# Evaluating the Performance of Tree-Based Model in Predicting Haze Events in Malaysia

Mahiran Muhammad, Ahmad Zia Ul-Saufie*, Fadhilah Ahmad Radi
Faculty of Computer Science and Mathematics, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia

***Abstract*—Predicting haze is crucial in controlling air pollution to reduce its impact, especially on human health. Accurate prediction of extreme values is vital to raising public awareness of this issue and better understanding of air quality management. Extreme values in air pollution refer to unusually high measurements of pollutants that diverge significantly from the normal range of observed values. Extreme values are normally caused by haze from various factors. Neglecting extreme values can cause unreasonable predictions. Therefore, this study aims to evaluate the performance of a tree-based algorithm in predicting haze events. Predictive analytics were based on hourly air pollution data from 2013 to 2022 in Shah Alam, Malaysia. The ten parameters are chosen Relative Humidity (RH), Temperature (T), Wind Direction (WD), Wind Speed (WS), $PM_{10}$, NOx, $NO_2$, $SO_2$, $O_3$ and CO. Decision Tree (DT), Gradient Boosting Regression (GBR) and Extreme Gradient Boosting (XGBoost) are compared in determining the best approach for modeling $PM_{10}$ concentrations for the next 24 hours ($PM_{10,t+24h}$) for overall air quality data and three air quality blocks: Good air quality (Block 1), Moderate air quality (Block 2) and Extreme air quality (Block 3). The performance of RMSE, MAE and MAPE indicate that XGBoost outperforms GBR and DT with the RMSE(21.5921), MAE(14.2396) and MAPE(0.4816). When evaluating the performance across the three air quality blocks, XGBoost remains as the top-performing model. However, XGBoost faces challenges in accurately predicting extreme values.**

***Keywords*—Extreme Gradient Boosting (XGBoost); Gradient Boosting Regression (GBR); Decision Tree (DT), extreme values; Particulate Matter (PM)***

## I. Introduction

Malaysia has experienced air pollution, which creates an environmental health threat. Air pollution can lead to a variety of critical illnesses in humans, including bronchitis, heart disease, pneumonia and lung cancer [1]. Bad air quality gives rise to other current environmental issues like global warming, acid rain, reduced visibility, smog, aerosol formation, climate change and premature death [2]. Primary pollutants that impact on most countries constitute Particulate Matter (PM), Nitrogen Dioxide ($NO_2$), Carbon Monoxide (CO), Ozone ($O_3$), Sulphur Dioxide ($PM_{10}$) [3]. Particulate Matter (PM) are notable pollutant within the air and it has a greater effect on human beings compared to other pollutants. Two types of particulate matter are $PM_{10}$ and $PM_{2.5}$. The value of the $PM_{10}$ concentration usually represents the API in Malaysia. This is because $PM_{10}$ concentration in Malaysia is always higher than any other pollutants [4] [5]. Monitoring and predicting $PM_{10}$ concentration, especially in urban areas, has become a vital and challenging task with increasing motor and industrial developments.

According to study [6], extreme values are defined as events that occur less frequently than common events. Extreme values in air pollution refer to unusually high pollutant measurements that deviate significantly from the normal range. An uncommonly high $PM_{10}$ concentration level can result in the existence of extreme values in air pollution data. These anomalies are typically caused by haze events, coming from wildfires, industrial accidents, temperature inversion and are sometimes caused by measurement errors. There is research that demonstrates the meteorological influence on air quality [7]. The $PM_{10}$ concentrations and CO were found to have a strong to moderate correlation when episodic haze was recorded. Meanwhile, the relationship of $PM_{10}$ level with $SO_2$ was found to be significant in 2013 and negatively correlated with relative humidity (RH) [8]. Weak correlation between $PM_{10}$ and NOx was measured in study areas, likely because of low contribution of domestic artificial sources towards haze events in Malaysia [8]. Neglecting extreme values can cause unreasonable predictions when using the original data set directly. Therefore, it is fundamental to precisely cope with the problem of extreme values to boost the effectiveness of prediction models.

Various predictive models, spanning from statistical approaches to machine learning methods, have been employed to forecast $PM_{10}$ concentrations [9]. Several researchers have applied and developed machine learning models to predict $PM_{10}$ concentrations in Malaysia [12] [11] [14] [13] [15]. Based on the outcomes of all the ML models, machine learning effectively addresses the challenges of nonlinear and complex models in predicting $PM_{10}$ concentrations. Predictive models based on machine learning (ML) are more accurate and consistent [10].

Globally, several tree-based machine learning models have demonstrated high accuracy in predicting $PM_{10}$ concentrations compared with other machine learning models. These tree-based ML models are often used for classification and regression tasks because they require less time to train and tune the model parameters [17] [18] [19] [16] [20]. Tree-based machine learning models, such as Decision Tree (DT), Random Forest (RF), Gradient Boosting Regression (GBR), and Extreme Gradient Boosting (XGBoost) have gained prominence in air quality research due to the interpretability, robustness and strong predictive performance [21]. These models are particularly well-suited for handling complex, non-linear relationships that often found in environmental datasets [16].

Several studies have demonstrated the effectiveness of tree-based models in predicting air pollution. A study comparing various machine learning models found that XGBoost achieved the highest R² value (0.9985) and the lowest error metrics,

[*] Corresponding author.

outperforming Lasso regression in forecasting $PM_{10}$ concentrations [22]. The study in [19] compare a linear forecast technique (multiple linear regression) with proposed non-linear algorithms, Random Forest, Support Vector Machine (SVR) and Gradient Boosting Regression (GBR). Their results revealed that GBR outperformed others to predict $PM_{10}$ concentrations. The study in [17] showed that XGBoost performs better than Light GBM in terms of prediction estimation with RMSE of 12.846, but it takes longer to train and tune the model's parameters. DT is one of the simplest, yet powerful models used in environmental modelling. Studies have demonstrated that DT models can efficiently capture local pollution patterns [25]. The study [23] stated that XGBoost outperforms other deep learning models due to the consideration of small sample size in datasets. The study [24] found ANN requires extensive tuning parameters and longer computational time and were found to be less effective as XGB and RF to predict air pollution.

Numerous research studies have been conducted in comparing different machine learning methods in air pollution prediction. However, limited research has been conducted on comparing the performance of tree-based machine learning models in predicting extreme events of $PM_{10}$ concentrations across different air pollution levels. Therefore, this research is motivated by the intention to evaluate the performance of the DT, GBR and XGBoost models in determining the best approach in predicting extreme or haze events of $PM_{10,t+24h}$ concentrations. Three air quality blocks: $PM_{10,24h}$ concentration $\leq 50\,\mu g/m^3$ (Block 1, Good air quality status), $50\,\mu g/m^3 < PM_{10,24h}$ concentration $\leq 150\,\mu g/m^3$ (Block 2, moderate air quality status), and $PM_{10,24h}$ concentration $> 150\,\mu g/m^3$ (Block 3, extreme air quality status) will be compared to assess their impact on predicting haze events in air pollution data. In this study, the primary focus is on Block 3, which is characterized as extreme haze events when the $PM_{10}$ concentrations exceed $150\mu g/m^3$ In summary, the key contribution of this research work is a comparison and evaluation of the proposed machine learning model for predicting extreme or haze events in Malaysia. This paper is organized as follows: I. Introduction, II. Methodology, III. Results and Discussion. Followed by the conclusion in Section IV and the reference list.

## II. Methodology

### A. Research Flow

Fig. 1 presents a flowchart outlining this study to evaluate the performance of a tree-based algorithm in predicting haze events of $PM_{10}$ concentrations in Shah Alam, Malaysia. This study utilizes air quality data from 2013 to 2022 provided by the Department of Environment, Malaysia. The process begins with data extraction, followed by data preprocessing. Data extraction is the first step in the process, which is then followed by extensive data pre-processing. Next, the air quality data are then used to train the DT, GBR and XGBoost. The model performance is then evaluated based on the accuracy of RMSE, MAE and MAPE. Finally, the best model that provides the most accurate prediction for extreme values is identified.

### B. Data Description

This study obtained secondary data from the Malaysian Department of Environment (DOE) from 2013 to 2022. The stations are situated in Shah Alam, Selangor, and consist of 83,431 air quality data points for 10 variables, such as air pollutants and meteorological parameters. The air pollutants included: $PM_{10}$, $SO_2$, $NO_x$, $NO_2$, $O_3$ and CO, whereas meteorological parameters included: WS, WD, RH and T. Table I shows the variable in this study with their respective level of measurements and role. $PM_{10}$ concentration for the next 24 hours, $PM_{10,t+24h}$ serve as dependent variable. Meanwhile, the other variables serve as independent variables.

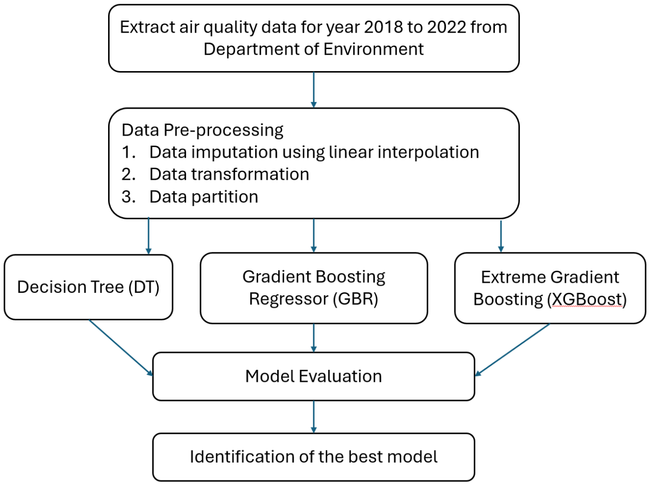
Extract air quality data for year 2018 to 2022 from Department of Environment
Data Pre-processing
1. Data imputation using linear interpolation
2. Data transformation
3. Data partition
Decision Tree (DT)
Gradient Boosting Regressor (GBR)
Extreme Gradient Boosting (XGBoost)
Model Evaluation
Identification of the best model

Fig. 1. Research flow.

TABLE I. Description of Variables

| Variable | Description | Level of measurement | Role |
|---|---|---|---|
| **Particulate Matter for the next 24h ($PM_{10,t+24h}$)** | Hourly concentration of $PM_{10,t+24h}$ (g/m$^3$) | Interval | Dependent |
| **Particulate Matter 10 ($PM_{10}$)** | Hourly concentration of $PM_{10}$ (g/m$^3$) | Interval | Independent |
| **Sulphur Dioxide ($SO_2$)** | Hourly concentration of Sulphur dioxide (ppb) | Interval | Independent |
| **Nitric Oxide and Nitrogen Dioxide ($NO_x$)** | Hourly concentration of nitric oxide and nitrogen dioxide (ppb) | Interval | Independent |
| **Nitrogen Dioxide ($NO_2$)** | Hourly concentration of nitrogen dioxide (ppb) | Interval | Independent |
| **Ozone ($O_3$)** | Hourly concentration of ozone (ppb) | Interval | Independent |
| **Carbon Monoxide (CO)** | Hourly concentration of carbon monoxide (ppb) | Interval | Independent |
| **Wind Speed (WS)** | Hourly wind speed (m/s) | Interval | Independent |
| **Wind Direction (WD)** | Hourly wind direction (°) | Interval | Independent |
| **Relative Humidity (RH)** | Hourly relative humidity (%) | Interval | Independent |
| **Ambient Temperature (T)** | Hourly temperature (°c) | Interval | Independent |

In this study, the dataset was segmented based on the Air Pollution Index (API) by [49]. Table II shows the API level, which is categorised as good, moderate, unhealthy, very unhealthy and hazardous, which can be of air quality management level or decision making for data interpretation processes. API is an effortless and encompassing technique for defining air quality conditions that is easily understood [5]. It is categorised based on the highest values of five main air pollutants. Meanwhile, Table III shows the calculation of the breakpoint concentration for $PM_{10}$ corresponding to each API category. For example, a $PM_{10}$ concentration between 50 μg/m$^3$ and 150 μg/m$^3$ falls into the API category of 51-100 (moderate air quality). Therefore, the air quality blocks are categorised into three blocks according to the breakpoint of $PM_{10}$ concentrations established in Malaysia. These blocks are defines as follows: For Good air quality status where $PM_{10,24h}$ concentration $\leq 50\,\mu g/m^3$ served as Block 1, for moderate air quality status which is $50\,\mu g/m^3 < PM_{10,24h}$ concentration $\leq 150\,\mu g/m^3$ served as Block 2, and for extreme air quality status which is $PM_{10,24h}$ concentration $> 150\,\mu g/m^3$ served as Block 3. The performance of each block is then compared to evaluate their influence on the prediction of extreme events in air pollution data.

TABLE II. THE API INDEX [49]

| API Range | Air Quality Status |
|---|---|
| 0–50 | Good |
| 51–100 | Moderate |
| 101–200 | Unhealthy |
| 201–300 | Very Unhealthy |
| > 300 | Hazardous |

### C. Data Preprocessing

Data preprocessing encompasses missing value imputation, data transformation and data partition. Missing values can originate from multiple sources, such as sensor malfunctions, environmental factors, or data transmission errors [26]. A high proportion of missing data may lead to biases or weaken the statistical power of the analysis [26]. According to study [27], Malaysian missing air pollution data belong to Missing at Random (MAR) and the linear interpolation method assumes that the pattern of missingness does not disrupt the underlying trends in the data. In most air pollution data, Linear interpolation is the most ordinary imputation method to treat short gaps of missing data in the air pollution dataset [13]. For data transformation, the units of air pollutants $SO_2$, $NO_2$, $NO_x$, $O_3$, and CO in ppm need to be converted to ppb since the ppm unit is too small, thus affecting the accuracy of the results. The WD variable, which is expressed in degrees, has been converted to wind direction index (dimensionless) [28]). According to study [10], the formula for conversion is

$$\text{Wind Direction Index (WDI)} = 1 + \sin(\theta - 45^\circ) \qquad (1)$$

In this study, data partition is conducted by employing splitting methods. Two subsets of the dataset are selected, with 80% ($n$ = 65,945) of the data going to training and 20% ($n$ = 16,486) to testing (80% for model development and 20% to measure the performance of the model). According to [29] , empirical studies show that the optimal results are achieved if 80% of the data is allocated for training and 20% is used for testing. Random sampling is applied to partition the data into train and test sections [30].

### D. Machine Learning Model

This part gives a brief introduction to DT, GBR and XGBoost. In this study, the machine learning model was used to evaluate the performance in predicting haze events. The general model for each machine learning model are shown in Table IV. The table shows the general model for each tree-based algorithms model, DT, GBR and XGBoost. The general model shows the prediction for $PM_{10}$ concentration for the next 24 hours. t represents time.

*1) Decision Tree (DT):* Decision Tree (DT) is a well-known machine learning model that falls under the category of supervised learning [31]. DT can be used for both classification and regression problems [32]. Additionally, DT can effectively handle both numeric and nominal data formats [33]. It constructs a tree-like structure of decisions and their potential outcomes, starting with a root node representing the entire dataset and branching into multiple internal and leaf nodes [25]. Each internal node represents a decision or feature test, and the edges leaving that node represent the possible outcomes of the test [33]. The path from the root node to the leaf node indicates a collection of decisions that leads to a prediction for each given sample [34]. The tree is constructed by recursively splitting the dataset based on the Mean Square Error (for regression trees) that provides the lowest variance [34]. To determine the optimal split, this algorithm applies the usual variance formula.

$$\text{Mean Squared Error} = \frac{\sum (y_i - \hat{y}_i)^2}{n} \qquad (2)$$

where, $y_i$ is the actual PM concentration for the next 24 hours and $\hat{y}_i$ is the predicted PM concentration for the next 24 hours.

DT are fast and easy to understand. However, the model tends to overfit if the tree is allowed to grow too deep or if there are many noisy features in the data [35].

*2) Gradient Boosting Regressor (GBR):* The Gradient Boosting Regressor (GBR) is a machine learning for regression or classification that provides better prediction models in the form of ensemble weak prediction models [36]. GBR is another ensemble model that is an iterative collection of sequentially ordered tree models so that the following model learns from the error of the previous model [37]. This machine learning approach provides predictions by 'boosting' the ensemble of weak prediction models, usually decision trees, to form a more robust model [38]. The objective function of GBR is described by study [39] as:

$$\hat{F}(x) = \arg\min \sum_{i=1}^{N} L(y_i, \hat{y}_i) \qquad (3)$$

$$L = \frac{1}{2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (4)$$

where, $y_i$ is the actual PM concentration for the next 24 hours and $\hat{y}_i$ is the predicted PM concentration for the next 24 hours. Meanwhile, L denotes the loss function.

TABLE III. BREAKPOINT OF $PM_{10}$ CONCENTRATION LEVELS [5]

| Breakpoint of Concentration | Air Quality Status | Equation for API |
|---|---|---|
| Conc. $\leq 50$ | Good | API = conc. |
| $50 <$ conc. $\leq 150$ | Moderate | API = 50 + (conc. − 50) × 0.5 |
| $150 <$ conc. $\leq 350$ | Unhealthy | API = 100 + (conc. − 150) × 0.5 |
| $350 <$ conc. $\leq 420$ | Very unhealthy | API = 200 + (conc. − 350) × 1.4286 |
| $420 <$ conc. $\leq 500$ | Hazardous | API = 300 + (conc. − 420) × 1.25 |
| Conc. $> 500$ | Emergency | API = 400 + (conc. − 500) |

TABLE IV. GENERAL MODEL FOR EACH ML MODEL

| Machine learning model | General model |
|---|---|
| Decision Tree (DT) | $PM_{10,t+24h}$ $\tilde{D}$T ($PM_{10,t}$, $SO_{2,t}$, $NO_{2,t}$, $NO_{x,t}$, $O_{3,t}$, $CO_t$, $WS_t$, $WDI_t$, $RH_t$, $T_t$) |
| Gradient Boosting Regression (GBR) | $PM_{10,t+24h}$ $\tilde{G}$BR ($PM_{10,t}$, $SO_{2,t}$, $NO_{2,t}$, $NO_{x,t}$, $O_{3,t}$, $CO_t$, $WS_t$, $WDI_t$, $RH_t$, $T_t$) |
| Extreme Gradient Boosting (XGBoost) | $PM_{10,t+24h}$ $\tilde{X}$GB ($PM_{10,t}$, $SO_{2,t}$, $NO_{2,t}$, $NO_{x,t}$, $O_{3,t}$, $CO_t$, $WS_t$, $WDI_t$, $RH_t$, $T_t$) |

A GBR with M number of trees can be stated as;

$$f_M(x_j) = \sum_{m=1}^{M} \gamma_m h_m(x_j) \tag{5}$$

where, $h_m$ is a weak learner that performs poorly individually and $\gamma_m$ is a scaling factor adding the contribution of a tree to the model.

GBR employs the gradient descent loss function to minimize errors by updating the starting estimation with a new estimation [40]. Thus, a final model is created by combining all preliminary estimations with appropriate weights [40].

*3) Extreme Gradient Boosting (XGBoost):* XGBoost is a decision tree ensemble based on gradient boosting that is highly scalable [39]. XGBoost is a powerful approach for developing supervised regression models [41]. The validity of this statement can be determined by deliberating about the objective function and base learners of XGBoost [41].

The objective function consists of a loss function and a regularization term [42]. Like gradient boosting, XGBoost develops an additive expansion of the objective function by minimizing a loss function [42]. XGBoost is one of the ensemble learning approaches that involves training and combining individual models (known as base learners) to produce a single prediction [43]. Considering that XGBoost is focused only on decision trees as base classifiers, a variation of the loss function is used to control the complexity of the trees [39]. Unlike gradient boosting, the XGBoost objective function includes a regularization term to avoid overfitting [39].

Assume that a dataset, D is $\{(x_i, y_i) : i = 1, \ldots, n\}$. Let $\hat{y}_i$ be defined as a result given by an ensemble represented by the generalized model as follows (Pan, 2018)

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i) \tag{6}$$

where $f_k$ is a regression tree, $f_k(x_i)$represents t he score given by the $k$-th tree to the $i$-th observation in the data.

To functions fk, the following regularized objective function should be minimized:

$$\text{Obj} = L(\phi) = \sum_i L(\hat{y}_i) + \sum_k \Omega(f_k) \tag{7}$$

where, $L$ is the custom loss function. The loss function $L$ is a differentiable convex loss function that measures the difference between the prediction $\hat{y}_i$ and the observation $y_i$ [42]. This loss function can be integrated into the split criterion of decision trees, leading to a pre-pruning strategy.

To prevent too large a complexity of the model, the penalty term or regularization term $\Omega$ is included as follows:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda w^2 \tag{8}$$

where, $\lambda$ and $\gamma$ are the parameters controlling penalty for the number of leaves T and magnitude of leaf weights w respectively. The purpose of $\Omega(f_k)$ is to prevent over-fitting and to simplify models produced by this algorithm [44]. The additional regularization term helps to smooth the final learnt weights to avoid overfitting [44].

*4) Parameter setting for baseline model evaluation:* All models were built using general parameter settings as provided by the respective libraries (scikit-learn) to evaluate and compare performance foundations for DT, GBR and XGBoost in predicting $PM_{10}$ concentrations during haze events in Malaysia. This technique gives an unbiased and consistent comparison among the models, emphasizing their inherent learning capabilities despite the effect of tuning. Furthermore, utilizing general parameters provides a clear baseline for future tuning operations and represents common practices in preliminary model evaluation. Using general parameter setting provides comparable performance to adjusted models, suggesting that general parameter settings can be an appropriate starting point for model evaluation [45].

Eventhough DT, GBR and XGBoost are all tree-based models, they do not share the same general parameters. Each algorithm is based on different concepts, and their application represents these differences. The DT from scikit-learn constructs a single decision tree using all available attributes, with no limitations on depth `max_depth=None` by general. This enables the tree to develop as deep as needed to fit the training data, potentially cause to overfitting if the data is noisy. Other important setting includes `min_samples_split=2` and min_samples_leaf=1 which manage the tree's branching criteria. In contrast, the GBR builds an ensemble of short decision trees in stages. By general parameter setting, it uses max_depth=3 to limit the complexity of each tree, along with a learning_rate=0.1 to figure out how much each tree serves to the final model. This restrictive configuration is created to avoid overfitting while maintaining flexibility.

XGBoost Regressor is also an ensemble method based on boosting but utilizes more aggressive parameter setting compared to scikit-learn's gradient boosting. It sets max_depth=6, learning_rate=0.3, and includes additional parameters such as min_child_weight=1, subsample=1, and colsample_bytree=1. These settings attempt to achieve a balance between speed and accuracy, with built-in regularization capabilities. While all three models fall under the category of tree-based methods, their general parameters differ significantly due to their mechanical nature. Recognizing these variations is essential when performing baseline effectiveness in prediction tasks.

### *E. Model Performance*

In predicting $PM_{10}$ concentrations, proper model evaluation is essential. According to a previous study by [46], 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. Three statistical evaluations will be used to evaluate the model performance: Root Mean Square Error (RMSE), Mean Square Error (MSE) and Mean Absolute Percentage Error (MAPE). 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 least value of each statistical evaluation. The criteria formulas are shown below:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2} \tag{9}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{Y}_i - Y_i| \tag{10}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i - \hat{Y}_i|}{Y_i} \tag{11}$$

where, $n$ is the total number of hourly measurements of a particular station, $\hat{Y}_i$ is the estimated value of $PM_{10,t+24h}$, $Y_i$ is the observed value of $PM_{10,t+24h}$ and $\bar{Y}_i$ is the mean of the observed value of $PM_{10,t+24h}$

## III. Results and Discussion

The descriptive statistics and boxplot for maximum hourly $PM_{10}$ concentrations in Shah Alam from 2013 to 2022 are shown in Table V and Fig. 2. The box plot in Fig. 2 visualizes the distribution of $PM_{10,t+24h}$ concentrations across different years (2013-2022). The boxed interquartile range (IQR) in 2013-2016 are larger, meaning higher variability in $PM_{10}$ concentrations. From 2017 onward, the boxes are smaller, indicating that $PM_{10}$ concentrations have become more stable. The boxplot indicates that Shah Alam experienced the highest $PM_{10}$ concentration in 2014, followed by 2013 with the second highest concentrations. Additionally, the year 2015 observed an increment in the number of extreme $PM_{10}$ concentration values. This is because Shah Alam serves as an industrial hub. Shah Alam hosts numerous manufacturing plants, factories and processing industries, leading to significant emissions of pollutants. Ongoing construction projects release dust and particulate matter into the air, adding to the pollution burden. In 2015, the extreme values were attributed to transboundary haze pollution [47]. A noticeable reduction in extreme $PM_{10}$ values occurred between 2016 to 2018. From 2016 to 2017, the overall air quality was generally good to moderate. Malaysia suffered moderate haze outbreaks in 2016 caused by localized and transboundary pollution, however, overall air quality improved throughout this year [48] [49] [50].

From the Table V, the mean concentration in Shah Alam for 10 years from 2013 to 2022 for 5 pollutants are $PM_{10}$ (41.7043 μgm$^{-3}$), CO (740.0108 ppb), $NO_2$ (18.3214 ppb), $SO_2$ (1.9875 ppb) and $O_3$ (19.4478 ppb). The concentrations of $PM_{10}$ were very high in Shah Alam, Selangor, with a maximum concentration of 575 μg/m$^3$. The skewness of $PM_{10}$ is 4.5730, indicating a highly positively skewed distribution, which shows the presence of extreme values in the data. The mean values for meteorological parameters are represented by Relative Humidity (RH) (78.1675%), Temperature (T) (27.8690$^o$c), wind direction (WD) (191.3082$^o$) and wind speed (WS) (2.7661 m/s). The mean values of $PM_{10}$ (41.7043) higher than the median (35.0000) indicates that the pollutant distributions are having right-skewed distribution. All the variables are positively skewed when the skewness presents positive values for each variable (CO=1.1710, $NO_2$=1.1650, $O_3$=1.3600, $PM_{10}$=4.5730, $SO_2$=6.8020, $NO_x$=1.6030, WS= 1.7360, T=0.7070) except for RH(-0.395) and WD(-0.0200), which shows negatively skewed when the skewness presents negative values. In summary, the box plot shows that extreme $PM_{10}$ concentration values occurred annually from 2013 to 2022. To accurately predict these extreme values, tree-based algorithms will be further analysed to evaluate the most effective model for forecasting haze events in Shah Alam. Fig. 3 shows the correlation heatmap in Shah Alam which represents the correlation coefficients between different air pollution parameters. The colour gradient ranges from green (strong negative correlation, -1) to blue (strong positive correlation, +1), with yellowish shades indicating weaker correlations. This heatmap provides valuable insights into the relationships between meteorological conditions and air pollutants. From the heatmap, there was a negative relationship between $PM_{10}$ and Relative Humidity (RH) (r=0.051) and a moderate correlation was found between $PM_{10}$ and CO in Shah Alam (r=0.46). This is supported by [8] when their study shows negative correlation between $PM_{10}$ and RH and strong to moderate correlation between $PM_{10}$ and CO. This suggest that as humidity increases $PM_{10}$ concentrations tend to decrease slightly. Understanding these correlations helps in air quality modelling, especially when predicting extreme pollution events.

Table VI shows a comparative analysis to evaluate the performance of the DT, GBR, and XGBoost models for hourly $PM_{10,t+24h}$ concentration predictions at Shah Alam station from the period 2013 to 2022. This table summarizes quantitatively the performance in terms of RMSE, MAE and MAPE. The

TABLE V. Descriptive Statistics of Air Pollution from 2013 to 2022 in Shah Alam, Selangor

| Parameter | Mean | Median | Standard deviation | Skewness | Kurtosis | Minimum | Maximum |
|---|---|---|---|---|---|---|---|
| **$PM_{10}$,24H ($\mu gm^{-3}$)** | 41.71 | 35.00 | 31.74 | 4.57 | 39.23 | 0.63 | 575.00 |
| **$PM_{10}$ ($\mu gm^{-3}$)** | 41.70 | 35.00 | 31.74 | 4.57 | 39.21 | 0.63 | 575.00 |
| **CO (ppb)** | 740.01 | 673.00 | 441.24 | 1.17 | 4.46 | 0.10 | 6550.00 |
| **$O_3$ (ppb)** | 19.45 | 11.90 | 20.92 | 1.36 | 1.73 | 0.10 | 161.00 |
| **$NO_2$ (ppb)** | 18.32 | 16.40 | 11.46 | 1.16 | 2.30 | 0.10 | 111.00 |
| **$SO_2$ (ppb)** | 1.99 | 1.30 | 2.17 | 6.80 | 122.10 | 0.10 | 100.00 |
| **$NO_x$ (ppb)** | 30.44 | 24.00 | 23.77 | 1.60 | 3.62 | 0.10 | 215.00 |
| **RH (%)** | 78.17 | 78.95 | 13.57 | -0.39 | -0.56 | 20.00 | 100.00 |
| **T ( $^o$c)** | 27.87 | 27.19 | 3.20 | 0.70 | -0.07 | 19.80 | 31.98 |
| **WD ( $^o$ )** | 191.31 | 192.26 | 82.07 | -0.02 | -0.28 | 0.05 | 317.93 |
| **WS (m/s)** | 2.77 | 1.25 | 3.26 | 1.74 | 2.59 | 0.02 | 24.30 |

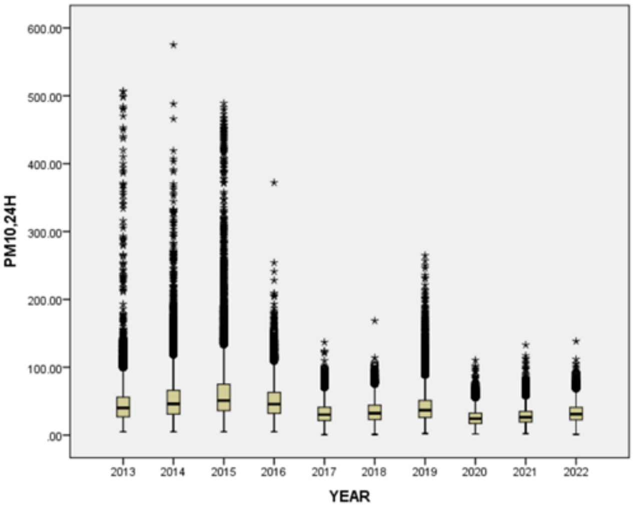

Fig. 2. Boxplot of PM10 concentrations in Shah Alam, Selangor.

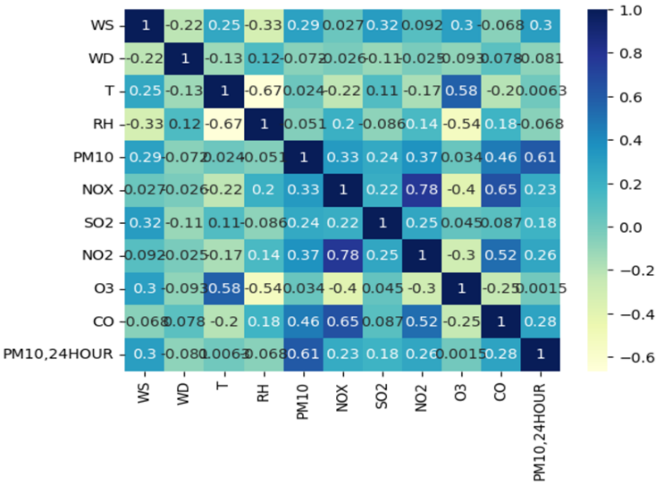

Fig. 3. The Correlation heatmap of air quality parameters in Shah Alam.

result shown are for overall air pollution data. According to the Table VI, it is observed that XGBoost outperforms all other models in the prediction of $PM_{10,t+24h}$ concentrations with the lowest value of RMSE, MAE and MAPE value. Meanwhile, DT generate the highest value of RMSE, MAE and MAPE value, indicating the DT is at lowest level of performance. This result aligns with the findings of [32] [44], which demonstrate that the XGBoost method is more effective than DT for predicting air pollution concentration. Fig. 4 presents the actual vs predicted graph for overall $PM_{10,t+24h}$ concentration using DT, GBR and XGBoost model. A) represents XGBoost model for the actual vs predicted $PM_{10,t+24h}$ concentration. B) represents GBR model for the actual vs predicted $PM_{10,t+24h}$ concentration and C) represents DT model for the actual vs. predicted $PM_{10,t+24h}$ concentration. From the graph, it can be concluded that XGBoost predictions are significantly more accurate than GBR and DT when the prediction line of XGBoost is closer to the actual line. In contrast, the GBR and DT models show a greater discrepancy, with their predicted lines deviating further from actual data. Therefore, for overall air pollution data, it can be concluded that XGBoost is the best model for $PM_{10,t+24h}$ air pollution predictions.

TABLE VI. Performance Results for Overall Air Quality

| Performance Evaluation | Model | | |
|---|---|---|---|
| | **XGBoost** | **GBR** | **DT** |
| RMSE | 21.5921 | 24.5051 | 24.7156 |
| MAE | 14.2396 | 15.7770 | 15.5915 |
| MAPE | 0.4816 | 0.5528 | 0.5164 |

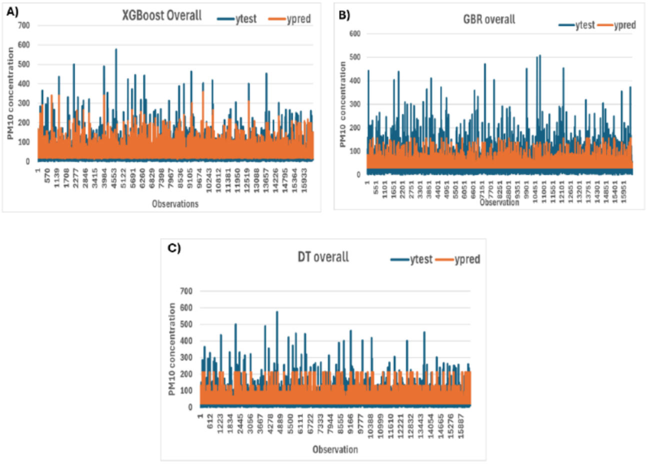

Fig. 4. Actual vs. Predicted for overall PM10 prediction A) XGBoost B) GBR C) DT.

The analysis is furthered through each block to evaluate the effectiveness of the tree-based model. From the result of the actual and predicted value of each model, the result is arranged and blocked through the actual values of the model. The three blocks of air quality data are $PM_{10,24h}$ concentration $\leq 50\,\mu g/m^3$ (Block 1), $50\,\mu g/m^3 < PM_{10,24h}$ concentration $\leq 150\,\mu g/m^3$ (Block 2), and $PM_{10,24h}$ concentration $> 150\,\mu g/m^3$ (Block 3). This analysis is extended to evaluate each model's ability to predict extreme events and determine whether they can accurately capture extreme values.

Table VII shows the results of the performance indicator by using the XGBoost, GBR and DT for the three blocks of air quality data. Based on the XGBoost results, the three blocks of air pollution data show that the $PM_{10,t+24h}$ concentration below 50μg/$m^3$ (Block 1) had the lowest RMSE value of 14.1875, compared to 27.2372 for $PM_{10,t+24h}$ concentrations between 50μg/$m^3$ and 150μg/$m^3$ (Block 2) and 106.0264 for $PM_{10,t+24h}$ concentrations above 150 μg/$m^3$ (Block 3) respectively. Similarly, the MAE is the lowest for the $PM_{10,t+24h}$ concentrations below 50μg/$m^3$ (Block 1) at 10.6636. While $PM_{10,t+24h}$ concentrations between 50μg/$m^3$ and 150μg/$m^3$ (Block 2) and those above 150 μg/$m^3$ (Block 3) recorded higher MAE values of 22.0516 and 83.4721, respectively.

GBR and DT also exhibited an increment in error measures, particularly for Block 3. For GBR, $PM_{10,t+24h}$ concentration below 50μg/$m^3$ (Block 1) had the lowest RMSE value of 14.1673, compared to 28.9145 for $PM_{10,t+24h}$ concentrations between 50μg/$m^3$ and 150μg/$m^3$ (Block 2) and 140.4993 for $PM_{10,t+24h}$ concentrations above 150 μg/$m^3$ (Block 3) respectively. Similarly, the MAE was the lowest for the $PM_{10,t+24h}$ concentrations below 50μg/$m^3$ (Block 1) at 11.3487. While $PM_{10,t+24h}$ concentrations between 50μg/$m^3$ and 150μg/$m^3$ (Block 2) and those above 150 μg/$m^3$ (Block 3) recorded higher MAE values of 23.9327 and 124.5067, respectively. Meanwhile For DT, $PM_{10,t+24h}$ concentration below 50μg/$m^3$ (Block 1) had the lowest RMSE value of 15.2412, compared to 30.6125 for $PM_{10}$ concentrations between 50μg/$m^3$ and 150μg/$m^3$ (Block 2) and 131.3885 for $PM_{10,t+24h}$ concentrations above 150 μg/$m^3$ (Block 3) respectively. Similarly, the MAE was the lowest for the $PM_{10,t+24h}$ concentrations below 50μg/$m^3$ (Block 1) at 11.3487. While $PM_{10,t+24h}$ concentrations between 50μg/$m^3$ and 150μg/$m^3$ (Block 2) and those above 150 μg/$m^3$ (Block 3) recorded higher MAE value of 24.5156 and 108.5374, respectively. From this table, XGBoost demonstrates the best performance compared to DT and GBR, as it achieves the lowest RMSE, MAE, and MAPE values. However, from this table, the key observation is Block 3, where the $PM_{10,t+24h}$ concentration exceeds 150μg/$m^3$. This indicates that Block 3 experiences significantly higher pollution levels. Additionally, the data suggests that this block exhibits the lowest performance in terms of air quality compared to Block 1 and Block 2.

Overall, the graph in Fig. 4 presents that tree-based model performs well in predicting normal events. Nevertheless, in Table VII, when the data exceeds 150μg/$m^3$, none of the model achieve accurate predictions. This is due to the presence of extreme values, which pose challenges for the models in predicting these extreme events effectively. This analysis is further illustrated in Fig. 5, which shows the Block 3 of $PM_{10}$ for the next 24 hours prediction graphs for all tree-based models. The figure reveals a significant discrepancy between the actual and predicted $PM_{10,t+24h}$ values for all the models.

TABLE VII. Performance Results for Three Blocks Air Quality

| Model | Performance Indicator | XGBoost | GBR | DT |
|---|---|---|---|---|
| $PM_{10,t+24h}$ concentration $\leq$ 50 μg/$m^3$ (Block 1) | | | | |
| | RMSE | 14.1875 | 14.1673 | 15.2412 |
| | MAE | 10.6636 | 11.3487 | 11.2781 |
| | MAPE | 0.5400 | 0.7274 | 0.5751 |
| 50 μg/$m^3$ $<$ $PM_{10,t+24h}$ concentration $\leq$ 150 μg/$m^3$ (Block 2) | | | | |
| | RMSE | 27.2372 | 28.9145 | 30.6125 |
| | MAE | 22.0516 | 23.9327 | 24.5156 |
| | MAPE | 0.2997 | 0.3166 | 0.3309 |
| $PM_{10,t+24h}$ concentration $>$ 150 μg/$m^3$ (Block 3) | | | | |
| | RMSE | 106.0264 | 140.4993 | 131.3888 |
| | MAE | 83.4721 | 124.5067 | 108.5374 |
| | MAPE | 0.3668 | 0.5553 | 0.4747 |

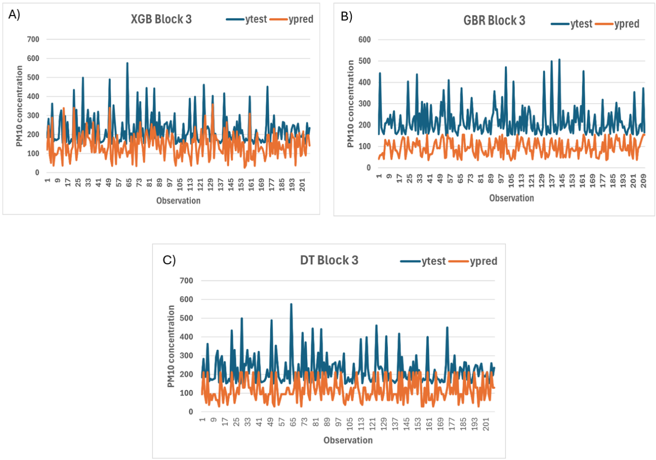

Fig. 5. Actual vs. Predicted for Block 3 of PM10 concentration for A) XGBoost B) GBR C) DT.

Table VIII shows the performance results of XGB, GBR and DT model for all air quality data. Based on the analysis of the overall data (without blocking) in Table VI, XGBoost outperforms the other two models, demonstrating greater efficiency in predicting $PM_{10,t+24h}$ concentrations (RMSE = 21.5921, MAE = 14.2396, MAPE = 0.4816). When comparing performance across blocks, XGBoost also surpasses DT and GBR. However, the performance gap is notably larger for the extreme air quality block compared to the good and moderate air quality blocks. This suggests that haze events have a significant impact on the models' accuracy. This disparity is due of the presence of extreme values in the extreme air quality block, leading to a highly skewed distribution and increased prediction error [51]. Previous studies have also highlighted this issue, where models effectively reduce overall error but struggle with accurately predicting extreme values [52] [53] [51] [54]. [55] further emphasized that another major challenge is the ability of standard learners to focus on the most important and extreme values. Therefore, neglecting these extreme values can result inaccurate model predictions.

## IV. Conclusion

The primary contribution of this study is to focus on extreme values using machine learning methods. Additionally, the evaluation of ML models is further explored through data blocking to assess whether the skewed data can be effectively modelled using the same ML approach. Based on

TABLE VIII. Performance Results of XGB, GBR and DT Model for All Air Quality Data

| Model | Performance Indicator | Overall | $PM_{10,24h}$ concentration $\leq$ $50\,\mu g/m^3$ (Block 1) | $50\,\mu g/m^3$ $<$ $PM_{10,24h}$ concentration $\leq$ $150\,\mu g/m^3$ (Block 2) | $PM_{10,24h}$ concentration $>$ $150\,\mu g/m^3$ (Block 3) |
|---|---|---|---|---|---|
| **Extreme Gradient Boosting (XGB)** | RMSE | 21.5921 | 14.1875 | 27.2372 | 106.0264 |
| | MAE | 14.2396 | 10.6636 | 22.0516 | 83.4721 |
| | MAPE | 0.4816 | 0.5400 | 0.2997 | 0.3668 |
| | N | 16488 | 12427 | 3855 | 208 |
| **Gradient Boosting Regression (GBR)** | RMSE | 24.5051 | 14.1673 | 28.9145 | 140.4993 |
| | MAE | 15.7770 | 11.3487 | 23.9327 | 124.5067 |
| | MAPE | 0.5528 | 0.7274 | 0.3166 | 0.5553 |
| | N | 16488 | 12367 | 3912 | 211 |
| **Decision Tree (DT)** | RMSE | 24.7156 | 15.2412 | 30.6125 | 131.38885 |
| | MAE | 15.5915 | 11.2781 | 24.5156 | 108.5374 |
| | MAPE | 0.5164 | 0.5751 | 0.3309 | 0.4747 |
| | N | 16488 | 12427 | 3855 | 208 |

the discussion of all comparisons between the overall data and the blocks (as discussed above), XGBoost outperforms the other two models with RMSE(21.5921), MAE(14.2396) and MAPE(0.4816), indicating XGB model is more efficient for predicting $PM_{10}$ concentration. The comparison Block 1, Block 2, and Block 3 air quality data blocks show a decline in performance, as indicated by RMSE, MAE, and MAPE. The performance gap is significantly larger for the Block 3 air quality block compared to the overall, Block 1, Block 2 air quality blocks. This disparity arises from the presence of extreme values in the Block 3 air quality block, making it as challenging for the model to generate accurate predictions. As a result, the error indicators become significantly high, leading to a high discrepancy between actual and predicted $PM_{10}$ concentrations. For further analysis, since XGBoost outperforms the other models, it will be further utilized and enhanced to better handle extreme data, as this is essential for improving $PM_{10}$ concentration predictions, particularly during haze events with elevated $PM_{10}$ levels.

## Acknowledgment

We would like to express our sincere gratitude to Faculty of Computer Sciences and Mathematics and Universiti Teknologi Mara (UiTM) for their continuous support throughout this study. Additionally, we extend our deepest appreciation to the DOE Malaysia for providing air quality data used in this study.

## References

[1] Dondi, A., Carbone, C., Manieri, E., Zama, D., Bono, C., Betti, L., Biagi, C. & Lanari, M. Outdoor Air Pollution and Childhood Respiratory Disease: The Role of Oxidative Stress. *International Journal Of Molecular Sciences*. **24** (2023,3)

[2] Gul, H. & Das, B. The Impacts of Air Pollution on Human Health and Well-Being: A Comprehensive Review. *Journal Of Environmental Impact And Management Policy*., 1-11 (2023,10)

[3] Gulati, S., Bansal, A., Pal, A., Mittal, N., Sharma, A. & Gared, F. Estimating PM2.5 utilizing multiple linear regression and ANN techniques. *Scientific Reports*. **13** (2023,12)

[4] Abdullah, S., Ismail, M. & Fong, S. MULTIPLE LINEAR REGRESSION (MLR) MODELS FOR LONG TERM PM 10 CONCENTRATION FORECASTING DURING DIFFERENT MONSOON SEASONS. *Article In Journal Of Sustainability Science And Management*. **12** pp. 60-69 (2017), https://www.researchgate.net/publication/318777794

[5] Rani, N., Azid, A., Khalit, S., Juahir, H. & Samsudin, M. Air pollution index trend analysis in Malaysia, 2010-15. *Polish Journal Of Environmental Studies*. **27**, 801-808 (2018)

[6] Jafarigol, E. & Trafalis, T. A Review of Machine Learning Techniques in Imbalanced Data and Future Trends. (2023)

[7] Birim, N., Turhan, C., Atalay, A. & Akkurt, G. The Influence of Meteorological Parameters on PM10: A Statistical Analysis of an Urban and Rural Environment in Izmir/Türkiye. *Atmosphere*. **14** (2023,3)

[8] Rahim, N., Noor, N., Jafri, I., Ul-Saufie, A., Ramli, N., Seman, N., Kamarudzaman, A., Zainol, M., Victor, S. & Deak, G. Variability of PM10 level with gaseous pollutants and meteorological parameters during episodic haze event in Malaysia: Domestic or solely transboundary factor?. *Heliyon*. **9** (2023,6)

[9] Lešnik, U., Mongus, D. & Jesenko, D. Predictive analytics of PM10 concentration levels using detailed traffic data. *Transportation Research Part D: Transport And Environment*. **67** pp. 131-141 (2019,2)

[10] Kumar, K. & Pande, B. Air pollution prediction with machine learning: a case study of Indian cities. *International Journal Of Environmental Science And Technology*. **20**, 5333-5348 (2023,5)

[11] Ali, Z., Abduljabbar, Z., Tahir, H., Sallow, A. & Almufti, S. eXtreme Gradient Boosting Algorithm with Machine Learning: a Review. *Academic Journal Of Nawroz University*. **12**, 320-334 (2023,5)

[12] Bakar, M., Ariff, N., Nadzir, M., Wen, O. & Suris, F. Prediction of Multivariate Air Quality Time Series Data using Long Short-Term Memory Network. — *Malaysian Journal Of Fundamental And Applied Sciences*. **18** pp. 52-59 (2022)

[13] Noor, N., Deak, G., Ul-Saufie, A., Mohd, Z. & Rozainy, R. MODELING OF PARTICULATE MATTER (PM10) DURING HIGH PARTICULATE EVENT (HPE) IN KLANG VALLEY, MALAYSIA. (2022), www.ijcs.ro

[14] Hong, W., Koh, D. & Yu, L. Development and Evaluation of Statistical Models Based on Machine Learning Techniques for Estimating Particulate Matter (PM2.5 and PM10) Concentrations. *International Journal Of Environmental Research And Public Health*. **19** (2022,7)

[15] Rahim, N., Noor, N., Jafri, I., Ramli, N., Kamaruddin, M. & Deák, G. Predicting Particulate Matter (PM10) during High Particulate Event (HPE) using Quantile Regression in Klang Valley, Malaysia. *IOP Conference Series: Earth And Environmental Science*. **1216** (2023)

[16] Kim, B., Lim, Y. & Cha, J. Short-term prediction of particulate matter (PM10 and PM2.5) in Seoul, South Korea using tree-based machine learning algorithms. *Atmospheric Pollution Research*. **13** (2022,10)

[17] Qadeer, K. & Jeon, M. Prediction of PM10 Concentration in South Korea Using Gradient Tree Boosting Models. *ACM International Conference Proceeding Series*. (2019,8)

[18] Sayegh, A., Munir, S. & Habeebullah, T. Comparing the performance of statistical models for predicting PM10 concentrations. *Aerosol And Air Quality Research*. **14**, 653-665 (2014)

[19] Barthwal, A., Acharya, D. & Lohani, D. Prediction and analysis of particulate matter (PM2.5 and PM10) concentrations using machine learning techniques. *Journal Of Ambient Intelligence And Humanized Computing*. **14**, 1323-1338 (2023,3)

[20] Suleiman, A., Tight, M. & Quinn, A. Applying machine learning methods in managing urban concentrations of traffic-related particulate matter (PM10 and PM2.5). *Atmospheric Pollution Research*. **10**, 134-144 (2019,1)

[21] Watpade, A., Thakor, S., Jain, P., Mohapatra, P., Vaja, C., Joshi, A., Shah, D. & Islam, M. Comparative analysis of machine learning models for predicting dielectric properties in MoS2 nanofiller-reinforced epoxy composites. *Ain Shams Engineering Journal*. **15** (2024,6)

[22] Mandvi, Patel, P. & Singh, H. Performance analysis of machine learning models for AQI prediction in Gorakhpur City: a critical study. *Environmental Monitoring And Assessment*. **196** (2024,10)

[23] 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 Journal Of Atmospheric Environment*. **17** (2023,12)

[24] Dao, T., Nhat, H., Trung, H., Dieu, V., Thu, N., Tran, D. & Tran, D. 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*. **33** pp. 89-94 (2023,3)

[25] Naveen, S., Upamanyu, M., Chakki, K., Chandan, M. & Hariprasad, P. Air Quality Prediction Based on Decision Tree Using Machine Learning. *International Conference On Smart Systems For Applications In Electrical Sciences, ICSSES 2023*. (2023)

[26] Arafin, S., Ul-Saufie, A., Ghani, N., Ibrahim, N. & Alam, S. Feature Selection Methods Using RBFNN Based on Enhance Air Quality Prediction: Insights from Shah Alam. *IJACSA) International Journal Of Advanced Computer Science And Applications*. **15** (2024), www.ijacsa.thesai.org

[27] Libasin, Z., Ul-Saufie, A. & Hasfazilah, A. identifying missing data mechanisms among incomplete air pollution datasets in Malaysia. (2024)

[28] Srivastava, C., Singh, S. & Singh, A. Estimation of air pollution in Delhi using machine learning techniques. *2018 International Conference On Computing, Power And Communication Technologies, GUCON 2018*. pp. 304-309 (2019,3)

[29] Gupta, N., Mohta, Y., Heda, K., Armaan, R., Valarmathi, B. & Arulkumaran, G. Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis. *Journal Of Environmental And Public Health*. **2023** pp. 1-26 (2023,1)

[30] Ditrich, J. PŮVODNÍ STATĚ DATA REPRESENTATIVENESS PROBLEM IN CREDIT SCORING. *ACTA OECONOMICA PRAGENSIA*. **23** (2015)

[31] Geurts, P., Irrthum, A. & Wehenkel, L. Supervised learning with decision tree-based methods in computational and systems biology. *Molecular BioSystems*. **5** pp. 1593-1605 (2009)

[32] Doreswamy, Harishkumar, K., Km, Y. & Gad, I. Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models. *Procedia Computer Science*. **171** pp. 2057-2066 (2020)

[33] Rokach, L. & Maimon, O. Decision Trees. *Data Mining And Knowledge Discovery Handbook*. pp. 165-192 (2006,5)

[34] Srihith, I., Thippna, G. & Srinivas, T. A Forest of Possibilities Decision Trees and Beyond. *Journal Of Advancement In Parallel Computing*. (2023)

[35] Hoarau, A., Martin, A., Dubois, J. & Gall, Y. Evidential Random Forests. *Expert Systems With Applications*. **230** (2023,11)

[36] Adamu, H., Muhammad, M. & Mohammed Application of Gradient Boosting Algorithm In Statistical Modelling. (Journal of Statistics,2019)

[37] Otchere, D., Ganat, T., Ojero, J., Tackie-Otoo, B. & Taki, M. Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. *Journal Of Petroleum Science And Engineering*. **208** (2022,1)

[38] Rao, H., Shi, X., Rodrigue, A., Feng, J., Xia, Y., Elhoseny, M., Yuan, X. & Gu, L. Feature selection based on artificial bee colony and gradient boosting decision tree. *Applied Soft Computing Journal*. **74** pp. 634-642 (2019,1)

[39] Martínez-Muñoz, G., Bentéjac, C. & Martínez-Muñoz, A. A Comparative Analysis of XGBoost. (2019), https://www.researchgate.net/publication/337048557

[40] Zemel, R. & Pitassi, T. A Gradient-Based Boosting Algorithm for Regression Problems. (2001)

[41] Shahani, N., Zheng, X., Liu, C., Hassan, F. & Li, P. 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*. **9** (2021,10)

[42] Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. (2016,3), http://arxiv.org/abs/1603.02754

[43] Mienye, I. & Sun, Y. A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. *IEEE Access*. **10** pp. 99129-99149 (2022)

[44] Pan, B. Application of XGBoost algorithm in hourly PM2.5 concentration prediction. *IOP Conference Series: Earth And Environmental Science*. **113** (2018,2)

[45] Weerts, H., Mueller, A. & Vanschoren, J. Importance of Tuning Hyperparameters of Machine Learning Algorithms. (2020,7), http://arxiv.org/abs/2007.07588

[46] Abdullah, S., Ismail, M., Ahmed, A. & Abdullah, A. Forecasting particulate matter concentration using linear and non-linear approaches for air quality decision support. *Atmosphere*. **10** (2019,11)

[47] DOE Department of Environment:Malaysia Environmental Quality Report 2015. Kuala Lumpur: Ministry of Energy, Science, Technology, Environment and Climate Change, Malaysia. (2015)

[48] DOE Department of Environment:Malaysia Environmental Quality Report 2016. Kuala Lumpur: Ministry of Energy, Science, Technology, Environment and Climate Change, Malaysia. (2016)

[49] DOE Department of Environment:Malaysia Environmental Quality Report 2017. Kuala Lumpur: Ministry of Energy, Science, Technology, Environment and Climate Change, Malaysia. (2017)

[50] DOE Department of Environment:Malaysia Environmental Quality Report 2019. Kuala Lumpur: Ministry of Energy, Science, Technology, Environment and Climate Change, Malaysia. (2019)

[51] Ribeiro, R. & Moniz, N. Imbalanced regression and extreme value prediction. *Machine Learning*. **109**, 1803-1835 (2020,9)

[52] Branco, P., Torgo, L. & Ribeiro, R. Pre-processing approaches for imbalanced distributions in regression. *Neurocomputing*. **343** pp. 76-99 (2019,5)

[53] Ren, J., Zhang, M., Yu, C. & Liu, Z. Balanced MSE for Imbalanced Visual Regression. *Proceedings Of The IEEE Computer Society Conference On Computer Vision And Pattern Recognition*. **2022-June** pp. 7916-7925 (2022)

[54] Sadouk, L., Gadi, T. & Essoufi, E. A novel cost-sensitive algorithm and new evaluation strategies for regression in imbalanced domains. *Expert Systems*. **38** (2021,6)

[55] Branco, P., Ribeiro, R., Torgo, L., Matwin, S., Japkowicz, N., Krawczyk, B. & Moniz, N. REBAGG: REsampled BAGGing for Imbalanced Regression. *Proceedings Of Machine Learning Research*. **94** pp. 67-81 (2018)