2]Faculty of Civil Engineering & Technology, Universiti Malaysia Perlis, Jejawi, Perlis 02600, Arau, Malaysia. [3]Department of Statistics, University of Leeds, Woodhouse Lane, Leeds, West Yorkshire LS2 9JT, UK. ✉email: ahmadzia101@uitm.edu.my

of extreme values in air pollution data[7]. Extreme values in air pollution result in imbalance problems, described by the combination of two factors: (i) skewed distribution of the target variable, and ii) user preferences towards underrepresented cases[8].

Concentrations of pollutants are often right-skewed. Most measurements are very modest; however, a small proportion of observations reach quite high levels during extreme events. According to[9], extreme values are defined as events that occur significantly less often than common events. Extreme values or extreme events are normally generated by haze from forest fires, industrial hazards, thermal inversion, and sometimes caused by monitoring faults[10]. Neglecting extreme values can lead to inaccurate predictions when using the original data set directly due to its skewed distribution[11]. The effect of imbalance and skewness is evident. Models trained on imbalanced data typically follow normal cases, as minimizing overall error inherently skews the algorithm toward the majority class. Hence, predictions for extreme pollution events are often inaccurate, with models struggling to cater for rare but important new environmental and social events[12,13]. Therefore, it is fundamental to correctly deal with the problem of extreme values in imbalanced distribution to boost the performance of prediction models, which has become a crucial topic in current research.

Imbalanced regression

In real life, data imbalance is an inherent, typical, encountering challenge encountered across various research disciplines. There are two categories of imbalanced datasets: classification and regression. In classification problems, an imbalanced dataset occurs through the presence of an underrepresented class (minority class) compared to another one (majority class)[14]. Meanwhile, in regression problems, the target or dependent is continuous, and indicates a complex definition, because the target observation is not bound to a limited set of discrete values, unlike in classification problems, where the target value represents specific categories or classes[14]. The imbalance regression problem occurs when the frequency of some target values in the regression dataset is extremely low, and these target values are often ignored by the model, resulting in poor prediction performance of the model on these samples[8,15]. The imbalanced regression requires studying the distribution of the data more thoroughly to predict the value more accurately[16].

Most of the existing strategies for dealing with imbalanced data are limited for classification problems, that is, the target value is a discrete index of different categories. Nonetheless, this issue also arises in numerical prediction tasks, for example, regression, when users are more interested in accurately predicting extreme (rare) values of the target variable[16]. In fields like finance, meteorology, or environmental sciences, the goal is often to predict uncommon events, also known as extreme cases[8]. According to [14], to better understand the different approaches for dealing with imbalance regression problems, the strategies are categorized into three main groups: (i) learning process modification, (ii) regression models, and (iii) evaluation metrics. Organizing these strategies into three distinct categories can yield a better understanding of the approaches and support the selection of the most effective strategy for handling imbalanced regression problems. One of the popular strategies that is often used by previous researchers in handling imbalanced regression is learning process modification or also known as data preprocessing, data preparation, or resampling[8,17–20]. The strategies operate by either removing samples from common cases or generating synthetic samples for extreme events. Data preprocessing techniques have the advantage of permitting the use of any learning algorithm simultaneously, without affecting the understandability of the model[17]. Data preprocessing or resampling procedures modify the original data distribution before implementing the learning algorithm[21]. The purpose of modifying the target variable distribution is to drive the learning algorithm to focus on the extreme value cases[22]. These solutions are popular because data preprocessing allows the use of any learning algorithm and are simple to implement because the modification is only on the original data set distribution. However,[12] emphasized that the effectiveness of such methods depends on how the data distribution is adjusted, which remains a challenge for researchers.

Previous studies have investigated several resampling approaches to address imbalanced regression problems. These approaches include Synthetic Minority Oversampling Technique for regression (SMOTER)[13], Synthetic Minority Oversampling Technique with Gaussian Noise (SMOGN)[12], SMOTEBoost[19], Resampled Bagging for Imbalanced Regression (REBAGG)[18], Weighted Relevance-based Combination Strategy (WERCS)[17], Geometric SMOTE[23] and others. SMOTE for regression (SMOTER), SMOTE with Gaussian Noise (SMOGN) and Geometric SMOTE generate synthetic samples through interpolation or noise rejection. Even though these methods boost the depiction of extreme values, these methods may potentially generate noisy and unrealistic data. This may result in misleading patterns that lower the generalisation capacity of machine learning models. Weighted Relevance-based Combination Strategy (WERCS) operates probabilistic oversampling and undersampling based on a relevance function that allocates more weights to extreme values[18]. Though this approach is customizable and maintains the entire target range covered, the data's temporal and spatial correlations are not incorporated. For time-dependent pollution data, this restriction can lead to an unrealistic training set that reduces model performance during sequence prediction.

The Resampled Bagging for Imbalanced Regression (REBAGG) algorithm has been proposed to address the issue of the imbalanced regression task by employing an ensemble method based on bagging and combined resampling techniques[18]. Although the ensemble method boosts robustness and often achieves outstanding performance, it is computationally costly and ignores temporal dependencies. By examining the association between the distribution of the target value and the test error of the prediction model[24], introduced the concept of deep imbalanced regression. According to the properties of imbalanced regression data, they implemented label distribution smoothing (LDS) and feature distribution smoothing (FDS). However, this approach still generates minority regression samples using the data interpolation method, which is prone to overfitting. Introducing synthetic data that lacks real-world representativeness, along with the elimination of common examples that provide important aspects for predictive tasks, may negatively impact the prediction performance.

Therefore, to avoid the synthetic artificial data and preserve the temporal dependencies inherent in the dataset, Moving Block Bootstrapping (MBB) is proposed as a resampling method in handling the imbalanced air pollution dataset. MBB is a well-known resampling technique[25]. MBB is one of the popular methods that originates from Block Bootstrapping. Unlike traditional bootstrapping, which resamples individual observations, MBB resamples blocks of successive data points[26]. This strategy is practicable for time series or spatial data, where maintaining dependencies between observations is vital. MBB can upgrade data quality by preserving data dependency within blocks[27]. Resampling entire blocks of data helps control the relationships between variables, offering the model a more realistic learning context. Furthermore, MBB avoids the addition of artificial data by resampling only from observed values, ensuring data originality while improving the training set with extreme but authentic observations.

Moving block bootstrapping (MBB)

Moving Block Bootstrapping (MBB) was introduced as an adaptation of the ordinary bootstrap for serially correlated data[2829], demonstrated that applying MBB helps to generate more reliable and robust parameter estimates for hydrological models by accounting that the calibration data represent a limited sample drawn from an unknown underlying distribution[26]. compare the standard optimisation model with the block bootstrap optimization models, and they conclude that block bootstrap provides more stable estimates of investment portfolios with a higher rate of diversification. Furthermore, when comparing the forecasts predicted by the ARMA model with the bootstrap forecasts and the actual price relations in the forecasted period, it was observed that each block bootstrap method forecasted the silver futures contract price closer to the actual execution than the ARMA model[30]. The MBB approach demonstrates its strength in preserving the temporal structure of time series data in its process and successfully boosts the estimates of missing rainfall data imputation by utilizing multiple imputation based on the MBB approach associated with normal ratio methods, compared to the conventional bootstrap approach[27]. suggested MBB for environmentalists as an alternative approach for better estimation of environmental datasets.

Although MBB has been demonstrated to be advantageous for time series data, it has its limitations.[31], demonstrated that the MBB resampling strategy can significantly enhance the performance of baseline forecasting models, which yield accurate predictions on normal events but perform poorly on extreme events. This is because the MBB approach puts equal weights on the blocks under the assumption that the blocks have the same underlying process. However, this is not the case when the data contains extreme observations due to temporary events such as a forest fire.

To address this challenge, we propose to improve the MBB approach by acknowledging that the blocks do not necessarily have the same process. This is done by integrating the MBB approach with a relevance weighting threshold (denoted MBB-RW), which allocates higher importance to target the extreme values. Relevance weighting is inspired by the relevance concept from utility-based learning[8,32]. In this approach, the extreme events are better represented during model training, and the inherent temporal structure is simultaneously preserved, resulting in more robust and accurate predictions for extreme cases.

For the development of the PM_{10} prediction model, Extreme Gradient Boosting (XGBoost) will be utilized, as previous studies on air pollution modelling have revealed XGBoost as one of the best machine learning models used for air pollution prediction[33–35]. Specifically, in tree tree-based model,[4] demonstrated that XGBoost is particularly well-suited for predicting extreme pollution values. Our results indicate that this novel approach (MBB-RW) gives a superior performance compared to the standard MBB approach and shows good operating characteristics.

Methodology

Research flow

The overall workflow of this article is shown in Fig. 1, which attempts to develop an effective resampling strategy for dealing with extreme events in imbalanced air pollution data in Shah Alam, Malaysia, using the modified Moving Block bootstrapping with relevance weighting (MBB-RW) approach. This study employs air quality data from 2013 to 2022 provided by the Department of Environment (DOE), Malaysia. The process begins with data retrieval, followed by extensive data preprocessing, with a particular focus on addressing imbalanced data treatment, MBB-RW. Next, a machine learning model called Extreme Gradient Boosting (XGBoost) is applied to compare the effectiveness of MBB-RW resampling approach against a baseline without resampling. The model performance is evaluated based on the accuracy of Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Finally, the MBB-RW approach is examined to determine whether it can serve as an effective resampling approach or vice versa.

Data description

This study obtained secondary data from the Malaysian Department of Environment (DOE), Malaysia, from 2013 to 2022. The station is in Shah Alam, Selangor, and comprises 83,431 hourly air quality data points for 10 variables, air pollutants and meteorological parameters. The air pollutants featured: Particulate Matter with an aerodynamic diameter of less than or equal to 10 μg (PM_{10}), Sulphur Dioxide (SO_2), Nitric Oxide and Nitrogen Dioxide (NO_x), Nitrogen Dioxide (NO_2), Ozone (O_3) and Carbon Monoxide (CO), whereas meteorological parameters involved: Wind Speed (WS), Wind Direction (WD), Relative Humidity (RH) and Temperature (T). These variables were the independent variables (IV) used to predict the PM_{10} concentrations for the next 24 h ($PM_{10,t+24\,h}$). $PM_{10,\,t+24\,h\,serves}$ as a dependent variable.

The incorporation of both meteorological parameters and air pollutants as input variables gives a broader framework for prediction. Air pollutants such as SO_2 and NO_2 operate as trackers of burning sources that regularly contribute to particulate matter, enhancing the model's ability to detect emission-associated variability

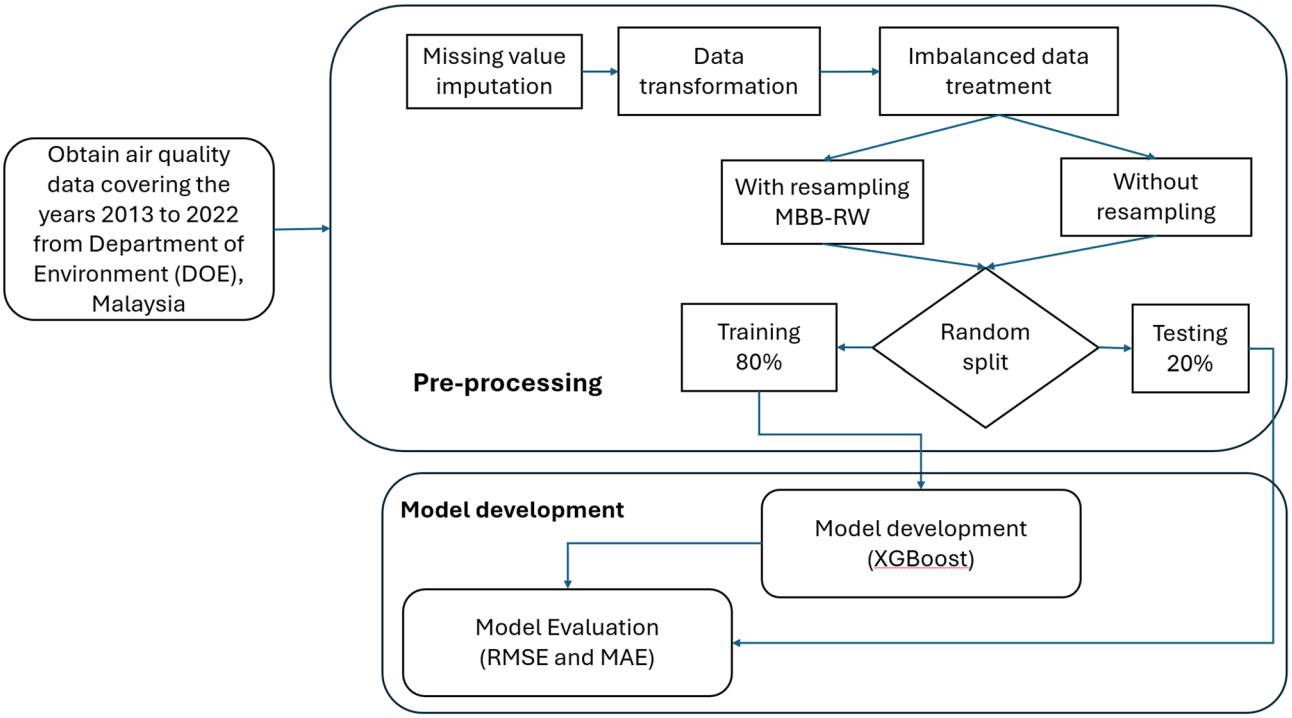

Fig. 1. Research Flowchart.

Variable	Unit	Level of measurement	Role
$PM_{10,t+24\,h}$	μg/m³	Continuous	Target/Dependent
PM_{10}	μg/m³	Continuous	Independent
SO_2	ppb	Continuous	Independent
NO_x	ppb	Continuous	Independent
NO_2	ppb	Continuous	Independent
O_3	ppb	Continuous	Independent
CO	ppb	Continuous	Independent
WS	m/s	Continuous	Independent
WD	°	Continuous	Independent
RH	%	Continuous	Independent
T	°c	Continuous	Independent

Table 1. Description of variables.

in PM_{10}[36–38]. Meteorological variables, including wind speed, wind direction, relative humidity, and temperature, control the dispersion and accumulation processes that identify whether emissions contribute to normal or extreme PM_{10} episodes[39,40].

Therefore, including both air pollutant and meteorological parameters ensures that the model supports the influence of emission signals and atmospheric dynamics, at the same time improving predictive performance and robustness for extreme PM_{10} events. Table 1 presents the variables in this study, along with their respective levels of measurement and roles.

Data preprocessing

Data preprocessing incorporates missing value imputation, data transformation, imbalanced data treatment, and data partition (random split). A substantial amount of missing data may introduce biases or weaken the statistical power of the analysis[41]. Missing values may derive from multiple sources, such as sensor failures, environmental influences, or data transmission issues[41]. Linear interpolation is widely used as an imputation method, particularly to treat small gaps of missing data[42]. As noted by[43], the missing air pollution data in Malaysia is referred as Missing at Random (MAR), and the use of the linear interpolation method suggests that the pattern of missing values does not significantly alter the inherent trends in the data. Therefore, in handling missing values, missing data are imputed using the linear imputation method to ensure that these gaps do not affect the performance of the predictive models. For data transformation, the units of air pollutants SO_2, NO_2, NO_x, O_3,

and CO in ppm are transformed to ppb since the ppm unit is too diminutive, thus compromising the accuracy of the results. The temperature and relative humidity features went through the removal of invalid and abnormal values, where the observations that have the temperature more than 50°c, and the relative humidity below 40% were removed. According to the Malaysian Meteorological Department, the range for average daily minimum humidity can be as low as 40% during the dry months[44]. The WD variable, which is expressed in degrees, has been converted to wind direction index (dimensionless)[45]. According to[46], the formula for conversion is

$$\text{Wind Direction Index (WDI)} = 1 + \sin(\theta - \pi/4) \quad (1)$$

Where θ is the wind direction in radians.

In this study, data partition is performed by implementing standard splitting methods. The dataset is divided into two subsets, allocating 80% for training (model development) and 20% for testing (performance evaluation). [6] suggest that allocating 80% of the data for training and 20% for testing is empirically effective for generating optimal results. To ensure unbiased distribution, random sampling is applied to partition the data into train and test sections[47,48]. The training set (80%) is used to develop and fit machine learning models, giving the learning algorithms to understand patterns and relationships between PM_{10} concentrations and independent variables such as meteorological variables and gaseous pollutants. Meanwhile, the testing set (20%) is not revealed to the model during training. It is used for performance evaluation, yielding an unbiased assessment of the model's ability to generalize to unseen data. This approach helps to minimize the risk of overfitting, where the model performs well on training data but poorly on new data.

Imbalanced regression

Imbalanced regression refers to problems where the target variable exhibits skewness.[49] noted that most regression tasks tend to assume a uniform evaluation of target values' relevance. In discrete imbalanced classification tasks, discovering which category is more relevant is relatively straightforward, specifically when one class is significantly underrepresented, making it as the main focus of prediction. However, in an imbalanced regression task, this becomes more difficult due to the continuous inherent of the target variable. For such scenarios, the researchers need to define which value ranges are considered crucial, as these conditions significantly influence the modelling approach.

To determine the important of target values, [32] propose the definition of relevance function

$$\phi(Y) : y \rightarrow [0,1] \quad (2)$$

where, y represents the domain of the target variable Y, and a relevance value of 1 represents priority area.

A relevance threshold, t_E is launched to represent a threshold or boundary for the definition of relevant values[19]. Its purpose is to identify the values that researcher considers most relevant in their domain environment. This approach enables the learning process to concentrate more on extreme events by defining a relevance threshold that distinguishes extreme and normal events[8]. Given this threshold, the set of extreme events, D_E and the set of normal events, D_N, as follows:

$$D_E = \{\langle x, y\rangle \epsilon D : (y) \geq t_E\} \text{ and } D_N = \{\langle x, y\rangle \epsilon D : (y) < t_E\}, \text{ where } |D_E| << |D_N| \quad (3)$$

where, D_E represents the set of extreme events, D_N represents the set of normal events, y is the target variable and x is the independent variable.

This study deals with the challenge of imbalanced regression in predicting extreme PM_{10} air pollution concentrations. At this stage, it is crucial to provide a good definition of the relevance threshold. Table 2 shows the calculation of the breakpoint concentration for PM_{10} corresponding to each Air Pollution Index (API) category, which is categorised as good, moderate, unhealthy, very unhealthy, and hazardous, which can be an air quality management level for data interpretation processes. API is an effortless and encompassing technique for defining air quality conditions that is easily understood[3]. API with 101–200 μg/m³, indicating that the air quality status is unhealthy. This unhealthy API corresponds to the PM_{10} breakpoint concentration, which is 155 μg/m³. Therefore, in this study, a PM_{10} concentration for extreme threshold, t_E of 155 μg/m³ is marked

API	Air Quality Status	Breakpoint of concentration	Equation for Sub-Index (SI)	Relevance Threshold
0–50	Good	**0 < y < 54**	SI= ($\frac{50-0}{54-0}$) X (y-0) + 0	Normal cases, D_N (assign as 0)
51–100	Moderate	**55 ≤ y < 155**	SI= ($\frac{100-51}{154-55}$) X (y-55) + 51	
101–150	Unhealthy	**155 ≤ x < 255**	SI= ($\frac{150-101}{254-155}$) X (y-155) + 101	Extreme cases, D_E (assign as 1)
151–200	Unhealthy	**255 ≤ x < 355**	SI= ($\frac{200-151}{354-255}$) X (y-255) + 151	
201–300	Very unhealthy	**355 ≤ y < 454**	SI= ($\frac{300-201}{424-355}$) X (y-355) + 201	
301–400	Hazardous	**455 ≤ y < 554**	SI= ($\frac{400-301}{504-425}$) X (y-425) + 301	
401–500	Hazardous	**555 ≤ y < 654**	SI= ($\frac{500-401}{604-505}$) X (y-505) + 401	

Table 2. Breakpoint of PM_{10} concentration.

as the relevance threshold, since the established breakpoint of PM_{10} concentration in Table 2 indicates that a concentration that is greater than or equal to 155 μg/m³ corresponds to an unhealthy API. For air pollution modelling, defining a relevance threshold based on air quality breakpoints ensures that the model emphasises predictions that are most critical for public health interventions. Defining an appropriate relevance threshold is important for the development of the moving block bootstrapping (MBB) stages. Integrating this threshold into the resampling strategy enables the MBB-RW procedure to concentrate more informative blocks, thus enhancing model performance.

Block bootstrapping

Given the situation of imbalanced regression above, the importance of this study is to produce methods that are efficient of developing the representation of the decision regions regarding extreme events. Resampling strategies are the most widely used method for addressing imbalanced datasets, as they change the data distribution to balance the targets[50]. In this study, one of the popular block bootstrapping methods, the Moving Block Bootstrapping (MBB) or also known as Overlapping Block Bootstrapping, is introduced to solve the problems of imbalance regression.

Moving block bootstrapping (MBB)

The MBB is a resampling method used to assess the accuracy of statistical estimates when dealing with observations that are in the form of a finite time series of correlated data[51]. The MBB performs the sampling procedure only within a row of formed blocks. The initial assumption is that the general block length will continue to increase with additional observations in the original sample. Unlike the traditional bootstrap method, the MBB resampled the data in contiguous blocks, rather than by individual values[52]. As a result of such a procedure, the time series structure of the original data remains unchanged within every single block of data[53].

Table 3 shows the summary of the MBB process, which consists of 3 steps. Meanwhile, Fig. 2 shows the visualization of MBB process. According to [25], the MBB method splits the original sample $(Y_1,\ldots,Y_n)$ into overlapping blocks of size l, defined as:

$$B_i = (Y_i, \ldots, Y_{i+l-1}) \tag{4}$$

For i = 1,…., n-l+1, the set of blocks are formed:

$$\{B_1, \ldots, B_{n-l+1}\} \tag{5}$$

where B_i represents the i_{th} block that was obtained from the original sample, Y is the target variable, n is the total number of observations, and l is the length of each block. The number of sampled blocks, B is given by:

$$B = n - l + 1 \tag{6}$$

Let $B_1^*,\ldots, B_b^*$ be a random sample drawn with replacement from original blocks $\{B_{1,\ldots,}B_{n-l+1}\}$,

$$b = n/l \tag{7}$$

b is the number of resampled blocks that will be pasted together to form a pseudo-time series. The i_{th} resampled block, B_i^* consist of the following observations:

$$Y^*_{(i-1)l+1}, \ldots, Y^*_{il}, \text{for } 1 \leq i \leq b \tag{8}$$

Then the full MBB sample then constructed by concatenating these resampled blocks, written as:

$$\underbrace{Y_1^*, \ldots, Y_1^*}_{B_1^*}, \underbrace{Y_{l+1}^*, \ldots, Y_{2l}^*}_{B_2^*}, \underbrace{Y_{2l+1}^*, \ldots, Y_{(b-1)l}^*}_{B_3^* \ldots B_{B-1}^*}, \underbrace{Y_{(b-1)l+1}^*, \ldots, Y_{bl}^*}_{B_b^*} \tag{9}$$

A more detailed illustration of MBB procedure is shown in Fig. 2. The process started with the original dataset, which is then divided into overlapping blocks of fixed length. The total number of blocks is determined by Eq. (6).

Step 1	**The moving block bootstrap begins by dividing the original sample $(Y_1,\ldots,Y_n)$ into overlapping blocks of size l, where each block is defined as $B_i=(Y_i,\ldots,Y_{i+l-1})$, together constituting a set $\{B_{1,\ldots,}B_{n-l+1}\}$. using B = *n*-l+1 is the number of blocks formed.**
Step 2	A bootstrapped sample $B_1^*,\ldots, B_b^*$ is drawn randomly with replacement from original blocks, where b = n/l is the number of resampled blocks that will be pasted together to form a pseudo-time series.
Step 3	Then the moving block bootstrap sample is the concatenation of the resampled blocks, written as: $\underbrace{Y_1^*, \ldots, Y_1^*}_{B_1^*}, \underbrace{Y_{l+1}^*, \ldots, Y_{2l}^*}_{B_2^*}, \underbrace{Y_{2l+1}^*, \ldots, Y_{(b-1)l}^*}_{B_3^* \ldots B_{B-1}^*}, \underbrace{Y_{(b-1)l+1}^*, \ldots, Y_{bl}^*}_{B_b^*}$

Table 3. Steps of moving block bootstrapping (MBB).

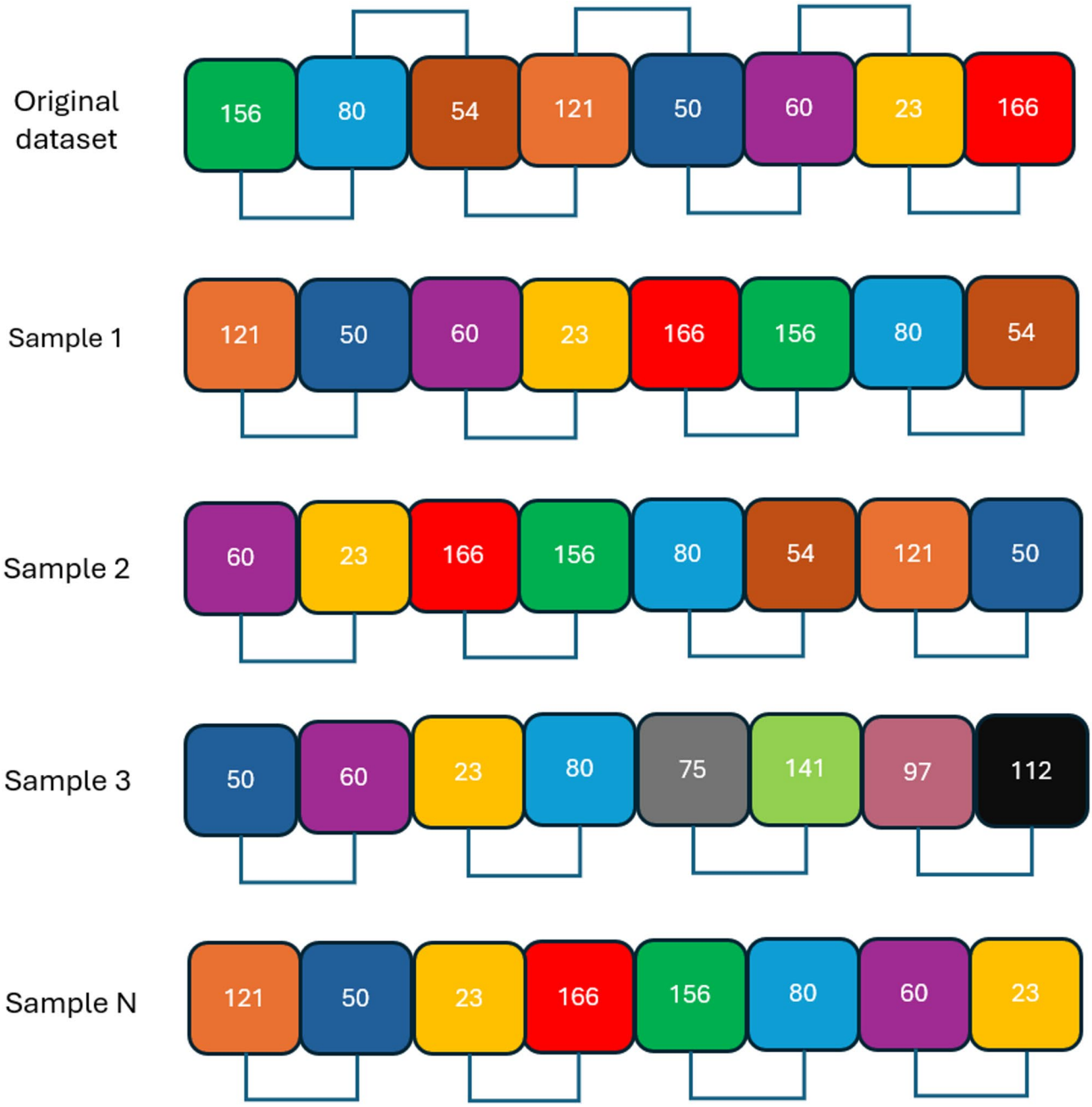

Fig. 2. Moving Block Bootstrapping (MBB) visualization process.

The next process is the number of resampled blocks. A random sample of each selected block is then drawn with replacement. The number of resampled blocks is determined by the Eq. (7). Finally, all these resampled blocks are concatenated to form a new MBB dataset.

Modified moving block bootstrapping with relevance weighting (MBB-RW)

Relevance weighting is a technique used to assign varying levels of importance (weights) to different target values in a regression problem based on their relevance to the study objective[13]. Relevance-weighted modifications of resampling approaches, inspired by the relevance concept from utility-based learning[8,32] are proposed to handle both temporal dependencies and target imbalance. Values below the extreme threshold, t_E (155 μg/m^3) are considered as normal or moderate events. Meanwhile, values above the t_E are considered extreme events. Once the PM_{10} has been set, the values which is below the threshold are assigned as 0, and the values that are above the threshold are assigned as 1. During the MBB-RW process, relevance-weighted sampling is implemented to address the rarity of these extreme events in the dataset. Each overlapping block of 24-hour observations is allocated a binary relevance score, where blocks comprising at least one extreme value are considered relevant. According to [18], sampling weights are introduced to provide a higher diversity in the modified block bootstrapping

dataset, which will include samples that can be either balanced or more favourable to the rare or normal cases. In this study, sampling weights at which a higher weight of 5 is assigned to extreme blocks and a lower weight of 1 is allocated to normal blocks. The 5:1 weight ratio represents a compromise designed to increase exposure to extreme events and avoid bias. The selected ratio is chosen as the target percentage of extreme cases to include in the new training set, with the remaining cases drawn from the set of normal observations[18]. This increases a higher diversity in the adjusted training sets, which will include samples that are either balanced or more favourable to the extreme or normal cases. Previous research on relevance-based resampling provides empirical validation for this method, such as REBAGG[18] and WERCS[8]. This step can effectively increase the chance of selecting rare but important events. These sampling weights are scaled to create a probability distribution to guide the resampling process, thus ensuring a balanced representation of extreme and normal pollution episodes in the bootstrapped dataset while preserving temporal dependencies. The probability weight for extreme events and normal events is derived as below:

Let:
Weight (extreme) (w_e) = 5 (weight assigned to extreme blocks).
Weight (normal) (w_n) = 1 (weight assigned to normal blocks).
N_e: Number of extreme blocks.
N_n: Number of normal blocks.
Then,

$$\begin{aligned}&\text{Total Weighted Count, } T = [\text{Number of extreme blocks.weight(extreme)}] + \\ &[\text{Number of normal blocks.weight(normal)}] = (N_e \,.\, w_e) + (N_n \,.\, w_n)\end{aligned} \tag{10}$$

$$\text{Probability weight for extreme blocks, } P_e = \text{Weight (extreme)} \,/\, \text{Total weighted count} = w_e/T \tag{11}$$

$$\text{Probability weight for normal blocks, } P_n = \text{Weight (normal)} \,/\, \text{Total weight} = w_n/T \tag{12}$$

Table 4 shows that the algorithm of MBB-RW, that integrated with relevance weight. First, the air pollution dataset with PM_{10} for the next day as a target variable, y is inputted. The block length of 24 h was chosen when this study aimed to predict PM_{10} concentrations after 24 h. Set the threshold t_E=155 (for upgrading extreme relevance) and sampling weight, w_s (Extreme block weight w_e=5, normal block weight w_n=1). In the MBB-RW procedure, each observation is assigned a relevance weight, 1 if the PM_{10} concentration is more than t_E=155, else assigned as 0. Next, the blocks are constructed, with the total number of blocks given by B = N-⌊+1. After forming these blocks, each block is assigned a sampling weight, w_s. Each block is assigned as w_e=5 if it contains at least one observation with relevance weight 1, otherwise, it is assigned as sampling weight w_n=1. When dealing with weighted resampling, especially in a random sampling process, the algorithm requires sampling probabilities, not raw weights, to determine how likely each block is to be chosen. In the context of moving block bootstrapping with relevance-based weighting, converting sampling weights into probabilities is a crucial step to maintain the functionality of the random sampling process. Next, a set of indices $\{s_1, s_2, \ldots s_N\}$ is sampled from the full list of possible blocks. Each index is selected based on the probability that corresponds to its assigned sampling weight, reflecting the associated importance of the corresponding block. Higher sampling probability may appear multiple times in the resampled dataset. Once all the sampled blocks have been extracted, they are combined sequentially to form a new MBB-RW dataset. This combined dataset improves the representation of rare but important events.

Figure 3 provides a clearer illustration of the MBB-RW procedure. The key differences between standard MBB and MBB-RW stem from the integration of relevance weights, sampling weights and sampling probabilities. Those elements are introduced in the MBB-RW to increase the representation of extreme events in the dataset. The procedure started with the original dataset, which was then divided into overlapping blocks of fixed length. The total number of blocks is determined by Eq. (6). Then, the relevance threshold, t_e is applied to each observation within the blocks. Next, a sampling weight of 5 is assigned to each block that contains at least one extreme observation, reflecting their importance. Afterwards, blocks are sampled with replacement from the original set, using probabilities corresponding to their assigned weights, prioritising blocks containing extreme events. Finally, all these resampled blocks are concatenated to form a new MBB-RW dataset.

Model development

Extreme gradient boosting (XGBoost)

In this study, XGBoost will be applied for model development of air pollution dataset. XGBoost is a decision tree ensemble based on a gradient boosting framework that is highly scalable, designed for supervised learning tasks, including both classification and regression problems[54,55]. Several studies utilise XGBoost for predicting air pollution in environmental modelling[4,34,56–58].

Unlike traditional gradient boosting algorithms, XGBoost incorporates several novel enhancements. The objective function of XGBoost comprises a loss function and a regularisation term[59]. XGBoost build an additive expansion of the objective function by minimising a loss function[59]. XGBoost exclusively focuses on decision trees as its base classifiers, and a variation of the loss function is used to control the complexity of the trees[54]. It operates by sequentially building an ensemble of decision trees, where each new tree attempts to correct the prediction errors made by the previous ensemble to produce a single prediction[60]. Different from gradient boosting, the XGBoost objective function involves a regularisation term to avoid overfitting[54].

Objective Function- Assume that a dataset is $D = \{(x_i, y_i): i = 1 \ldots n\}$ where x_i represents the input features and y_i the corresponding target values. Let $\widehat{y}_I$ denote the predicted output produced by an ensemble model, represented by the generalised model as follows[61]

Algorithm: Moving Block Bootstrapping with relevance weight (MBB-RW)
Input: Data set D={$(x_1,y_1),\ldots, x_n,y_1)$} with y_i as PM_{10} values Block size ⌊=24 hours Threshold t_E=155 (for refining extreme relevance) Sampling weight (Extreme block weight w_e=5, normal block weight w_n=1) Number of bootstrap blocks B=N-⌊+1 **Procedure for MBB-RW:** **Assign relevance weight** Relevance weight=$\begin{cases} 1 \text{ if } y_i > t_e \\ 0 \text{ otherwise} \end{cases}$ **Assign sampling weights to blocks** Assign w_s=$\begin{cases} w_e \text{ if any relevance weight} = 1 \\ w_n \text{ otherwise} \end{cases}$ **Convert sampling weights to sampling probabilities** Probability weight for extreme blocks, $P_e=w_e/T$ Probability weight for normal blocks, $P_n=w_n/T$ **Create sampled blocks based on sampling probabilities** Sample S_i block indices {$s_1,s_2,\ldots.s_N$} with replacement **Concatenation of the bootstrapped dataset** $D^*=\sum_i^n Si$

Table 4. Moving block bootstrapping Algorithm.

$$\widehat{y}_i = \theta\ (\mathrm{xi})\ =\ \sum\nolimits_{k=1}^{K} f_k(x_i) \tag{11}$$

here,f_k is a regression the k_{th} regression tree in the ensemble.

$f_k(x_i)$ is the prediction (or score) given by the k_{th} tree for the i-th observation in the data.

K is the total number of trees in the model.

To learn the functions f_k, XGBoost aims to minimise the following regularised objective function.

$$\mathrm{Obj}\ =\ L(\theta) = \sum\nolimits_i L(\widehat{y}_i, y_i) + \sum\nolimits_k \Omega(f_k) \tag{12}$$

where,

$\sum_i L(\widehat{y}_i, y_i)$ is the custom loss function. Loss function, L is a differentiable function that measures the difference between the prediction $\widehat{y}_I$ and the actual y_i [59]. This loss function can be embedded into the split criterion of decision trees, leading to a pre-pruning strategy.

$\Omega(f_k)$ is a regularisation term that limits the complexity of each tree f_k to prevent overfitting[61]. The penalty term or regularisation term Ω is included as follows:

$$\Omega(f_k) = \gamma\, T + \frac{1}{2}\lambda\, w^2 \tag{13}$$

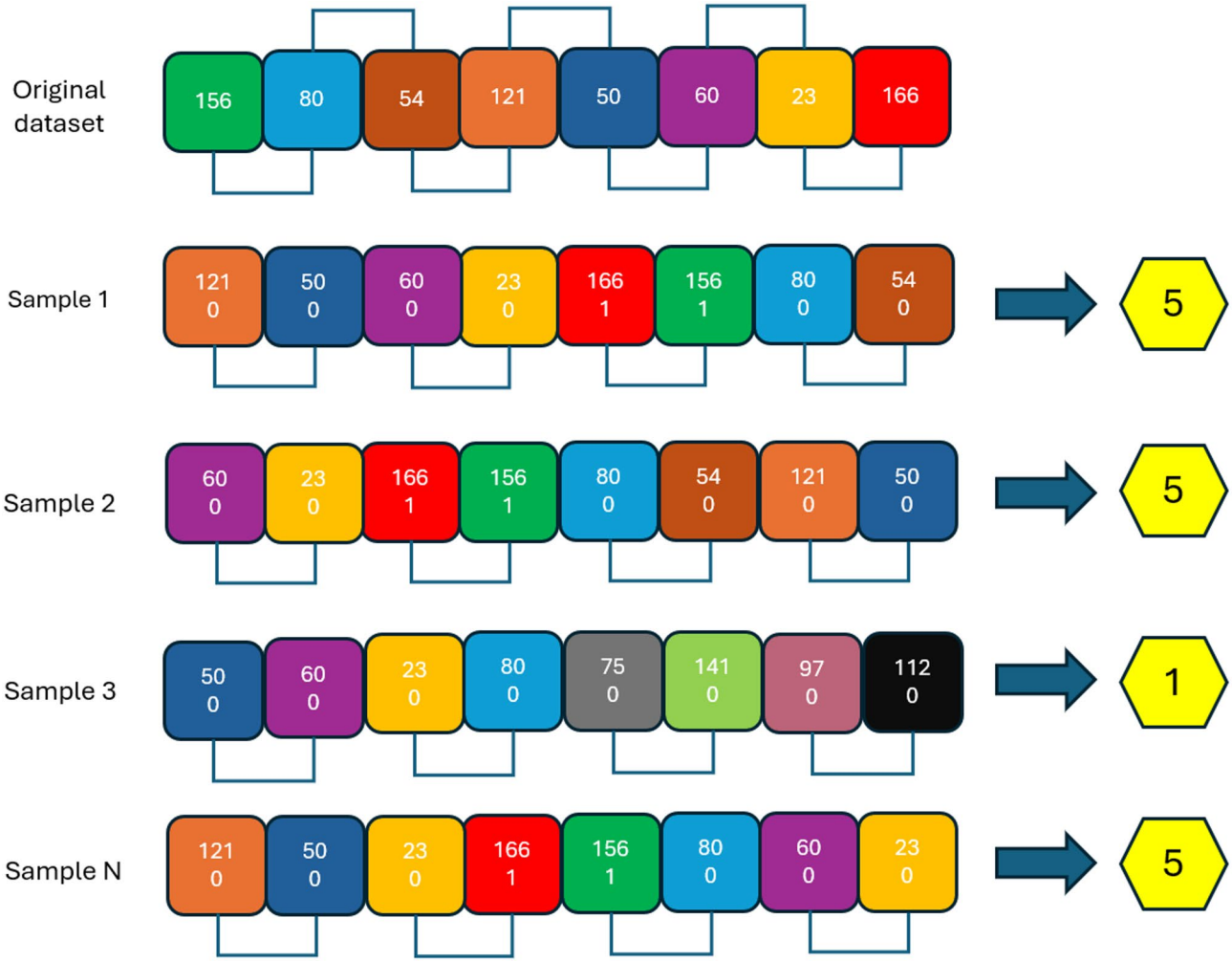

Fig. 3. MBB-RW illustration process.

where,

λ and γ are the parameters controlling the penalty for the number of leaves T and the magnitude of leaf weights w, respectively.

Parameter setting for XGBoost

All models were developed using the default parameter settings provided by the respective libraries (scikit-learn) to evaluate the performance of XGBoost in predicting PM_{10} concentrations in Malaysia. Using general parameters allows for a fair comparison and provides a reasonable baseline to adjusted models, suggesting that general parameter settings are an appropriate starting point for model evaluation. The XGBoost regressor is an ensemble algorithm method based on boosting and employs relatively aggressive parameter settings. In this study, the parameter setting (max_depth = 6, learning_rate = 0.3, seed = 100 and n_estimator = 100) was obtained through empirical tuning to achieve the best adjustment between accuracy and generalization. These parameters were evaluated, and the selected configuration that produced the lowest RMSE effectively captured the nonlinear behaviour of PM_{10}. These configurations tries to balance computational efficiency between speed and accuracy, with built-in regularisation capabilities to prevent overfitting.

Model performance

In the context of air pollution modelling, accurate prediction of PM_{10} concentrations is crucial. To evaluate the accuracy of prediction models, several performance indicators are commonly used. Among them are Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). According to[62], a comparison of the best statistical PM_{10} forecasting methods with the lowest values of RMSE was conducted to select the best fit prediction model. The difference between the estimated and observed values is obtained to investigate the performance of each estimation method. The most appropriate methods are selected based on the lowest value of each statistical evaluation. The criteria formulas are shown below:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{Y}_i - Y_i\right)^2} \tag{14}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\widehat{Y}_i - Y_i\right| \tag{15}$$

where,

n is the total number of hourly measurements of a particular station.

$\widehat{Y}_i$ is the estimated value of $PM_{10,t+24\,h}$.

Y_i is the observed value of $PM_{10,t+24\,h}$.

Result and discussion

Descriptive statistics summary for PM_{10} concentration before and after resampling MBB-RW

Table 5 shows the descriptive statistics for PM_{10} concentrations in Shah Alam from 2013 to 2022 without and with resampling using the modification of Moving Block Bootstrapping with relevance weighting (MBB-RW) approach. From the table, the number of observations is approximately the same, which is without resampling ($N = 82431$) and when resampling with MBB-RW ($N = 82416$). After the MBB-RW approach, the number of normal events decreases from 81,499 to 78,243, and the number of extreme events increases from 932 to 4173. The mean, median, and standard deviation show an increase following the application of the MBB-RW approach. Specifically, the mean value rose from 41.7078 μg/m^3 to 54.35 μg/m^3. The median increased from 35 μg/m^3 to 38 μg/m^3 and the standard deviation widened from 31.7368 μg/m^3 to 55.394 μg/m^3. These changes are due to the additional extreme PM_{10} values introduced from the MBB-RW strategy. The skewness of the PM_{10} concentrations was evaluated to understand the symmetry and presence of extreme values with and without the MBB-RW. Following the MBB-RW procedure, the skewness decreases from 4.574 to 3.6010. This reduction is attributed to the introduction of additional extreme PM_{10} values through the MBB-RW strategy, which increased the frequency of extreme values in the upper tail.

Figure 4 demonstrates the boxplot of hourly PM_{10} concentrations in Shah Alam, Malaysia, from 2012 to 2022. During years 2013 to 2016, these years reveal higher median concentrations, larger variability, the broadest interquartile ranges (IQRs) (Upper quartile Q_3 – Lower quartile Q1) and various extreme values, with some exceeding 500. This shows that regular and extreme pollution episodes frequently occur throughout these years. These frequent and extreme values in 2013–2016 are likely due to transboundary haze or industrial sources[63–65]. Meanwhile, from 2017 to 2019, the medians and interquartile ranges (IQRs) decreased substantially. Outliers are still present, but less extreme and frequent. Air quality is improving, but continues to fluctuate. The most marked improvement is observed in the year 2020 to 2022, which reveals the lowest median values, narrowest IQRs, and minimal extreme values. These PM_{10} concentration improvements may be linked to enhanced pollution control measures, changes in emission sources, and reduced industrial and vehicular activity during the COVID-19 pandemic[66]. All distributions are positively skewed, indicated by extended upper whiskers and high-value outliers. This means that while most hourly values are low and regular, there are occasional spikes in pollution.

Figure 5 shows the correlation heatmap in Shah Alam, representing the correlation coefficients between different air pollution variables. The colour gradient scales from blue (strong negative correlation, −1) to red (strong positive correlation, +1), with pale shades indicating weaker correlations. This heatmap yields valuable discoveries into the relationship between meteorological and air pollutants parameters. Regarding our main variable of interest, PM_{10}, we found that all variables related exhibit positive correlation with PM_{10}, except for Relative Humidity. From the correlation heatmap, there was a negative relationship between PM_{10} and Relative Humidity (RH) ($r = -0.03$). These findings are supported by[67] when their study shows a negative correlation between PM_{10} and RH. This suggests that as the humidity level rises, PM_{10} is prone to decrease slightly. Understanding these correlations helps in air quality analysis, especially when developing a model and predicting extreme pollution events.

Figure 6 and Fig. 7 illustrate the distribution of PM_{10} without resampling and after applying the MBB techniques. From **Fig. 6**, the original dataset (blue line) exhibits a sharp peak at lower concentrations and a thin

	Without resampling	Resampling with MBB-RW
N of air pollution dataset	82,431	82,416
N of normal events	81,499	78,243
N of extreme events	932	4173
Mean	41.7078	54.3500
Median	35.0000	38.0000
Std.dev	31.7368	55.3940
Variance	10007.226	3068.5300
Skewness	4.574	3.6010
Minimum	0.63	1.0000
Maximum	575.0000	575.0000

Table 5. Descriptive statistics for PM_{10} concentrations in Shah Alam without resampling and with MBB-RW resampling approach.

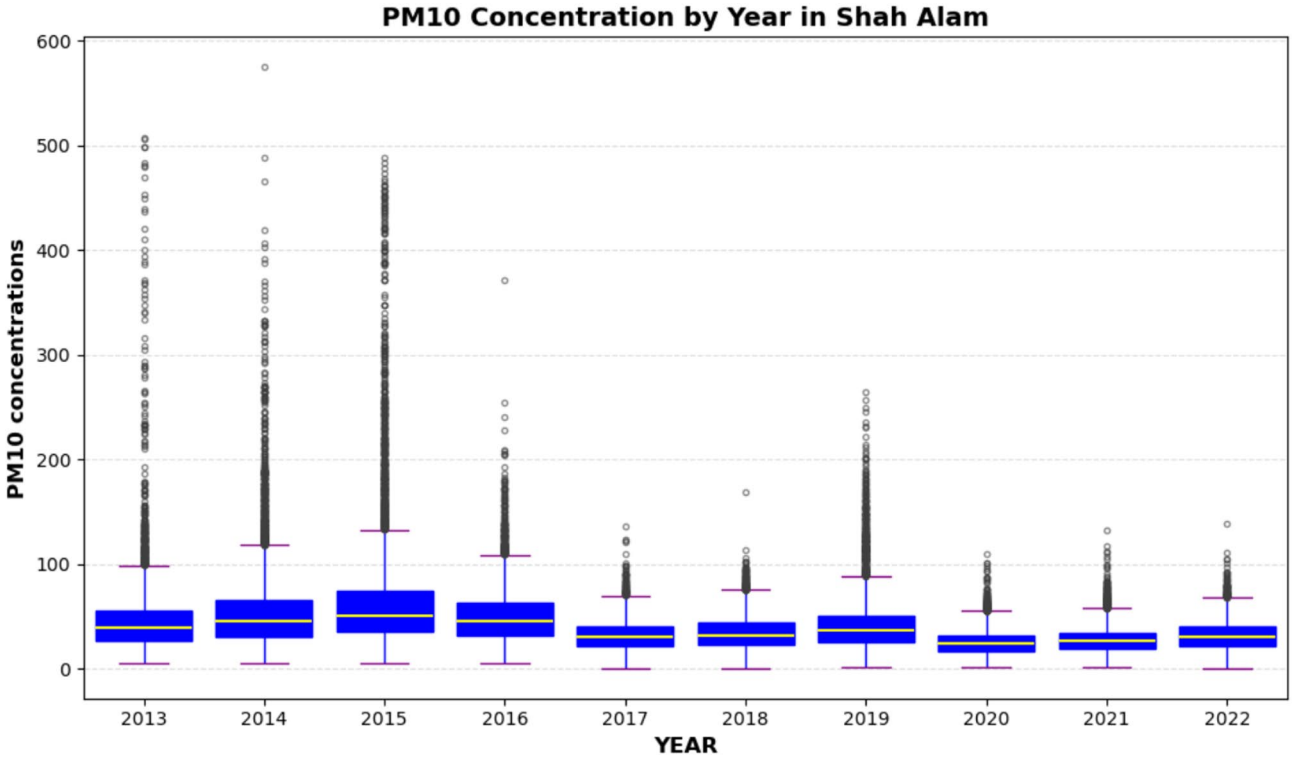

Fig. 4. Boxplot of hourly PM_{10} concentration in Shah Alam.

right tail, reflecting a highly right-skewed distribution, consistent with a skewness of 4.574. After MBB-RW (red line), the distribution remains right-skewed. However, the peak flattens, indicating a moderate reduction in skewness to 3.6010. This visual change suggests that the MBB-RW strategy introduced an additional observation of extreme values, resulting in a distribution with reduced skewness in the air pollution dataset. As a result, the distribution appears more balanced with reduced skewness. Figure 7 shows a Two-Dimensional Kernel Density Estimation (2D KDE) of PM_{10} contour plot, demonstrating the joint distribution between the original PM_{10} concentrations and the PM_{10} values generated through MBB-RW approach. The density contours illustrate regions of differing concentration frequency, with the red contour indicating the top density and the blur regions indicating gradually lower densities.

A red contour showing that both the original (without resampling) and with MBB-RW of PM_{10} values range approximately between 0 and 50 $\mu g/m^3$. This high-density area implies that most of the PM_{10} observations in both datasets (with and without MBB-RW) reflect to low PM_{10} concentrations, corresponding with normal air quality patterns in non-extreme conditions. However, a significant difference is seen in the MBB-RW values compared to the original data. The MBB-RW process generated a broader range of air pollution events, particularly skewed towards higher concentrations. This vertical extension of the density contours reveals that the MBB-RW approach effectively generates wider variability of extreme PM_{10} values. This pattern shows that the MBB-RW strategy enhances the representation of extreme air pollution events in the dataset, providing a more balanced distribution for further analysis.

Performance of XGBoost without resampling

Table 6 presents the performance of machine learning model, Extreme Gradient Boosting (XGBoost) evaluated on the overall dataset before the imbalanced data treatment, Moving Block Bootstrapping with Relevance Weight (MBB-RW) for hourly PM_{10} concentration predictions in Shah Alam station from the period of 2013 to 2022. The performance indicator, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are provided for overall air pollution dataset, normal and extreme events, along with the total number of observations (N) in each category. Overall dataset referring to air pollution dataset, without distinguishing between normal and extreme events. Normal events represent observations below the threshold, t_e, typically corresponding to common air pollution levels. Meanwhile, extreme events denote observations exceeding the relevance threshold, t_e, reflecting rare and extreme episodes of air pollution events.

The results reveal a distinct discrepancy in model performance across different levels. Overall, the RMSE and MAE values for the original air pollution dataset are 19.8421 and 14.2396, respectively, reflecting the average difference between predicted and actual values. When split by event type, the model performs better during normal events, with a reduced RMSE of 18.2636 and an MAE of 13.4048. However, performance extensively worsens during extreme events, with the RMSE increasing to 108.3010 and the MAE rising to 85.1041. These sharp increases in error values suggest that the model struggles to accurately predict extreme pollution levels. These findings are strongly supported by [4], whose study shows that extreme values affect the model's performance.

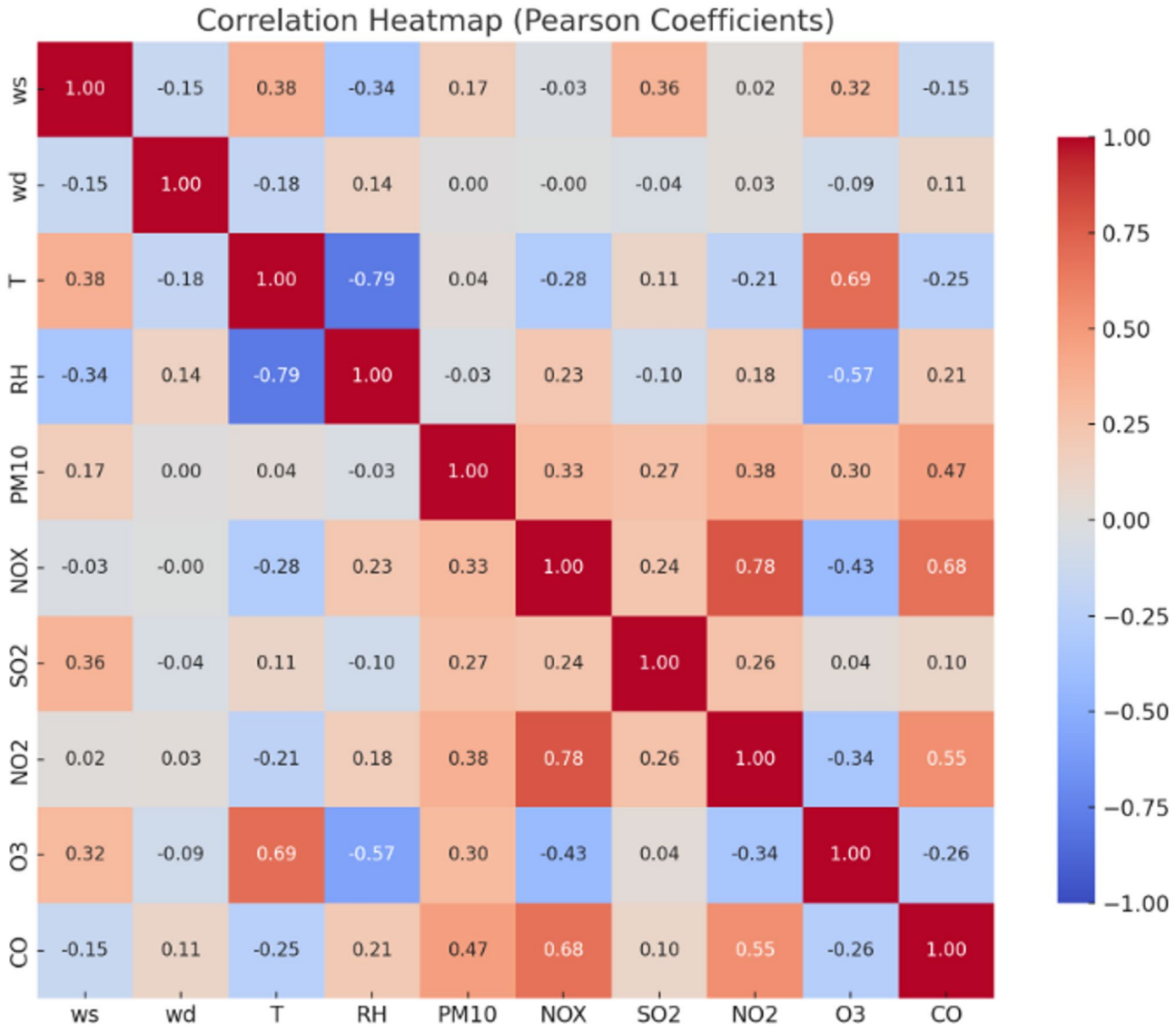

Fig. 5. The correlation heatmap of the air pollution dataset in Shah Alam.

Furthermore, the distribution of events within the dataset is highly imbalanced. Out of 16,486 total observations for testing XGBoost model, only 192 observations (approximately 0.23%) correspond to extreme events. Conversely, normal events account for 16,295 instances. This major imbalance in distribution causes the model's poor predictive performance in extreme cases, as it is more biased toward the more frequently occurring normal events during model's training[8,17,68].

The model indicates good performance in predicting normal air pollution concentrations but presents limitations in catering extreme pollution events. These results emphasize the necessity of utilizing appropriate resampling strategies to enhance model robustness and accuracy, especially for underrepresented extreme cases.

Performance of XGBoost with resampling – moving block bootstrapping with relevance weight (MBB-RW) resampling approach

Table 7 presents the model's performance after applying the MBB-RW technique enhanced with relevance weighting. This table summarizes quantitatively the performance in terms of RMSE and MAE, measured separately for the full dataset, normal events, and extreme events. Additionally, the sample sizes (N) for each category are reported.

The after MBB-RW results demonstrate improvements in model performance, especially in the prediction of extreme events. The RMSE for extreme events reduced significantly from 108.3010 (without resampling) to 39.1846, constituting a reduction of approximately 63.8% in prediction error. Similarly, the MAE declines sharply from 85.1041 to 27.1082, aligning closely with the MAE observed for normal events. These results portray a notable improvement in the model's ability to capture extreme pollution patterns. Simultaneous reduction in RMSE and MAE validates the conclusion that model accuracy for extreme events improved meaningfully.

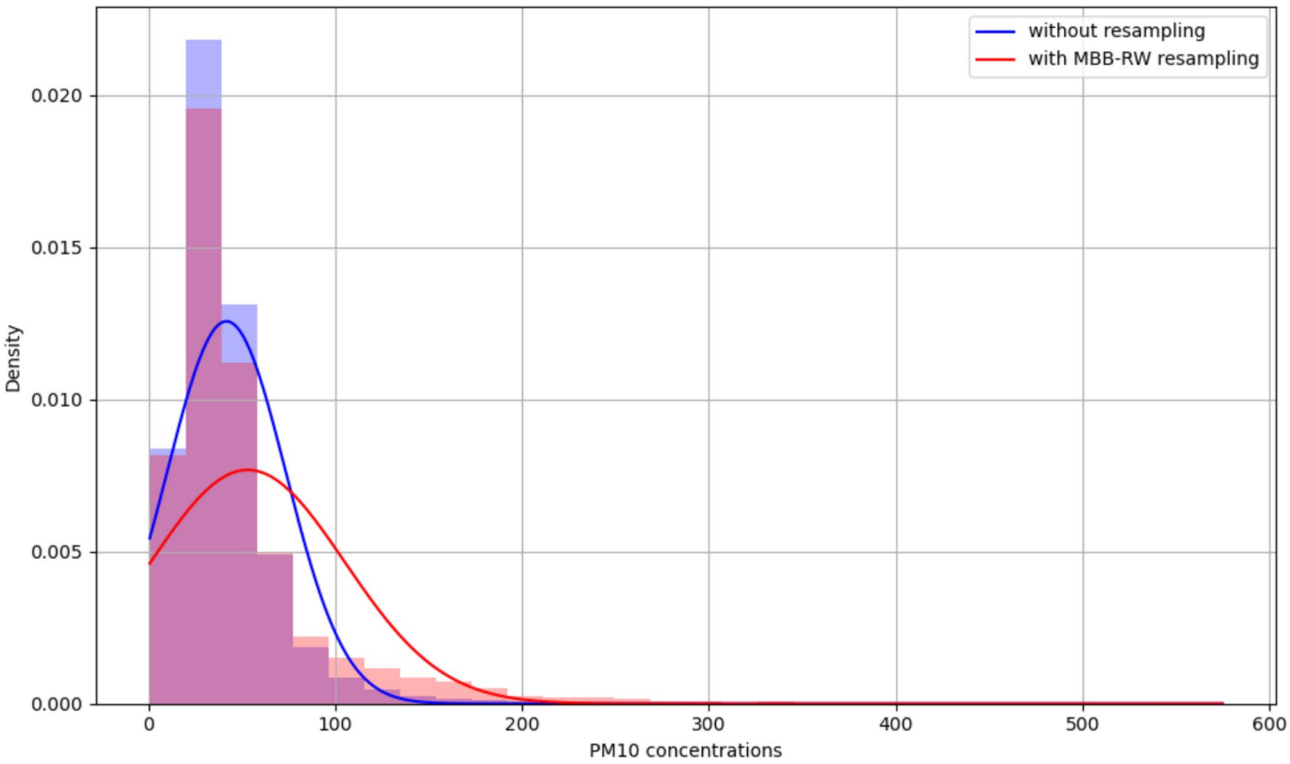

Fig. 6. Distribution plot without resampling and with MBB-RW resampling strategy.

The enhancements can be attributed to the increased representation of extreme events in the dataset after MBB-RW strategy. The number of extreme cases rose from 192 to 769, enhancing the model's exposure to rare events during training. This more balanced distribution, enhanced by relevance weighting, allowed the model to effectively capture and generalize the underlying characteristics of both normal and extreme events. This is supported by [31] when the Block Bootstrapping approach can enhance the performance on extreme events prediction. Moreover, several studies have proven that block bootstrapping can serve as a powerful resampling method when it can generate robust estimates, producing stable estimation, especially in environmental analysis[27,53].For normal events, performance also improved slightly. The RMSE decreased from 18.2636 to 16.8257, and the MAE decreased from 13.4048 to 12.3299. These enhancements suggest that the MBB-RW technique minimally affects the model's performance on frequent cases while simultaneously addressing the underperformance in extreme scenarios. The modification of MBB with relevance weighting (MBB-RW) substantially mitigated the effects of data imbalance and enhanced the model's predictive capability, particularly for extreme pollution events. This finding highlights the effectiveness of resampling strategies in improving model prediction in the context of imbalanced regression tasks.

Table 8 shows how the model's performance compares with and without the modified MBB-RW method. The results are compared for the overall dataset, normal events, and extreme events using three error measures: RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error). The number of data points (N) in each group is also presented.

After applying MBB-RW, the model's performance improved overall. For the full dataset, the RMSE reduced from 19.84 to 18.44, and MAE dropped from 14.24 to 13.16. This shows that the model became more accurate overall. For normal events, the model also achieved slightly better. RMSE reduce from 18.26 to 16.83 and MAE from 13.40 to 12.33. This suggests a small but consistent improvement for predicting normal pollution events.

The most obvious improvement occurred in extreme events. without resampling approach, the model faced high error performance—RMSE was 108.30 and MAE was 85.10. After MBB-RW, these reduced sharply to 85.1041 and 27.1082, respectively. Moreover, the number of extreme cases increased from 192 to 769, giving the model more data to learn from. A clearer visualization of each performance evaluation with and without resampling is presented in Fig. 8, where the RMSE and MAE for extreme events dropped dramatically when the MBB-RW method was applied.

Figure 9 presents the actual vs. predicted plot for PM_{10} concentrations without and with data resampling using MBB-RW. (A) represents the actual vs. predicted of $PM_{10,24\ H}$ using XGBoost model for extreme events without data resampling, and (B) represents the actual and predicted of $PM_{10,24\ H}$ using XGBoost model for extreme events with data resampling using MBB-RW. The actual event observations are depicted in blue, while the predicted extreme observations are shown in orange. Referring to plot A, the predicted values mostly underestimate the actual PM_{10} concentrations for extreme events. The orange prediction line lies significantly below the blue line for most of the observations. This underestimation suggests that the model, trained on the

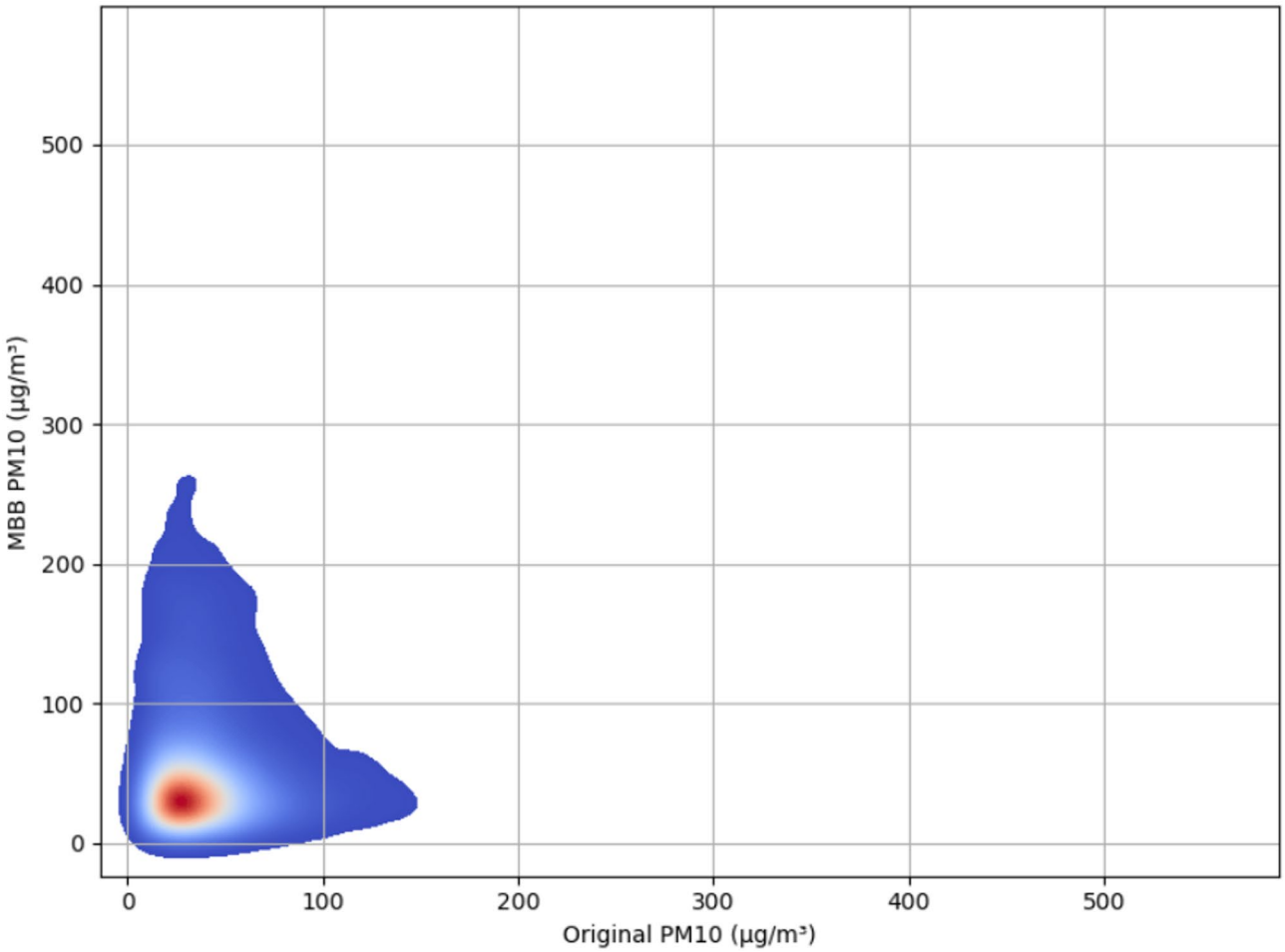

Fig. 7. Two-Dimensional Kernel Density Estimation of PM_{10} concentrations.

Performance Indicator	Overall air pollution dataset (without resampling)	Normal event	Extreme event
RMSE	19.8421	18.2636	108.3010
MAE	14.2396	13.4048	85.1041
N	82,431 (split 20% for testing = 16486)	16,295	192

Table 6. Performance result for XGBoost of air pollution modelling without resampling.

Performance Indicator	Overall air pollution dataset with MBB-RW resampling approach	Normal event	Extreme event
RMSE	18.4400	16.8257	39.1846
MAE	13.1567	12.3299	27.1082
N	82,416 (split 20% for testing = 16484)	15,715	769

Table 7. Performance result for XGBoost of air pollution modelling with MBB-RW resampling approach.

original imbalanced air pollution dataset, is biased toward frequent events and lacks exposure to high extreme events. From plot B, model predictions after the training data were improved using the MBB-RW approach. The visual difference is outstanding. The predicted value demonstrates closer consistency with actual observations. There is a significant improvement in the model's ability to produce the peaks of the true PM_{10} concentrations, indicating that MBB-RW helped mitigate the imbalance problems by increasing the extreme cases in the training set. This improvement suggests that modification of MBB-RW enabled the model to better analyse and respond to complex patterns associated with extreme pollution events. In summary, applying MBB-RW helped the model make better predictions, especially for extreme pollution events, while also slightly improving accuracy for normal events. This shows that MBB-RW is an effective method to handle imbalanced data in this study.

		Without resampling	with MBB-RW resampling approach
Full dataset	RMSE	19.8421	18.4400
	MAE	14.2396	13.1567
	N	82,431	82,416
Normal events	RMSE	18.2636	16.8257
	MAE	13.4048	12.3299
	N	16,295	15,715
Extreme events	RMSE	108.3010	39.1846
	MAE	85.1041	27.1082
	N	192	769

Table 8. Comparison of the performance of XGBoost without resampling and with MBB-RW resampling approach.

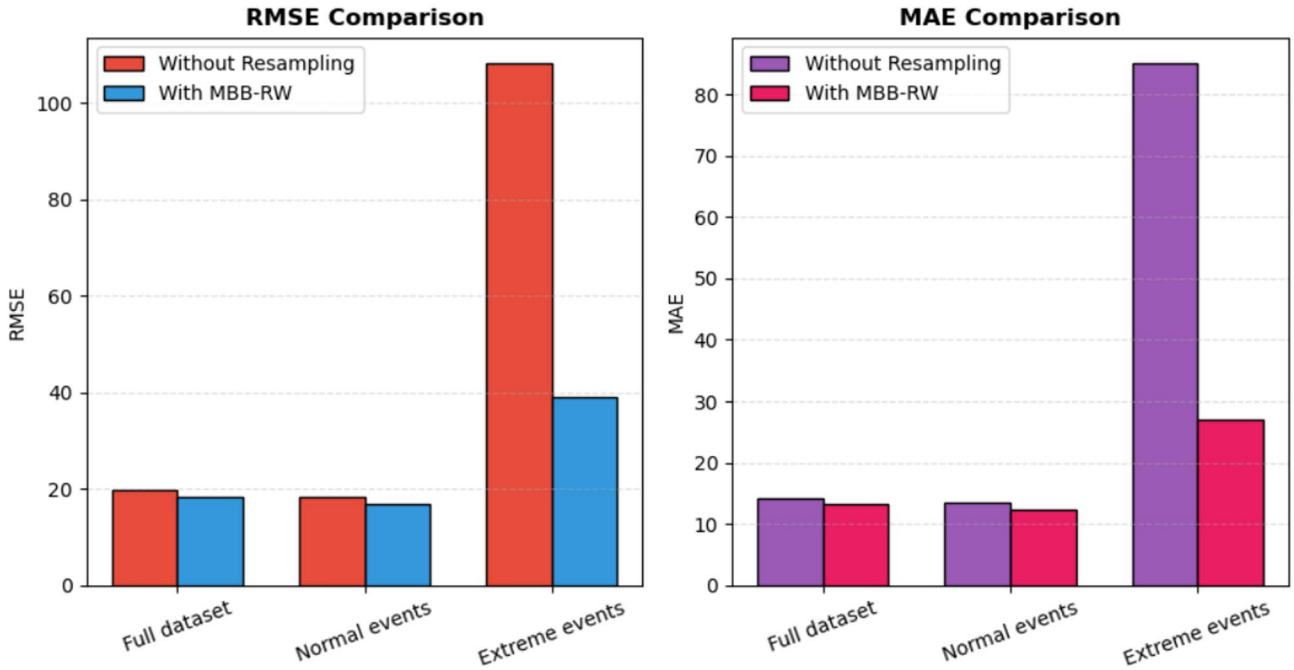

Fig. 8. Comparison of XGBoost Performance Metrics with and without MBB-RW Resampling.

Conclusion and recommendations

This study highlights the challenges of predicting extreme air pollution events using XGBoost machine learning model trained on imbalanced datasets. Initial results before the MBB showed that the model consistently underpredicted extreme PM_{10} concentrations. The modification of the Moving Block Bootstrap with relevance weighting (MBB-RW) resampling technique significantly improved the model's performance by better representing the extreme events in the training data when the RMSE, and MAE shown by the reduction of error before and after MBB-RW from 108.3010 to 39.1846 and 85.1041 to 27.1082. Visual comparisons with and without resampling demonstrated significant improvements in peak alignment, variability, and magnitude accuracy of predictions. These findings emphasize the effectiveness of MBB-RW in addressing data imbalance and enhancing model prediction towards extreme events. Overall, integrating appropriate resampling methods such as MBB-RW is essential for improving the air quality forecasting models.

This study has several limitations that should be considered. First, the resampling method is designed specifically for datasets with similar statistical properties to air pollution data, such as positive, skewed distributions with extreme events. Second, the analysis was limited to PM_{10} data from a single monitoring station in Shah Alam, Malaysia, which may restrict the generalizability of the findings to other regions with different meteorological profiles.

Based on the findings of this study, several recommendations can be built for future work. Comparative studies using multi-station data may highlight robustness in imbalanced data studies and improve air quality forecasting frameworks. Future research could apply the proposed MBB-RW framework beyond PM_{10} to include other air pollutants such as $PM_{2.5}$, NO_2, SO_2, and O_3, together with different stations in Malaysia and other regions. Moreover, the method shows potential for adaptation to other imbalanced time-series problems beyond

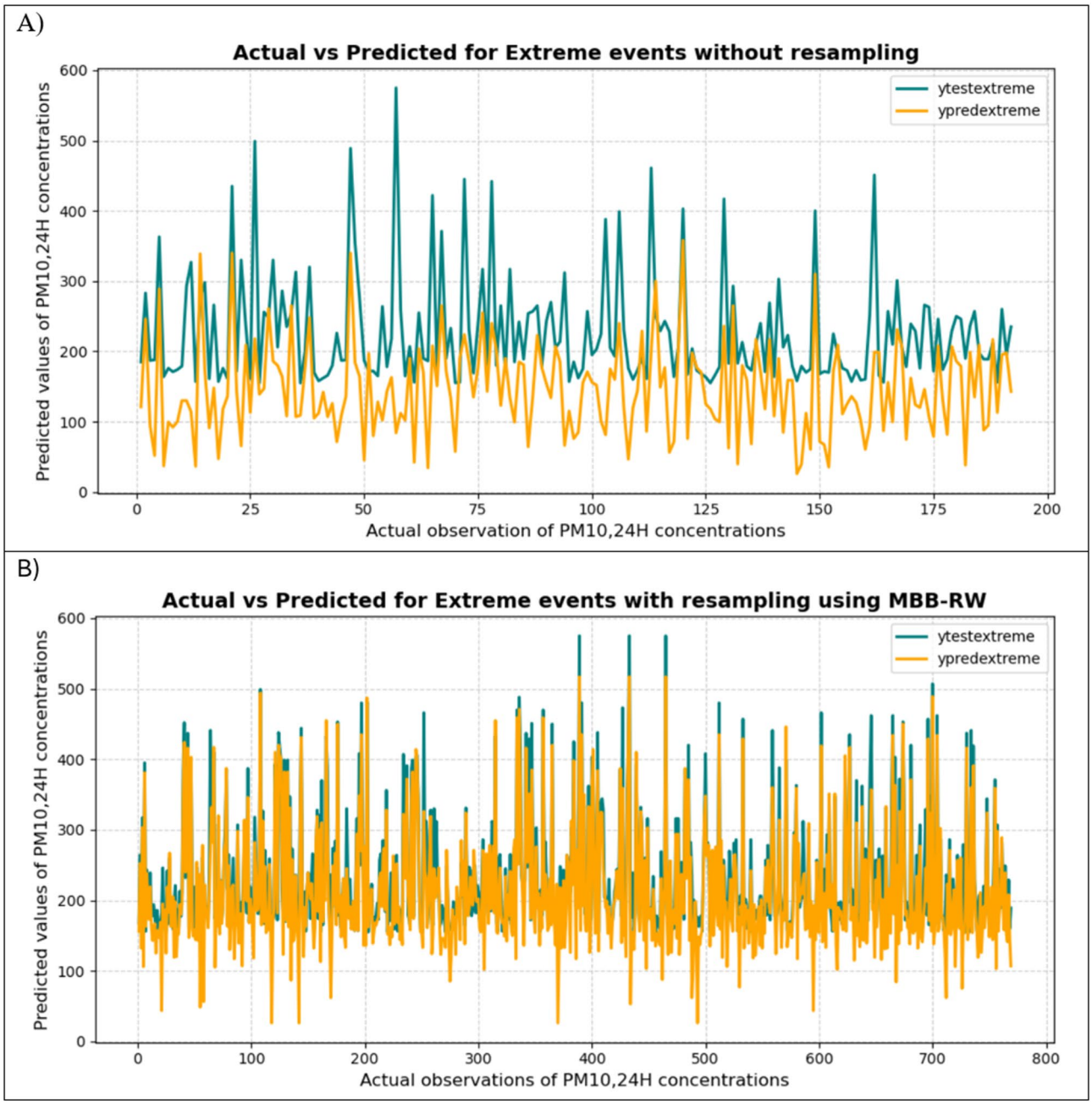

Fig. 9. Actual vs. predicted of the XGBoost model before and after MBB-RW.

the air pollution area, including extreme rainfall prediction, energy demand forecasting, and financial time-series analysis. Such extensions would demonstrate the versatility of the MBB-RW approach and strengthen its contribution to modelling extreme events in diverse domains.

A specific comparison of MBB-RW against these established previous methods would further demonstrate whether weighting improves the handling of rare, extreme events, other than what traditional resampling strategies can achieve. Beyond XGBoost, the proposed resampling method, MBB-RW should be integrated with other machine learning models such as Random Forest, Artificial Neural Network and Light GBM. This would show the adaptability of the approach and help identify whether certain algorithms are more effective in handling extreme value prediction under different conditions.

Data availability

The air quality dataset analyzed in this study was obtained from the Department of Environment (DOE) Malaysia and is subject to confidentiality restrictions. As such, the data cannot be publicly shared. Researchers interested in accessing the data may submit a formal request to the Department of Environment Malaysia www.doe.gov.my. Approval is subject to DOE Malaysia's data-sharing policies and compliance requirements. Further

information about the dataset and how it was used in this study can be obtained from the corresponding author, Dr. Ahmad Zia Ul-Saufie (email: ahmadzia101@uitm.edu.my).

Received: 1 July 2025; Accepted: 10 November 2025
Published online: 12 December 2025

References

1. Aini, Q. et al. Factors that contribute to air pollution in Malaysia. *Malaysian J. Bus. Econ.* **8**, 43–58 (2023).
2. Gulati, S. et al. Estimating PM2.5 utilizing multiple linear regression and ANN techniques. *Sci Rep.* **13**, 22578 (2023).
3. Rani, N. L. A., Azid, A., Khalit, S. I., Juahir, H. & Samsudin, M. S. Air pollution index trend analysis in Malaysia, 2010-15. *Pol. J. Environ. Stud.* **27**, 801–808 (2018).
4. Muhammad, M., Ul-Saufie, Z. & Radi, A. A. Evaluating the Performance of Tree-Based Model in Predicting Haze Events in Malaysia. *International Journal of Advanced Computer Science and Applications.* **16**, 1127-1135 (2025).
5. He, Z., Guo, Q., Wang, Z. & Li, X. A. Hybrid Wavelet-Based Deep Learning Model for Accurate Prediction of Daily Surface PM2.5 Concentrations in Guangzhou City. *Toxics*. 13, (2025).
6. Gupta, N. S. et al. Prediction of air quality index using machine learning techniques: A comparative analysis. *J. Environ. Public. Health*. **2023**, 1–26 (2023).
7. Martins, L. D. et al. Extreme value analysis of air pollution data and their comparison between two large urban regions of South America. *Weather Clim. Extrem*. **18**, 44–54 (2017).
8. Ribeiro, R. P. & Moniz, N. Imbalanced regression and extreme value prediction. *Mach. Learn.* **109**, 1803–1835 (2020).
9. Jafarigol, E. & Trafalis, T. B. *A Review of Machine Learning Techniques in Imbalanced Data and Future Trends*. (2023).
10. Jaffe, D. A. et al. Wildfire and prescribed burning impacts on air quality in the United States. *Journal of the Air and Waste Management Association* **70**, 583–615 https://doi.org/10.1080/10962247.2020.1749731Preprint at (2020).
11. Noor, N. M., Deak, G., Ul-Saufie, A. Z., Mohd, Z. & Rozainy, R. *Modeling of Particulate Matter (PM10) during High Particulate Event (HPE) in Klang Valley, Malaysia*. www.ijcs.ro (2022).
12. Branco, P., Ribeiro, R. P., Torgo, L., Krawczyk, B. & Moniz, N. SMOGN: a Pre-processing Approach for Imbalanced Regression. in *Proceedings of Machine Learning Research*. **74**, 36–50 (2017).
13. Torgo, L., Ribeiro, R. P., Pfahringer, B. & Branco, P. SMOTE for regression. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8154 LNAI 378–389 (2013).
14. Avelino, J. G., Cavalcanti, G. D. C. & Cruz, R. M. O. Resampling strategies for imbalanced regression: a survey and empirical analysis. *Artif Intell. Rev.* **57**, 1-42 (2024).
15. Wang, C., Deng, C., Wang, S. & Imbalance-XGBoost Leveraging Weighted and Focal Losses for Binary Label-Imbalanced Classification with XGBoost. http://arxiv.org/abs/1908.01672 (2019).
16. Liu, X. & Tian, H. Research on Imbalanced Data Regression Based on Confrontation. *Processes.* 12, (2024).
17. Branco, P., Torgo, L. & Ribeiro, R. P. Pre-processing approaches for imbalanced distributions in regression. *Neurocomputing* **343**, 76–99 (2019).
18. Branco, P. et al. REBAGG: Resampled Bagging for Imbalanced Regression. in *Proceedings of Machine Learning Research*. **94** 67–81 (2018).
19. Moniz, N., Ribeiro, R., Cerqueira, V. & Chawla, N. SMOTEBoost for regression: Improving the prediction of extreme values. in *Proceedings – 2018 IEEE 5th International Conference on Data Science and Advanced* Analytics, *DSAA* 150–159 (Institute of Electrical and Electronics Engineers Inc., 2018). 150–159 (Institute of Electrical and Electronics Engineers Inc., 2018) https://doi.org/10.1109/DSAA.2018.00025 (2018).
20. Silva, A., Ribeiro, R. P. & Moniz, N. Model Optimization in Imbalanced Regression. in *International Conference on Discovery Science* https://doi.org/10.48550/arXiv.2206.09991 (2022).
21. Felix, E. A. & Lee, S. P. Systematic literature review of preprocessing techniques for imbalanced data. *IET Software* **13**, 479–496 https://doi.org/10.1049/iet-sen.2018.5193 Preprint at (2019).
22. Torgo, L. & Ribeiro, R. Predicting rare extreme values. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 3918 LNAI 816–820 (2006).
23. Fonseca, J. & Bacao, F. Geometric SMOTE for imbalanced datasets with nominal and continuous features. *Expert Syst. Appl.* **234**, 121053 (2023).
24. Yang, Y., Zha, K., Chen, Y. C., Wang, H. & Katabi, D. Delving into Deep Imbalanced Regression. inInternational Conference on Machine Learning. https://doi.org/10.48550/arXiv.2102.09554 (2021).
25. Kuffner, T. A., Lee, S. M. S. & Young, G. A. Block bootstrap optimality and empirical block selection for sample quantiles with dependent data. *Biometrika Trust*. **103**, 1–18 (2018).
26. Radovanov, B. & Marcikic, A. A comparison of four different block bootstrap methods. *Croatian Oper. Res. Rev.* **5**, 189–202 (2014).
27. Burhanuddin, D. Shaadan. Controlled sampling approach in improving multiple imputation for missing seasonal rainfall data. *Res. Sq*. https://doi.org/10.21203/rs.3.rs-679692/v1 (2021).
28. Mader, M., Mader, W., Sommerlade, L., Timmer, J. & Schelter, B. Block-bootstrapping for noisy data. *J. Neurosci. Methods*. **219**, 285–291 (2013).
29. Ebtehaj, M., Moradkhani, H. & Gupta, H. V. Improving robustness of hydrologic parameter Estimation by the use of moving block bootstrap resampling. *Water Resour. Res* **46**, W07515 (2010).
30. Vogel, R. M. *The Moving Blocks Bootstrap versus Parametric Time Series Models* (1996).
31. Chen, X., Gupta, L. & Tragoudas, S. Improving the forecasting and classification of extreme events in imbalanced time series through block resampling in the joint Predictor-Forecast space. *IEEE Access.* **10**, 121048–121079 (2022).
32. Torgo, L. & Ribeiro, R. *LNAI 4702 - Utility-Based Regression* (in (Springer, 2007). https://doi.org/10.1007/978-3-540-74976-9_63
33. Ayus, I., Natarajan, N. & Gupta, D. Comparison of machine learning and deep learning techniques for the prediction of air pollution: a case study from China. *Asian J. Atmospheric Environment.* **17**, 48732–48745 (2023).
34. Dao, T. H. et al. PTI,. Analysis and Prediction for Air Quality Using Various Machine Learning Models. in *Proceedings of the Seventh International Conference on Research in Intelligent and Computing in Engineering*. **33** 89–94 (2023).
35. Verma, A., Ranga, V. & Vishwakarma, D. K. Combating Respiratory Health Issues with Intelligent NO2 Level Prediction from Sentinel 5P Satellite. in *IEEE 20th India Council International Conference, INDICON 2023* 882–886 (Institute of Electrical and Electronics Engineers Inc., 2023). 882–886 (Institute of Electrical and Electronics Engineers Inc., 2023) https://doi.org/10.1109/INDICON59947.2023.10440910 (2023).
36. Azid, A. et al. Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in Malaysia. *Water Air Soil. Pollut.* **225**, 1-14 (2014).
37. Guo, Q., He, Z. & Wang, Z. The characteristics of air quality changes in Hohhot City in China and their relationship with meteorological and Socio-economic factors. *Aerosol Air Qual. Res.* **24**, 230274 (2024).
38. Kumar, A. S. et al. wiley,. Recent Developments of Bioethanol Production. in *Bioenergy Research: Evaluating Strategies for Commercialization and Sustainability* 175–208 (2021). https://doi.org/10.1002/9781119772125.ch9

39. Syed, A. et al. Spatial and Temporal air quality pattern recognition using environmetric techniques: A case study in Malaysia. *Environ. Sciences: Processes Impacts*. **15**, 1717–1728 (2013).
40. Latif, M. T. et al. Long term assessment of air quality from a background station on the Malaysian Peninsula. *Sci. Total Environ.* **482–483**, 336–348 (2014).
41. Khadijah Arafin, S., Ul-Saufie, Z., Azura Md Ghani, A., Ibrahim, N. & Alam, S. N. *Feature selection methods using RBFNN based on enhance air quality prediction: insights from Shah Alam. IJACSA) Int. J. Adv. Comput. Sci. Applications.* **15,** 509-514 (2024).
42. Noor, N. M., Bakri Abdullah, A., Yahaya, M. M. & Ramli, N. A. A. S.Trans Tech Publications Ltd,. Comparison of linear interpolation method and mean method to replace the missing values in environmental data set. in *Materials Science Forum.* **803**, 278–281 (2015).
43. Libasin, Z., Ul-Saufie, Z. & Hasfazilah, A. A. identifying missing data mechanisms among incomplete air pollution datasets in Malaysia. https://doi.org/https://doi.org/10.1007/978-3-031-43922-3_18 doi:https://doi.org/10.1007/978-3-031-43922-3_18. (2024).
44. Malaysian Meteorological Department. Malaysia's climate. *Malaysian Meteorological Department* (2025).
45. Srivastava, C., Singh, S. & Singh, A. P. Estimation of air pollution in Delhi using machine learning techniques. in. *International Conference on Computing, Power and Communication Technologies, GUCON 2018* 304–309 (Institute of Electrical and Electronics Engineers Inc., 2019) https://doi.org/10.1109/GUCON.2018.8675022 (2018).
46. Kumar, A. & Goyal, P. Forecasting of air quality index in Delhi using neural network based on principal component analysis. *Pure Appl. Geophys.* **170**, 711–722 (2013).
47. Ditrich, J. *Data representativeness problem in credit scoring. ACTA OECONOMICA PRAGENSIA*. **23,** 1-17 (2015).
48. Verma, A., Ranga, V. & Vishwakarma, D. K. Forecasting of Satellite Based Carbon-Monoxide Time-Series Data Using a Deep Learning Approach. in *International Conference on Innovative Trends in Information Technology, ICITIIT 2023* (Institute of Electrical and Electronics Engineers Inc., 2023). (Institute of Electrical and Electronics Engineers Inc., 2023) https://doi.org/10.1109/ICITIIT57246.2023.10068609 2023).
49. Crone, S. F., Lessmann, S. & Stahlbock, R. Utility based data mining for time series analysis - Cost-sensitive learning for neural network predictors. in *Proceedings of the 1st International Workshop on Utility-Based Data Mining, UBDM '05* 59–68 https://doi.org/10.1145/1089827.1089835 (2005).
50. Moniz, N., Branco, P., Torgo, L. & Krawczyk, B. *Evaluation of Ensemble Methods in Imbalanced Regression Tasks. Proceedings of Machine Learning Research* 74 http://www.kdd.org/kdd-cup (2017).
51. Mignani, S. & Rosa, R. *The moving block bootstrap to assess the accuracy of statistical estimates in Ising model simulations. Computer Phys. Communications.* **92,** 203-213 (1995).
52. Sroka, Ł. Applying block bootstrap methods in silver prices forecasting. *Econometrics* **26**, 15–29 (2022).
53. Radovanov, B. & Marcikic, A. Testing the performance of the investment portfolio using block bootstrap method. *Economic Themes*. **52**, 166–183 (2014).
54. Martínez-Munoz, G., Bentejac, C. & Csorg, O. B. Gonzalo Martínez-Munoz, A. *A Comparative Analysis of XGBoost*. https://www.researchgate.net/publication/337048557 (2019).
55. Shahani, N. M., Zheng, X., Liu, C., Hassan, F. U. & Li, P. Developing an XGBoost regression model for predicting young's modulus of intact sedimentary rocks for the stability of surface and subsurface structures. *Front Earth Sci. (Lausanne)* **9**, 761990 (2021).
56. Jing, H. & Wang, Y. Research on Urban Air Quality Prediction Based on Ensemble Learning of XGBoost. in *E3S Web of Conferences* 165EDP Sciences, (2020).
57. Kumar, K. & Pande, B. P. Air pollution prediction with machine learning: a case study of Indian cities. *Int. J. Environ. Sci. Technol.* **20**, 5333–5348 (2023).
58. Nguyen, A. T., Pham, D. H., Oo, B. L., Ahn, Y. & Lim, B. T. H. Predicting air quality index using attention hybrid deep learning and quantum-inspired particle swarm optimization. *J Big Data.* **11**, 49-58(2024).
59. Chen, T., Guestrin, C. & XGBoost: A Scalable Tree Boosting System. https://doi.org/10.1145/2939672.2939785 doi:10.1145/2939672.2939785 (2016).
60. Mienye, I. D., Sun, Y. A. & Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. *IEEE Access* **10**, 99129–99149 https://doi.org/10.1109/ACCESS.2022.3207287 Preprint at (2022).
61. Pan, B. Institute of Physics Publishing,. Application of XGBoost algorithm in hourly PM2.5 concentration prediction. in *IOP Conference Series: Earth and Environmental Science.* 113 (2018).
62. Abdullah, S., Ismail, M., Ahmed, A. N. & Abdullah, A. M. Forecasting particulate matter concentration using linear and non-linear approaches for air quality decision support. *Atmosphere (Basel).* **10**, 1-24 (2019).
63. Shaziayani, W. N., Ul-Saufie, Z., Ahmat, H. & Al-Jumeily, D. Ahmad, Coupling of quantile regression into boosted regression trees (BRT) technique in forecasting emission model of PM 10 concentration. https://doi.org/10.1007/s11869-021-01045-3/Published (2021).
64. DOE. Department of Environment:Malaysia Quality Report 2016. Kuala Lumpur: Ministry of Energy, Science, Technology, Environment and Climate Change, Malaysia. (2016).
65. Mohd Shafie, S. H. et al. Influence of urban air pollution on the population in the Klang Valley, malaysia: a Spatial approach. *Ecol Process.* **11**, 1-16 (2022).
66. DOE. Department of Environment:Malaysia Environmental Quality Report. Kuala Lumpur: Ministry of Energy, Science, Technology, Environment and Climate Change, Malaysia. (2021). (2021).
67. Rahim, N. A. A. A. et al. Institute of Physics,. Predicting Particulate Matter (PM10) during High Particulate Event (HPE) using Quantile Regression in Klang Valley, Malaysia. in *IOP Conference Series: Earth and Environmental Science*. 1216 (2023).
68. Ren, J., Zhang, M., Yu, C. & Liu, Z. Balanced MSE for Imbalanced Visual Regression. in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* vols -June 2022 7916–7925 (IEEE Computer Society, 2022).

Acknowledgements

We would like to express our sincere thanks to Faculty of Computer Science and Mathematics and 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.

Author contributions

M.M, A.Z.U.S performed the method development, data analysis, and writing the manuscript. N.F.A.R and A.G provided guidance and reviewed the manuscript N.M.N reviewed the manuscript. All authors reviewed and approved the final version of the manuscript.

Funding

This research was funded by the Malaysian Ministry of Higher Education, grant number UiTM.800-3/2FRGS

(003/2025).

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to A.Z.U.-S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.