CHAPTER 4

AIR POLLUTION FORECASTING BASED ON ENSEMBLE CLASSIFICATION MODEL

4.1 Introduction

The prediction air quality data accurately identifies the dynamic nature, instability and high inconsistency of pollutants from air pollution. Air pollution has recently been considered the most critical issue in developing countries due to hasty mechanization. Predicting air pollution reduces the adverse effect on the atmosphere and people's well-being. For predicting air pollution from the dataset, machine learning techniques are designed. The developed learning process forecast accurate pollutant result for both individual and society. The application of a feature selection procedure improves data forecasting performance. Relevant features that lower the risk encountered during the AQI prediction are identified throughout the feature selection procedure.

The presence of unrelated features causes more severe problems and lacks accuracy in air pollution forecasting. For efficient air pollution prediction, LR-MSV model is described in the previous section. At first, several air quality data with different features are considered from the dataset. For each input air data, pre-processing is performed to remove noise data for attaining efficient air forecasting. After that, the feature selection process is performed by using LRC models. According to the estimated correlation coefficient value, significant optimum relevant features are determined. Lastly, accurate air quality data classification is presented through multiclass support vectors to improve the performance of prediction with minimum time and error rate. Here, the entire air quality data failed to consider at a particular time to select significant relevant features. Thus, it consumes more space for storing air data for pollution forecasting.

Recently, a few studies have been created employing data mining and ML techniques to forecast and predict different air quality pollutants. However, accurate pollutant and particulate air quality levels failed to present at the required level. Air quality contamination makes it difficult to estimate pollution levels in the early stages. The ensemble classification technique was presented to classify air data from the dataset for forecasting. It effectively classifies data with increased accuracy but needs to execute the forecasting modeling effectively. The air quality dataset presents a vast amount of data on air quality, making it challenging to identify the information that is crucial for forecasting air pollution. For efficient forecasting, pre-processing and feature selection are performed to analyze the risk factors. However, the time complexity involved during data classification is more due to irrelevant features of data. The existing classification techniques could not handle the early-stage accurate air pollution forecasting performance with minimum time.

In order to predict air pollution, the BTBSR-QWEBC model is developed. The primary goal of the suggested method is to choose features and classify data in order to predict air pollution more accurately. The feature section, classification procedure, and pre-processing are the foundations upon which the suggested technique is built. Using the BDZWT technique, pre-processing is done for each input data set to eliminate noise data from the dataset. With obtained pre-processed data, feature selection is presented to select significant features. Here, relevant significant features are selected by applying Otsuka indexive Broken-stick regression process. The Otsuka similarity coefficient value is measured between features to identify relevant and irrelevant features. Eliminating irrelevant features minimizes time to forecast air data and helps attain accurate data classification. Ultimately, data is categorized into distinct classes using the weighted emphasis increase technique. By building several weak learners, the considered ensemble classification technique produces a strong learner by combining the results. Next, in

order to estimate quadratic error, a weight is assigned to each weak learner. Strong classification outcomes with reduced error and precise air quality forecasts are thus attained. As a result, in the shortest amount of time, the BTBSR-QWEBC model generated air pollution forecasts with higher accuracy. This experiment tests a number of variables related to air pollution forecasting, including accuracy, time, mistake rate, and spatial complexity. The experiment's results show that the recommended approach beats cutting-edge methods for predicting air pollution.

4.2 Architecture of Proposed Ensemble Model

The accurate forecast of air pollution in the environment is achieved by proposing a BTBSR-QWEBC technique. The developed technique predicted accurate air pollution with enhanced accuracy and minimized time complexity. From the considered air quality dataset, a subset of relevant features is selected to reduce space complexity that occurred during the forecasting process. The presence of irrelevant features may cause a risk during air pollution prediction. The feature selection process is the method used to identify important features from the dataset and also it helps to minimize the dimensionality of the dataset. Several methods are developed to predict air pollution early by removing noisy air pollutant data. Accurate prediction is still a risk due to improving data. It failed to select significant features with minimum time, and the forecasting performance result needed to be improved.

As a result, the suggested BTBSR-QWEBC model is created to improve air pollution forecasting accuracy while requiring the least amount of memory and time. Here, noisy air data is removed for a better classification process. While selecting the essential features and eliminating irrelevant features, dimensionality reduction is achieved. Data in the suggested model Preprocessing is done to get pre-processed data and to eliminate duplicate data. Selecting the most relevant features from the dataset is the next step in the process, known as feature selection. The complexity of pollution forecasting is reduced using the acquired features.

By calculating the Otsuka similarity index coefficient between the features, this objective is accomplished. More significant features are obtained for air pollution prediction based on the similarity index. Following this, ensemble boost classification technique is performed using selected relevant features. Based on the amount of data with relevant features, a weak classifier is constructed, and results are combined to form strong classification results. To increase accuracy, the quadratic error is estimated for every weak learner. For the purpose of efficiently classifying data into distinct classes, a weak learner with the least amount of error is chosen. The categorized data contributes to faster and more accurate air pollution forecasting.

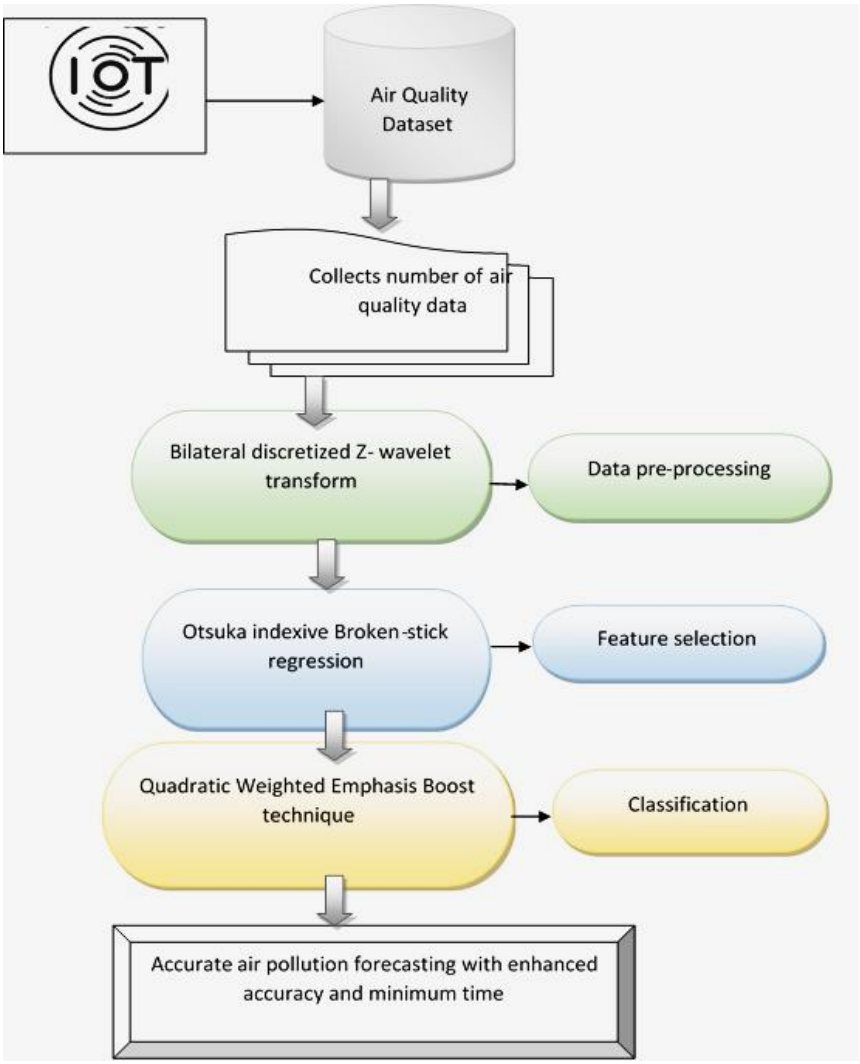

Figure 4.1: Architecture of proposed BTBSR-QWEBC model

The process of the suggested BTBSR-QWEBC model for precise air pollution prediction is shown in above Figure 4.1. The air quality dataset is considered, and IoT devices are used to gather the air quality data. At first, data pre-processing is performed to remove noisy data from dataset by applying bilateral discretized Z- transform. With obtained pre-processed data, more relevant features are selected through Otsuka Indexive Broken-stick Regression procession. The selected significant features are utilized to forecast air pollution with minimized complexity and error rate.

At last, Quadratic Weighted Emphasis Boost technique is implemented to classify data with the support of selected relevant features. Thus, strong classification result with minimum error is attained for accurate air pollution forecasting.

4.2.1 Data pre-processing using Wavelet Transform

The proposed model initially performs data pre-processing that removes noisy data collected from dataset. Here, BDZWT technique is used to perform pre-processing. The process of converting the raw input air quality data into a readable format or structure is known as transformation. The modified data is fed into the subsequent training model to improve forecasting performance. The considered wavelet transformation converts data format, structure or values into clean and usable data. The obtained pre-processed data helps to forecast air pollution.

Initially, air quality dataset 'DS' is considered and it includes several different numbers of features '$\beta_{1,}\beta_{2,}\beta_{3,}\ldots\beta_{n,}$' that are arranged into the row and columns. The arrangement of features is named as feature matrix and given,

$$F_a = \begin{matrix} - \\ R_1 \\ R_2 \\ R_3 \\ \vdots \\ R_n \end{matrix} \begin{bmatrix} C_1 & C_2 & C_3 & \ldots & C_n \\ \beta_{11} & \beta_{12} & \beta_{13} & \ldots & \beta_{1n} \\ \beta_{21} & \beta_{22} & \beta_{23} & \ldots & \beta_{2n} \\ \beta_{31} & \beta_{32} & \beta_{33} & \ldots & \beta_{3n} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ \beta_{n1} & \beta_{n2} & \beta_{n3} & \ldots & \beta_{nn} \end{bmatrix} \quad \ldots\ldots \text{Eqn (4.1)}$$

The feature matrix 'F_a' is demonstrated in the Equation (4.1) with the combination of rows and columns. Here, number of rows '$R_1, R_2, R_3, .. R_n$' denotes the features and number of columns '$C_1, C_2, C_3, .. C_n$' specifies the instances. After considering input features, the BDZWT is applied to transfer input data and mathematical expressed as given,

$$Y(z) = \sum_{n=-\infty}^{\infty} y\,[n]\, z^{-n} \qquad \ldots\ldots \text{Eqn (4.2)}$$

In Equation (4.1), the BDZWT is expressed and indicated as '$Y(z)$'. Here, 'n' indicates an integer and 'z' denotes a complex variable '$z = re^{j\omega}$'. By replacing complex variable, the equation is re-phrased as

$$Y(z) = \sum_{n=-\infty}^{\infty} y\,[n]\, re^{-j\omega n} \qquad \ldots\ldots \text{Eqn (4.3)}$$

Equation (4.3), which divides the input data into low-frequency and high-frequency feature components, is used to determine the BDZWT. In the given Equation, 'ω' denotes an angular frequency '$\omega = 2\pi f$' where 'f' is temporal frequency.

$$Y(z) = \sum_{n=-\infty}^{\infty} y[n]\ re^{-2j\pi fn} \qquad \ldots\ldots \text{Eqn (4.4)}$$

The result of wavelet Z- transform is achieved in Equation (4.4). Based on obtained result, air quality data is decomposed into low and high frequency feature components.

$$\varphi_l(n) = \frac{1}{2}\sum_{n=-\infty}^{\infty} y\,[n]\, re^{-2j\pi fn} \qquad \ldots\ldots \text{Eqn (4.5)}$$

A low frequency feature component '$\varphi_l(n)$' is described in Equation (4.5) and high frequency feature component '$\varphi_h(n)$' is illustrated in Equation (4.6).

$$\varphi_h(n) = \frac{1}{2}\sum_{n=-\infty}^{\infty} y\,[n]\, re^{-2j\pi fn} \qquad \ldots\ldots \text{Eqn (4.6)}$$

Figure 4.2 shows the intended BDZWT flow. The high-frequency and low-frequency feature components are identified using the wavelet decomposition. Low-frequency feature components are identified as noise-air quality data based on the decomposed result. The dataset is cleansed of the identified noise data in order to improve air pollution forecasts.

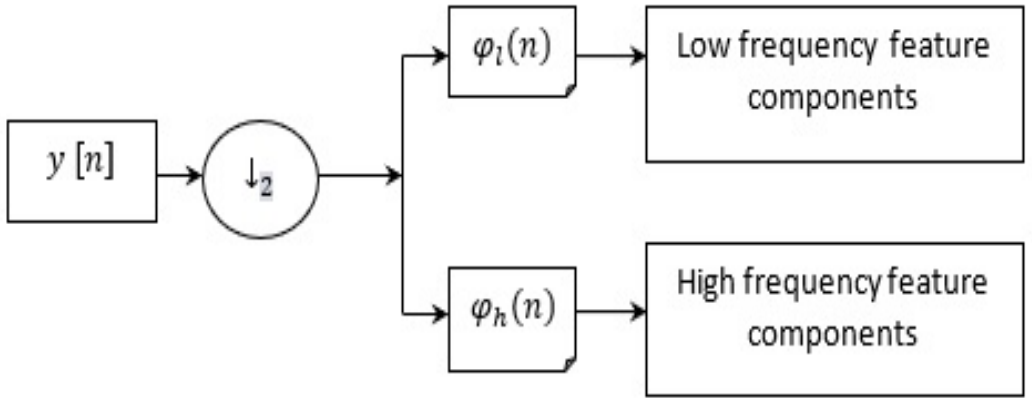

Figure 4.2: Flow of BDZWT technique

After that, pre-processed data related to frequency feature components are chosen in order to carry out air pollution forecasting. Algorithm 4.1 outlines the algorithmic procedure for BDZWT-based pre-processing.

Initially, number of air quality data considered from dataset with various features. For each input air data, BDZWT is applied to remove noise data. The considered features and instances are arranged in form of row and column for construction of feature matrix. After that, wavelet decomposition is performed on each feature to convert data into low and high-frequency components. From that, low frequency component is denoted as noisy data and it is removed to attain efficient pre-processed data.

4.2.2 Feature Selection based on Regression Model

Next, the feature selection procedure is completed using the Otsuka Indexive Regression using broken sticks and pre-processed data collected. A machine learning model enables the intended regression method. In the least amount of time, forecasting accuracy can be improved by carefully choosing pertinent data aspects. Here, the similarity coefficient value is used to distinguish between relevant and irrelevant features. The dimensionality of the air quality data is reduced through feature selection from the dataset. In order to identify pertinent information that shortens the time needed to anticipate air pollution, the suggested technique evaluates the Otsuka similarity index. Figure 4.3 illustrates the relevant feature selection process flow.

The chosen features contribute to reducing the temporal complexity involved in air pollution forecasting. When we look at the quantity of features from the dataset as '$\beta_{1,}\beta_{2,}\beta_{3,}$ ….$\beta_{n,}$' of air quality data. Then, the relationships between a dependent variable (features) are determined by using Broken-stick regression analysis. Based on breakpoint, the input characteristics are divided into two segments using the suggested Broken-stick regression. In this case, the breakpoint designates the threshold value above or below which the intended effects materialize.

When making decisions based on the Otsuka similarity index, breakpoint is employed. By applying Otsuka similarity index, the similarity between the features is measured and mathematically formulated as given:

$$\rho = \frac{\sum \beta_{i,}\beta_{j}}{\sqrt{\sum B_i{}^2 \sum \beta_j{}^2}} \qquad \text{……. Eqn (4.7)}$$

Equation (4.7) is used to calculate the similarity coefficient between two characteristics '$\beta_{i,}$' and '$\beta_{i,}$'. The similarity coefficient has a range of values from -1 to +1 and is represented by the symbol 'ρ'.

$$\rho = \begin{cases} +1 \ ; relevant features \\ -1 \ ; irrelevant fatures \end{cases} \quad \text{…… Eqn (4.8)}$$

By using Equation (4.8), relevant and irrelevant features are represented based on estimated similarity coefficient value. If the similarity coefficient value is '-1', then it indicated negative similarity.

Input: Dataset 'DS', feature '$F = \beta_1, \beta_2, \beta_3, \ldots . \beta_n,$', Air Quality data '$D = D_1, D_2, \ldots, D_n$'

Output: Noise reduced pre-processed air quality data 'PD'

Begin

Step 1: For each Dataset 'DS' (Air Quality data '$D = D_1, D_2, \ldots, D_n$')

Step 2: Apply bilateral discretized wavelet Z- transform

Step 3: Create feature matrix 'F_a'

Step 4: **For** each Features 'F'

Step 5: Perform wavelet decomposition using (4.1)

Step 6: Obtain Low and high-frequency components $\varphi_l(n), \varphi_h(n)$

Step 7: Remove Low-frequency components '$\varphi_l(n)$'

Step 8: **Return** pre-processed air quality data

Step 9: **End for**

Step 10: End for

End

Algorithm 4.1: Process of Bilateral discretized Z- wavelet transform

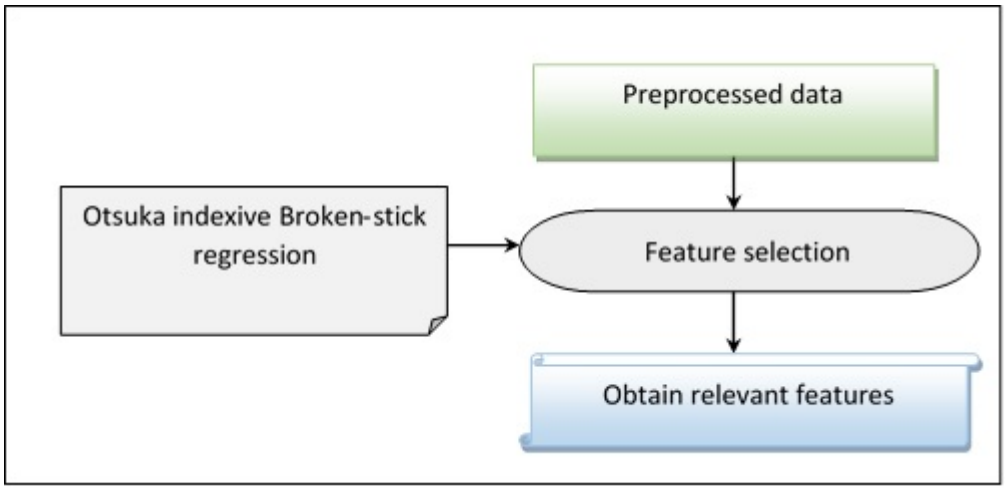

Figure 4.3: Flow of Otsuka Indexive Broken-Stick Regression

Similarly, coefficient value provided with '+1' specifies the positive similarity. Here, positive similarity is considered as relevant features and negative similarity is indicated as irrelevant features. Based on measured similarity coefficient value, the relevant features are selected, and irrelevant features are removed. The selected relevant features help to attain accurate air data classification with minimized time consumption. Thus, the algorithmic process of features selection based on regression process is defined in Algorithm 4.2.

Finding the pertinent data features is the primary goal of OIBSR Technique. Initially, the dataset's number of features is considered. Subsequently, BSR is used to distinguish between important and irrelevant aspects in the input feature.

Between characteristics, the Otsuka similarity index coefficient value is calculated. Using the coefficient value as a guide, relevant features are extracted from the dataset and superfluous features are removed. To improve air pollution forecasting, data is classified using the appropriate features that have been chosen.

4.2.3 Ensemble Boost Classification Technique

The proposed BTBSR-QWEBC model finally performs Quadratic weighted emphasis boost classification to classify data into different classes. The developed

weighted emphasis boost classification is a machine learning ensemble algorithm. The ensemble classification technique enhances the air pollution monitoring and controlling. Here, number of weak learners is constructed, and results are combined to attain strong classification result.

Thus, it achieves strong classification result with minimum error. Figure 4.4 illustrates the flow of designed emphasis boost classification technique to obtain an accurate air pollution prediction with minimum time consumption. Initially, training data with extracted relevant features are considered as input for classification process. Based on considered training data, number of weak classifiers is constructed.

Input: Dataset 'DS', feature '$F = \beta_{1,}\beta_{2,}\beta_{3,} \ldots . \beta_{n,}$',

Output: Selected relevant features

Begin

Step 1: **For** each feature in 'DS'

Step 2: Apply Broken-stick regression

Step 3: Measure the Otsuka similarity index 'ρ'

Step 4: If $(\rho = +1)$ **then**

Step 5: The feature is said to be relevant

Step 6: **Else**

Step 7: The feature is said to be irrelevant

Step 8: **End if**

Step 9: Return relevant feature

Step 10: End for

End

Algorithm 4.2: Procedure of Otsuka Indexive Broken-Stick Regression

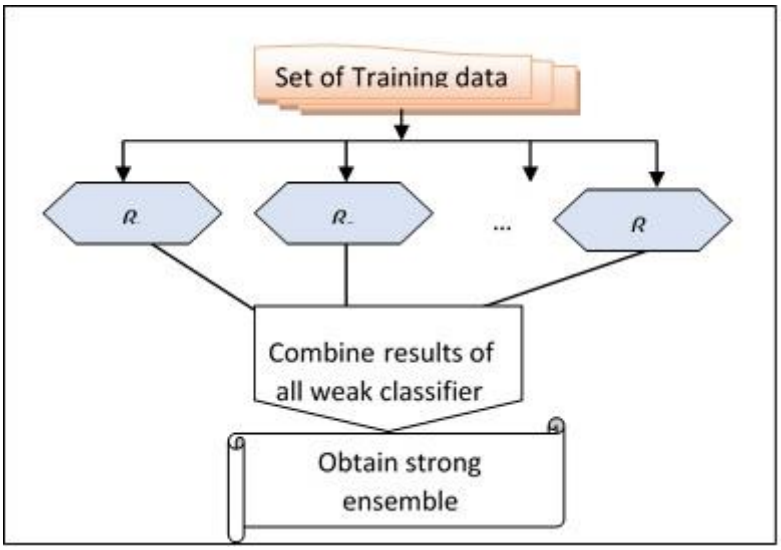

Figure 4.4: Architecture of QWEBC technique

Then, the result of weak classifier is combined to form strong classification result. Based on obtained result, accurate prediction of air pollution is attained with minimum time. Let us consider input training sample set specified as '$\{D_i, Z\}$'. Here, number of sample air data from dataset specified as '$D_i = D_1, D_2, \ldots, D_m$' and ensemble classification result is denoted as 'Z'. With considered input training data, weak learners 'β' is constructed and represented as '$C_1, C_{2,} C_{3,} \ldots . C_B$'. As a result, ensemble classification attains strong classification result for forecasting the air pollution with higher accuracy.

For efficient data classification, Kernelized support vector classifier is utilized in developed classification technique. It constructs the best line or decision boundary represented in n-dimensional space. It separates data into different classes, and the decision boundary is represented as hyperplane. Here, hyperplane is denoted based on the quantity of output classes. The data point closest to hyperplane position is referred to as support vector. With the set of input data, hyperplane is attained by using the following expression.

$$H \rightarrow \alpha . D_i + d = 0 \qquad \ldots\ldots \text{ Eqn (4.9)}$$

By using Equation (4.9), hyperplane or decision boundary 'H' is demonstrated. The hyperplane is attained based on a training sample 'D_i' (i.e., input air pollutant data), a bias indicated as 'd' and a normal weight vector to hyperplane denoted as 'α'. The input data is separated into different classes when the training samples are linearly separable. In hyperplane, data is presented in the above and below side of the decision boundary which is expressed as below.

$$K_1 \rightarrow \alpha . D_i + d > 0 \qquad \ldots\ldots \text{Eqn (4.10)}$$

$$K_2 \rightarrow \alpha . D_i + d < 0 \qquad \ldots\ldots \text{Eqn (4.11)}$$

From above Equations (4.10) and (4.11), support vector that positioned above and below the boundary is represented and symbolized as 'K_1' and 'K_2'. After separating data into different classes, Kernel function is applied to predict output of support vector classifier. The predicted output is given below.

$$Z = \sum \alpha y_i \vartheta \left(AQI_t, AQI_r\right) \qquad \ldots\ldots \text{Eqn (4.12)}$$

The predicted classification results are illustrated in Equation (4.12) and denoted as 'Z'. Here, '$\vartheta \left(AQI_t, AQI_r\right)$' specifies a Kernel function relationship that measures relationship between Testing AQI value (i.e.,AQI_t) and Training AQI (i.e.,AQI_r), the weight of training sample is denoted as 'α'.

For input sample air data, the AQI value is estimated according to the selected features during feature selection process. It is calculated with average of air pollutant concentration from selected features '$PM2.5, PM10, SO2, NOx, NO2$', highest value of CO and O3 respectively.

The formulation of AQI is given below.

$$AQI = Avg(PM2.5, PM10, SO2, NOx, NO2) + Max(CO, O3) \qquad \ldots\ldots \text{Eqn (4.13)}$$

The estimation of Air Quality Index 'AQI' is measured using Equation (4.13). Based on estimated result, Laplace RBF Kernel function is applied next for determining relationship between training and testing sample air data. The formulation of Laplace RBF kernel is expressed as given below.

$$\boldsymbol{\vartheta}\,(AQI_t, AQI_r) = exp\left(-\frac{\|AQI_t - AQI_r\|^2}{v^2}\right) \qquad \ldots \text{ Eqn (4.14)}$$

The relationship between training and testing sample air data '$\boldsymbol{\vartheta}\,(AQI_t, AQI_r)$' is measured in Equation (4.14) using Laplace RBF Kernel. Here, 'v' specifies a deviation. The data is categorized into a certain class when the AQI of the training and testing sets are equal. Similarly, a specific class is assigned to the training AQI value that is closest to the testing AQI value. Based on estimation result, air data is classified into six diverse classes for better air pollution prediction.

The classified classes are namely good, satisfactory, moderate, poor, very poor, and severe. Furthermore, the Laplace RBF Kernel function yields similarity values between 0 and 1. Accurate classification of data is achieved based on similarity value. A higher similarity value, or 1, indicates that the conclusions of the data classification are accurate. The classification of data into different classes using kernel function is demonstrated in Figure 4.5.

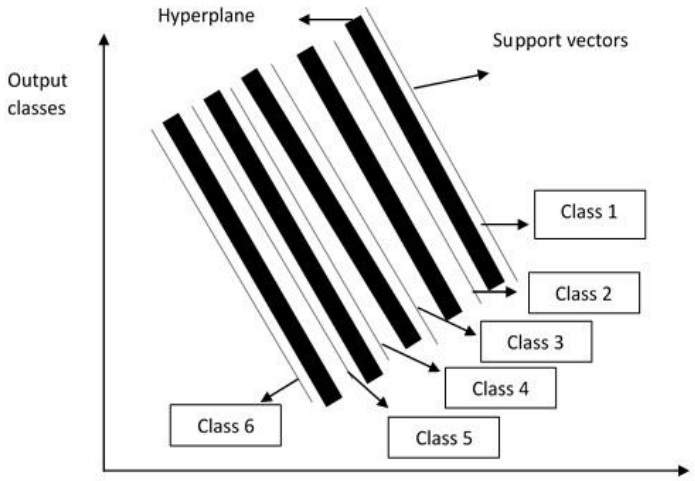

Figure 4.5: Construction of a Kernelized support vector classifier

Here, hyperplane (H) is used as a decision boundary to classify the input data into six diverse classes through the kernel function. Thus, set of training data is used to construct number of weak learners. The obtained weak learner results include some training error while classifying data. For better strong classification result, the results of all weak learners are combined and expressed as below.

$$Z = \sum_{i=1}^{B} C_i \quad \text{..... Eqn (4.15)}$$

The ensemble classification output 'Z' is determined using Equation (4.15) based on the output of weak learner's indicated as 'C_i'. Following this, weight is assigned for each weak learner result to achieve strong classification results with minimum error. The classification output with weight is given below.

$$Z = \sum_{i=1}^{B} C_i * \gamma_i \quad \text{..... Eqn (4.16)}$$

The findings of the weak learners are marked with a weight of 'γ_i' and it is provided in a random integer number. To estimate error occurred during classification results, the weighted emphasis function is utilized in proposed ensemble technique. The quadratic error of weak classification result is expressed below.

$$E_q = \exp\left[p\left((\sum_{i=1}^{B} C_i\,\gamma_i - Z)^2 - (1-p)\left(\sum_{i=1}^{b} C_i\right)^2\right)\right] \quad \text{..... Eqn (4.17)}$$

The quadratic error is estimated by the weighted emphasis function 'E_q' by using Equation (4.17). Here, 'p' denotes a weighting parameter, 'Z' denotes actual ensemble classification results, '$\sum_{i=1}^{B} C_i\gamma_i$' indicates a predicted classification result of weak learner with the weight 'γ_i' and the without weight '$\sum_{i=1}^{B} C_i$'. The weighting parameter value is set to 1 and final quadratic error is estimated as follows.

$$E_q = \exp[(\sum_{i=1}^{B} C_i\gamma_i - Z)^2] \quad \text{..... Eqn (4.18)}$$

The final quadratic error 'E_q' is calculated using Equation (4.18). When there is an inaccuracy in the classification result, the weak learner's weight is modified. The weight is reduced if the weak learner is appropriately identified. If not, the starting weight value is raised. As a result, strong classification result with minimum error is selected to forecast air pollution with higher accuracy. The process of developed Quadratic weighted emphasis boost classification is illustrated in Algorithm 4.3 to obtain enhanced accuracy on air pollution forecasting with minimum error. Initially, number of training sample data 'D' is considered from data. Based on input sample data, number of weak learners is constructed using Kernelized support vector classifier. For each weak leaner, class is initialized to estimate air quality index. Then, the Laplace RBF kernel function is applied to measure relationship between training and testing air quality data. Based on estimated kernel function value, data is classified into a particular class. After that, weight for each weak learner is assigned and the quadratic error of each weak learner result is measured using emphasis function. Then, the results of all weak learners are combined to form strong classification result with minimum error. Hence, proposed BTBSR-QWEBC model achieves accurate air quality forecasting with minimum time and memory consumption. According to dataset attributes, the air quality indices (AQI) and bucket values are tabulated as given in Table 4.1.

Table 4.1: AQI Classes.

S. No.	AQI	AQI classes
1	0 and 50	Good
2	51 and 100	Satisfactory
3	101 and 200	Moderate
4	201 and 300	Poor
5	301 and 400	Very poor
6	Greater than 401	Severe

Input: Selected features, data $D_1, D_2, D_3, \ldots D_n$

Output: Improve air pollution forecasting accuracy

Begin

Step 1: For each data 'D_i'

Step 2: Construct 'B' number of weak learners

Step 3: Initialize the classes $S_1, S_2, S_3, S_4, S_5, S_6$

Step 4: Construct optimal hyperplane $\boldsymbol{H} \rightarrow \alpha . D_i + d = 0$

Step 5: Find two supporting vectors $K_1 \rightarrow \alpha . D_i + d > 0$,$K_2 \rightarrow \alpha . D_i + d < 0$

Step 6: Compute '$\boldsymbol{AQI}$'

Step 7: Measure kernel '$\boldsymbol{\vartheta}$ (AQI_t, AQI_r)'

Step 8: If $(\boldsymbol{\vartheta}(AQI_t, AQI_r) = 1)$

Step 9: Classifies the data into a particular class

Step 10: End if

Step 11: Combine all weak learner ' $Z = \sum_{i=1}^{B} C_i$ '

Step 12: For eachC_i

Step 13: Assign the weight 'γ_i'

Step 14: Measure the quadratic error 'E_q'

Step 15: Update the weight 'γ_i'

Step 16: Find the weak learner with minimum error

Step 17: Return (strong classification results)

Step 18: End for

End

Algorithm 4.3: Process of QWEBC technique

4.3 Performance Analysis

The simulation result of proposed BTBSR-QWEBC model is evaluated by comparing with two different existing methods. The compared existing methods are such as IMD-VAE designed by Abdelkader Dairi et al. (2021) and Convolutional neural network-Long Short-Term Memory, Sparse Denoising Auto encoder (CNN–LSTM–SDAE) (CLS) model developed by K. Krishna Rani Samal et al. (2021) respectively. The experimental evaluation is carried out with the different parameters and the analysis is carried out with the help of following table values and graphs.

4.3.1 Evaluation of Forecasting Accuracy

The accuracy is determined by considering the quantity of air quality sample data from the dataset and calculating the amount of sample data that are accurately predicted to that number. The accuracy is computed using the following formula and given as a percentage (%).

$$APF_{acc} = \frac{D_{\text{accurately forecasted}}}{D_i} * 100 \qquad \ldots\ldots \text{Eqn (4.19)}$$

From Equation (4.19), Forecasting Accuracy 'APF_{acc}' is calculated depending on the 'D_i' amount of sample data. Here, '$D_{\text{accurately forecasted}}$' specifies number of data that is accurately forecasted.

Sample calculation:

Existing IMD-VAE: There are 20,000 total air quality data points and 15,486 successfully predicted air quality data points. Next, the accuracy of the air pollution forecast is computed as $APF_{acc} = \frac{15{,}486}{20{,}000} * 100 = 77.43\%$.

Existing CLS model: There are 16,428 air quality data points that have been properly predicted, out of a total of 20,000 data points. Next, the accuracy of the air pollution forecast is computed as $APF_{acc} = \frac{16{,}428}{20{,}000} * 100 = 82.14\ \%$.

Proposed BTBSR-QWEBC model: There are 20,000 total air quality data points and 17,250 properly predicted air quality points. Next, the accuracy of the air pollution forecast is computed as $APF_{acc} = \frac{17,250}{20,000} * 100 = 86.25\%$.

Table 4.2: Forecasting accuracy of existing methods vs BTBSR-QWEBC model

Number of air quality data	Air pollution forecasting accuracy (%)		
	Existing IMD-VAE	Existing CLS model	Proposed BTBSR-QWEBC
20,000	77.43	82.14	86.25
40,000	77.82	83.31	87.14
60,000	78.62	84.52	88.52
80,000	79.54	84.92	89.12
1,00,000	80.25	85.16	90.53
1,20,000	80.61	85.25	91.25
1,40,000	80.94	86.62	91.06
1,60,000	81.25	87.11	92.45
1,80,000	82.62	87.96	92.71
2,00,000	83.64	88.51	93.04

The experimental values on air pollution forecasting accuracy during data classification for air data prediction from dataset are shown in Table 4.2. The suggested BTBSR-QWEBC model is compared in the table to the IMD-VAE that was created by Abdelkader Dairi et al. (2021) and the CLS model that was created by K. Krishna Rani Samal et al. (2021), respectively.

Several air quality data points are taken into consideration for experimental purposes, ranging from 20,000 to 2,00,000 data points in the dataset. While increasing the number of data, the accuracy on forecasting air pollution also gets varied through a small difference. From the result, proposed model attained higher pollution forecasting accuracy on air data.

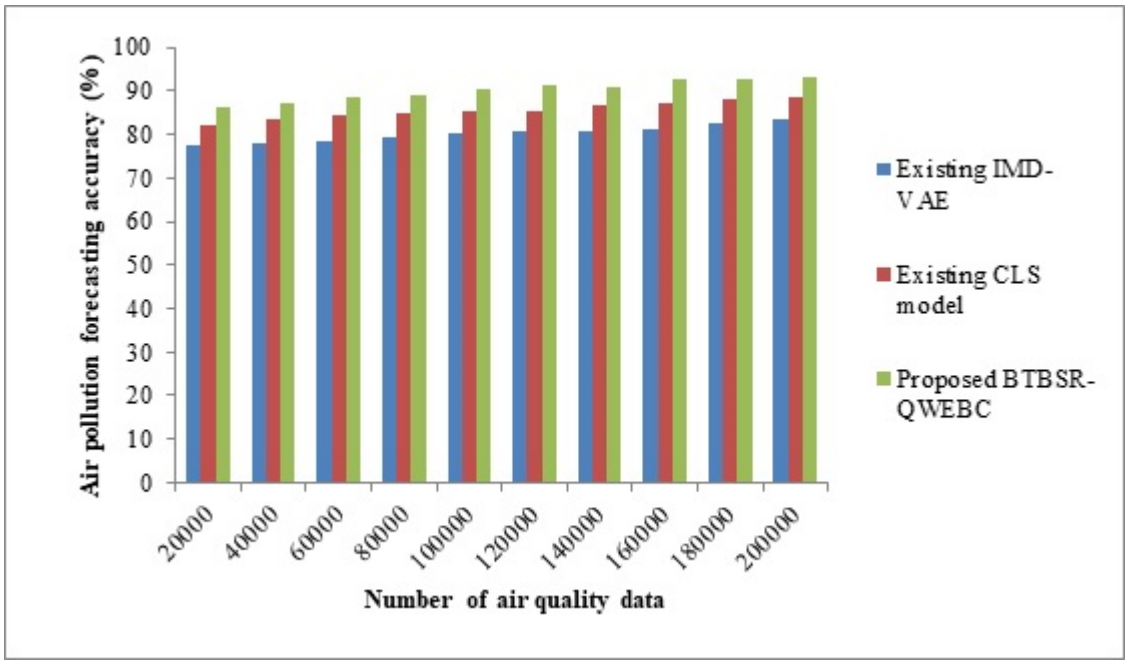

Figure 4.6: Forecasting Accuracy of BTBSR-QWEBC model

The results of the investigation of the accuracy of air pollution forecasting based on various numbers of air quality data are shown in Figure 4.6. The figure shows the comparison of proposed BTBSR-QWEBC model with existing IMD-VAE and CNN-LSTM-Sparse Denoising Auto encoder model. With the optimal selection of features, air pollution forecasting is significantly improved.

With the application of OIBSR, feature selection is performed to improve pollution forecasting with minimal time. With the selected optimal features, the classification algorithm is performed. The QWEBC technique is then used to categorize the data into various classes. Results for strong classification are obtained by evaluating feature values. Therefore, the air pollution prediction

accuracy is improved by 12% and 5% when compared to existing methods such as IMD-VAE designed by Abdelkader Dairi et al. (2021) and CLS model developed by K. Krishna Rani Samal et al. (2021) respectively.

4.3.2 Evaluation of Error Rate

The ratio of incorrectly projected air data to the total amount of input air quality sample data is known as the error rate. It is expressed mathematically as follows and is measured in percentages (%).

$$Error_{Rate} = \frac{D_{\text{forecasted wrongly}}}{D_i} * 100 \qquad \text{.... Eqn (4.20)}$$

From Equation (4.20), error rate '$Error_{Rate}$' is determined according to total number of data denoted as 'D_i'. Here, '$D_{\text{forecasted wrongly}}$' specifies number of data that forecasted wrongly.

Sample calculation:

Existing IMD-VAE: There are 20,000 air quality data points overall, of which 4,514 were incorrectly predicted. Next, the rate of error is ascertained as $Error_{Rate} = \frac{4,514}{20,000} * 100 = 22.57$ %.

Existing CLS model: There are 20,000 air quality data points overall, of which 3,572 were incorrectly predicted. Next, the rate of error is ascertained as $Error_{Rate} = \frac{3,572}{20,000} * 100 = 17.86$ %.

Proposed BTBSR-QWEBC Model: There were 2,750 incorrectly predicted air quality data out of a total of 20,000 data points. Next, the rate of error is ascertained as $Error_{Rate} = \frac{2,750}{20,000} * 100 = 13.75\%$.

Table 4.3: Error rate of existing methods vs BTBSR-QWEBC model

Number of air quality data	Error rate (%)		
	Existing IMD-VAE	Existing CLS model	Proposed BTBSR-QWEBC
20,000	22.57	17.86	13.75
40,000	22.18	16.69	12.86
60,000	21.38	15.48	11.48
80,000	20.46	15.08	10.88
1,00,000	19.75	14.84	9.47
1,20,000	19.39	14.75	8.75
1,40,000	19.06	13.38	8.94
1,60,000	18.75	12.89	7.55
1,80,000	17.38	12.04	7.29
2,00,000	16.36	11.49	6.96

The performance study of error rate for both existing and suggested approaches, considering varying numbers of air quality data, is presented in Table 4.3. To undertake experimental work, data in the range of 20,000 to 2,00,000 is obtained. The simulation is conducted by comparing proposed BTBSR-QWEBC model with existing methods such as IMD-VAE designed by Abdelkader Dairi et al. (2021) and CLS model developed by K. Krishna Rani Samal et al. (2021) respectively. It is evident from the above result that the suggested BTBSR-QWEBC model minimizes error rate more than the other techniques currently in use. Figure 4.7 shows the experimental error rate description for the suggested and current methods. Error rate is obtained with respect to various numbers of air quality data from the dataset.

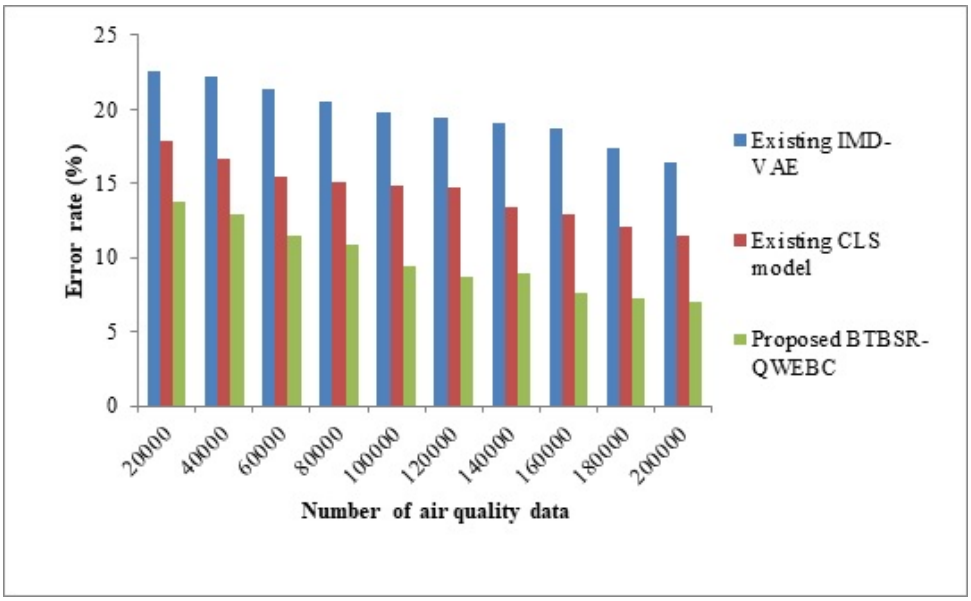

Figure 4.7: Error rate of BTBSR-QWEBC model

It provides the comparative result of proposed BTBSR-QWEBC model with existing IMD-VAE and CNN-LSTM-Sparse Denoising Auto encoder model. From the experimental analysis, proposed model provides reduced error rate compared to the existing techniques. By combining the results of weak learners, a strong classification result is obtained through the use of the ensemble technique. Based on the estimated quadratic error, strong data classification result is attained with minimum error. This helps to reduce error rate during data classification process. From the result analysis, the proposed BTBSR-QWEBC model reduces the error rate by 51% and 33% when compared with existing IMD-VAE designed by Abdelkader Dairi et al. (2021) and CLS model developed by K. Krishna Rani Samal et al. (2021) respectively.

4.3.3 Evaluation of Air Pollution Forecasting Time

The duration of time required for the classifier to categorize data in order to anticipate air pollution is known as the air pollution forecasting time. The unit of measurement is milliseconds (ms). The following formula is used to get the air pollution forecasting time.

$$APF_T = D_i * Time_{forecastingsingledata} \ldots\ldots \text{ Eqn (4.21)}$$

From Equation (4.21), forecasting time for air pollution 'APF_T' is calculated. Here, 'D_i' denotes sum of air quality sample data and '$Time_{forecastingsingledata}$' describes time consumed to forecast single data.

Sample calculation:

Existing IMD-VAE: One air quality value is anticipated in 0.1215 seconds out of 20,000 available data points. Next, the air pollution prediction time is given as $APF_T = 20{,}000 * 0.1215 = 2{,}430ms$.

Existing CLS model: Each air quality data point requires 0.099 seconds of processing time in total to forecast 20,000 data points. Next, the air pollution prediction time is given as $APF_T = 20{,}000 * 0.099 = 1{,}980\ ms$.

Table 4.4: Forecasting time of existing methods vs BTBSR-QWEBC model

Number of air quality data	Air pollution forecasting time (ms)		
	Existing IMD-VAE	Existing CLS model	Proposed BTBSR-QWEBC
20,000	2430	1980	1600
40,000	2360	1960	1560
60,000	2300	1860	1500
80,000	2200	1820	1470
1,00,000	2100	1780	1450
1,20,000	2020	1750	1430
1,40,000	1980	1700	1400
1,60,000	1920	1660	1380
1,80,000	1850	1630	1300
2,00,000	1700	1600	1250

Proposed BTBSR-QWEBC Model: There are 20,000 air quality data points and it takes 0.08 seconds to forecast one air quality data point. Next, the predicting time for air pollution is expressed as $APF_T = 20{,}000 * 0.08 = 1{,}600\ ms$.

The performance study of forecasting time for various data in the range of 20,000 to 2,00,000 is presented in Table 4.4. Using the dataset, the simulation is run by contrasting suggested and current approaches. Here, the evaluation of proposed BTBSR-QWEBC model is made with existing IMD-VAE designed by Abdelkader Dairi et al. (2021) and CLS model developed by K. Krishna Rani Samal et al. (2021) respectively. According to various data, the time taken for forecasting gets varied. From the results, the BTBSR-QWEBC model achieves minimized time than other existing methods.

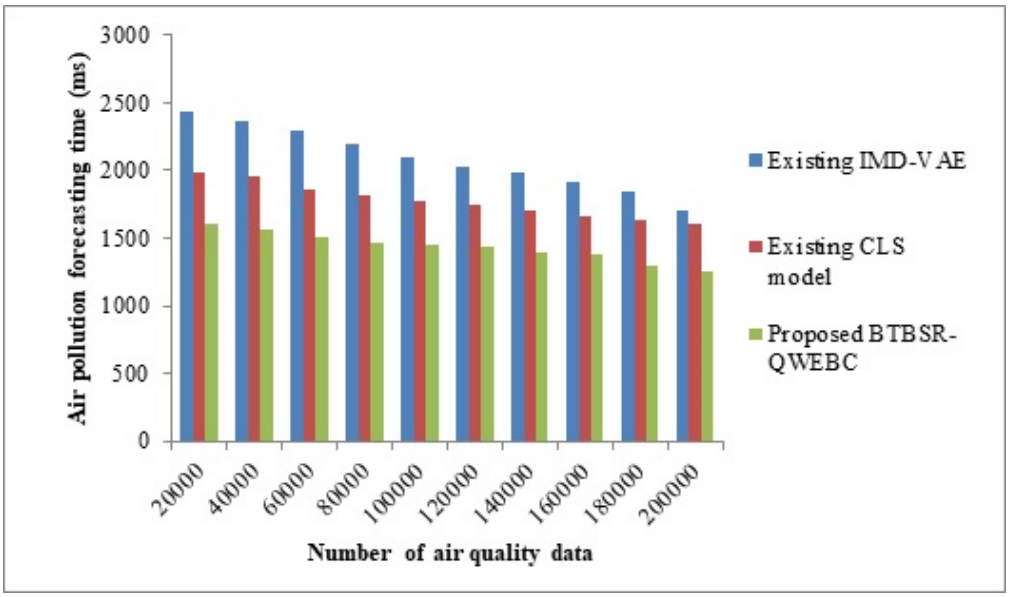

Figure 4.8: Forecasting time of BTBSR-QWEBC model

The performance analysis of the air pollution forecasting time for various quantities of air quality data is displayed in Figure 4.8. The figure shows the data forecasting time using both existing and recommended methodologies. Performance studies show that the proposed BTBSR-QWEBC model anticipates air

pollution more quickly than the existing methods. Pollution forecasting is completed faster because pre-processing and feature selection are present. Due to the application of BDZWT-based pre-processing and OIBSR-based feature selection, forecasted air data is accurately predicted. The wavelet transformation process removes the noisy data considered from dataset

Then the regression function selects significant relevant features by estimating Otsuka correlation index. The selected features help to efficiently reduce the time. Therefore, the forecasting time is considerably reduced by 31% and 19% when compared to existing IMD-VAE designed by Abdelkader Dairi et al. (2021) and CLS model developed by K. Krishna Rani Samal et al. (2021) respectively.

4.3.4 Performance Analysis of Memory Consumption

The amount of storage space needed to anticipate air pollution using air quality data from an input dataset is known as memory consumption. Thus, memory consumption is computed in megabytes (MB) and formulated as follows.

$$Memory_{Con} = D_i * SS\ (feorecastingsingledata) \quad \ldots \text{Eqn (4.22)}$$

From Equation (4.22), memory consumption '$Memory_{Con}$' is calculated with respect to 'D_i' number of air quality data and storage space for forecasting single data '$SS\ (forecastingsingledata)$'.

Sample calculation:

Existing IMD-VAE: Memory space consumed to forecast single air quality data is 0.0055 and the entire number of data is 20,000. Then, the memory consumption is calculated as $Memory_{Con} = 20{,}000 * 0.0055 = 110\ MB$.

Existing CLS model: Memory space consumed to forecast single air quality data is 0.006 and the entire number of data is 20,000. Then, the memory consumption is calculated as $Memory_{Con} = 20{,}000 * 0.006 = 120\ MB$.

Proposed BTBSR-QWEBC Model: Memory space consumed to forecast single air quality data is 0.0049 and the entire number of data is 20,000. Then, the memory consumption is calculated as $Memory_{Con} = 20{,}000 * 0.0049 = 98\ MB$.

Table 4.5: Memory consumption of existing methods vs BTBSR-QWEBC model

Number of air quality data	Memory consumption (MB)		
	Existing IMD-VAE	Existing CLS model	Proposed BTBSR-QWEBC
20,000	110	120	98
40,000	105	116	94
60,000	100	112	90
80,000	95	110	87
1,00,000	94	105	84
1,20,000	92	102	82
1,40,000	90	98	80
1,60,000	85	95	78
1,80,000	80	93	76
2,00,000	76	90	71

Table 4.5 displays the memory consumption performance study for various data amounts in the range of 20,000 to 2,000,000. Using an air quality sample dataset, the simulation is run by contrasting the suggested and current approaches. Here, proposed BTBSR-QWEBC model is compared with existing IMD-VAE designed by Abdelkader Dairi et al. (2021) and CLS model developed by K. Krishna Rani Samal et al. (2021) respectively. From the result, the proposed BTBSR-QWEBC model resulted with minimum memory consumption than the other existing methods. Figure 4.9 shows the memory usage simulation study based on the quantity of air quality data. It shows the evaluation result of proposed BTBSR-QWEBC model with current IMD-VAE and CNN-LSTM -Sparse

Denoising Auto encoder model. While increasing the number of data, the space required for storing data gets increased correspondingly. But comparatively, BTBSR-QWEBC model attains minimum consumption of space storage than the other existing methods.

The proposed model applies BDZWT technique to remove noisy data from dataset. Following this, OIBSR technique is presented to select pertinent features based on regression coefficient index. With selected features, ensemble boost classification is performed to classify data into different class for obtaining strong classification result. The relevant features help to minimizes memory for storing data. Thus, the memory consumption is considerably reduced by 9% and 19% using BTBSR-QWEBC model when compared to existing IMD-VAE designed by Abdelkader Dairi et al. (2021) and CLS model developed by K. Krishna Rani Samal et al. (2021) respectively.

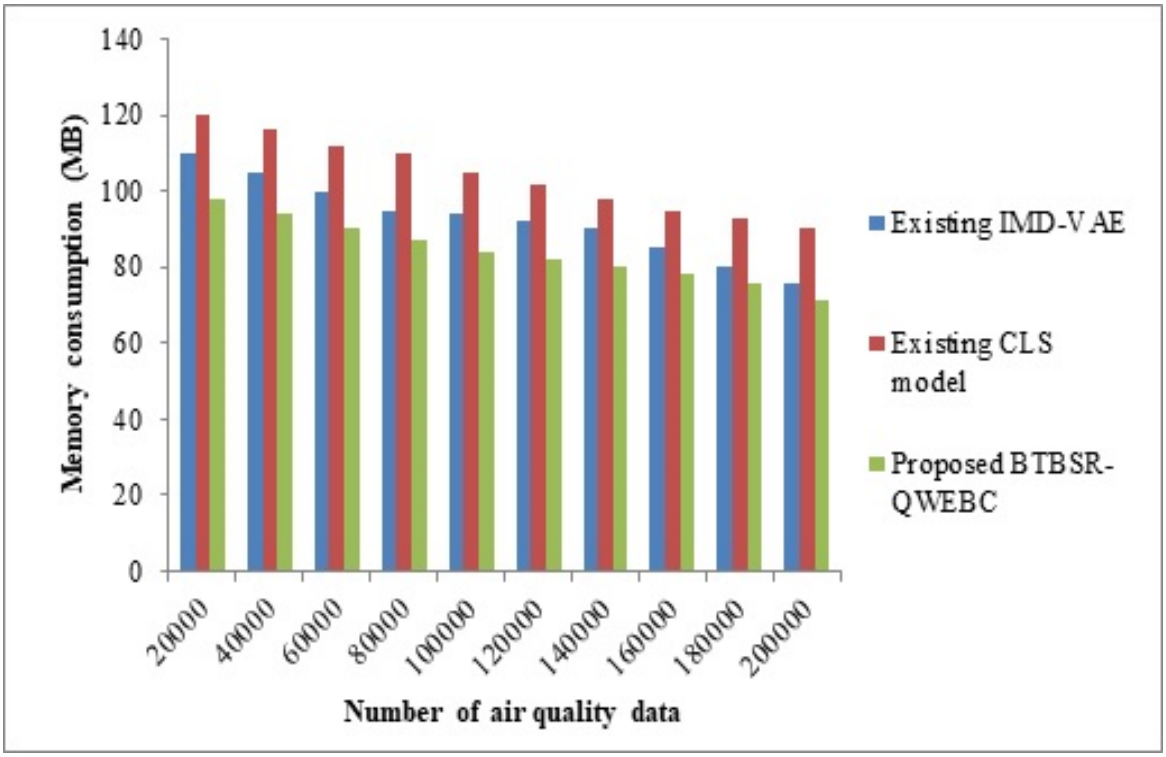

Figure 4.9: Memory consumption of BTBSR-QWEBC model

4.4 Summary

An efficient BTBSR-QWEBC model is proposed for efficient air pollution monitoring and controlling with higher accuracy. The key drive of the proposed model is to achieve enhanced result of forecasting with minimum time. Initially, the multiple amount of data is considered as input from database. At first, data pre-processing is performed using BDZWT technique to remove noise data from dataset for efficient data classification. With obtained pre-processed data, the feature selection procedure is achieved by applying OIBSR-based feature selection. Based on estimated regression coefficient value, relevant features are selected for efficient data classification. By using selected relevant features, QWEBC technique is performed at last to classify data into different classes. It constructs sum of weak learners and results are joined to form robust classification result with minimized error. Therefore, the proposed BTBSR-QWEBC model improves air pollution forecasting performance with enhanced accuracy and minimum memory consumption. It consumes more time for forecasting air pollution data with minimum error rate. For better forecasting air pollution with minimum time and error rate than BTBSR-QWEBC model, next model is proposed.