# Air Pollution Forecasting using Novel Feature Selection and Classification Model

M.Dhanalakshmi[1], Dr. V. Radha[2]
[1] *Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore.*
[2] *Professor, Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore.*

**ABSTRACT**
Quality of air prediction was the difficult function towards dynamic nature, instability, as well as better inconsistency within the time and space. Especially in urban areas, Air Pollution is becoming more and more, and the various pollutants affect the air quality. Hence, it is essential to accurately predict Air Pollution for providing hazardous impact earlier. The existing machine learning methods have been developed but it is difficult to forecast accurate pollutant and particulate levels and to predict the air quality index. Bilateral Transformative Broken-Stick Regression-based Quadratic Weighted Emphasis Boost Classification (BTBSR-QWEBC) technique is introduced for IoT-based Air Pollution Forecast with higher accuracy and minimum time consumption for increasing accuracy of air pollution forecasting. From the BTBSR-QWEBC, IoT devices are used to collect Air Quality data. The BTBSR-QWEBC technique includes three major processes namely pre-processing, Feature Selection, and classification. That Technique helps to improve the accuracy of the Air Pollution Forecast and to minimize time consumption. Experimental assessment is performed by various metrics namely Air Pollution Forecast accuracy, error rate, as well as Air Pollution Forecasting time and space complexity. The observed results display the BTBSR-QWEBC technique provides better accuracy as well as minimal time than conventional techniques.

**Keywords— Air Pollution Forecast, feature selection, classification Technique, AQI, Regression**

## 1. Introduction

Air is one of the most important factors for the entire living creatures on the earth. Due to rapid industrialization, air pollution has become a significant problem in all developed and developing countries. Air pollution forecasting is an important step for air quality pollution management to decrease pollution’s negative impact on the environment and people’s health conditions. The entire existing forecasting model generally performs Air pollution forecasting and fails to execute the forecasting modelling effectively.

## 2. Experimental Methodology

Air pollution monitoring is a significant and challenging problem since it manages the surroundings and strengthens air pollution. In the real dataset, the number of air pollutants and the instances within training set increases danger. Therefore, feature selection was necessary for reducing dimensionality of dataset. Moreover, the raw data collected with the help of IoT devices comprises noisy data resulting increase time and space complexity. These types of problems were developed by novel technique called BTBSR-QWEBC based on three different processes. These three processes of the BTBSR-QWEBC technique are described in this section.

Figure 1 given above illustrates the architecture of the BTBSR-QWEBC technique includes three different method such as pre-processing, feature selection, as well as classification. IoT devices are used to collect the air Quality data. First, pre-processing is performed using bilateral discretized Z-transform to obtain noise-free dataset. Second, Otsuka inducive Broken-stick regression was utilized for choosing the significant function for minimizing time complexity of air pollution forecasting. Finally, the Quadratic Weighted Emphasis Boost technique is applied for classifying the data by analysing the selected features. Based on the classification, accurate air pollution forecasting is performed by measuring the air quality index. An elaborate explanation of the proposed BTBSR-QWEBC technique is presented in the following subsection.

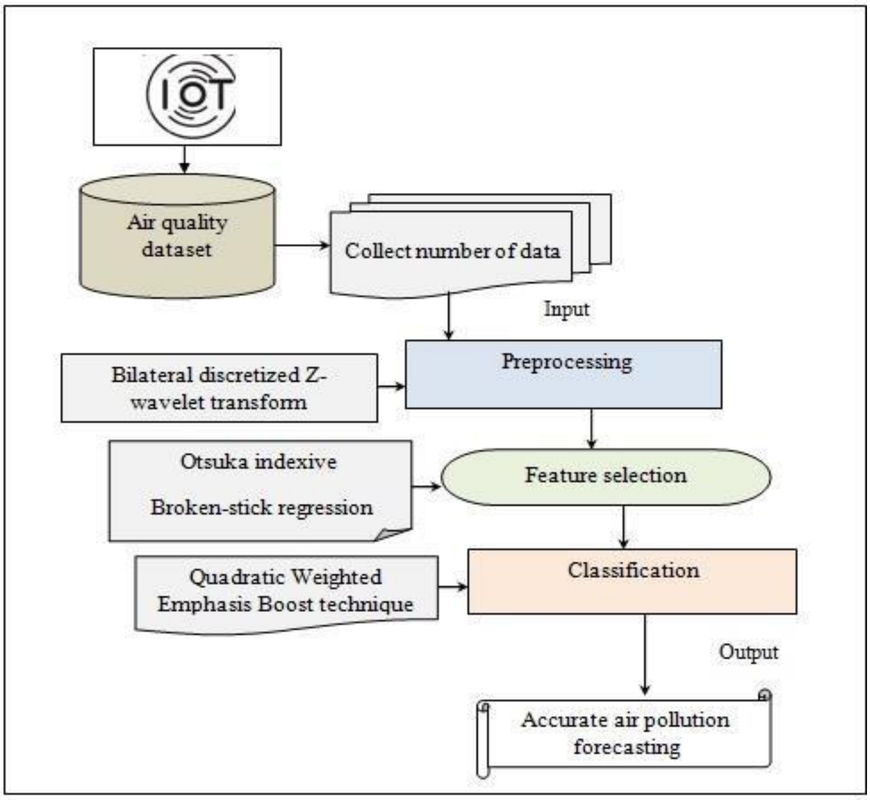

IOT
Air quality dataset
Collect number of data
Input
Bilateral discretized Z-wavelet transform
Preprocessing
Otsuka indexive Broken-stick regression
Feature selection
Quadratic Weighted Emphasis Boost technique
Classification
Output
Accurate air pollution forecasting

**Figure 1 architecture of proposed BTBSR-QWEBC technique**

## 3. Otsuka inducive Broken-stick regression

The second process of the BTBSR-QWEBC technique is to perform the feature selection using Otsuka inducive Broken-stick regression. Feature selection is the significant step while building machine learning. The major process was used for finding finest feasible features for building a

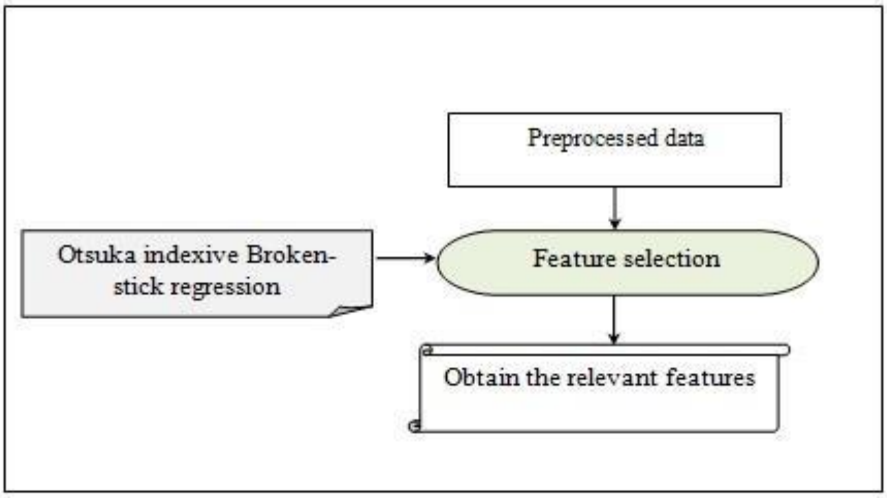

Preprocessed data
Otsuka indexive Broken-stick regression
Feature selection
Obtain the relevant features

**Figure 3 Flow diagram of Otsuka indexive Broken-stick regression**

machine learning model. Relevant feature selection was used for minimizing air pollution forecasting time.

Figure 3 illustrates the flow process of Otsuka inducive Broken-stick regression-based feature selection to obtain the relevant features for minimizing the time complexity. Let us consider the number of features '$\beta 1,2$, $\beta 3$, … . $\beta n$,', Then, the Broken-stick regression was the statistical procedures to evaluate interaction among dependent variable (features) using Otsuka similarity index. Broken-stick regression is used to segment the input into two parts based on breakpoint. It is significant for decision-making based on Otsuka similarity index. Otsuka similarity index was utilized for calculating similarity among features.

$$\rho = \frac{\sum \beta_{i,} \beta_j}{\sqrt{\sum B_i^2 \sum \beta_j^2}}$$

Where, '$\rho$' indicates similarity coefficient is measured between features '$\beta_i$,' and '$\beta_i$ '
The similarity coefficient values range between -1 and +1.

### 3.1 Quadratic weighted emphasis boost technique

Finally, the proposed BTBSR-QWEBC technique performs the Classification to forecast air pollution using the Quadratic weighted emphasis boost technique with the objective of improving both accuracy and time involved in air pollution monitoring. The quadratic weighted emphasis boost technique was the machine learning ensemble classification. Weak learner was the classifier which difficult for offering true classification. In contrast, strong learner is offering a true classification.

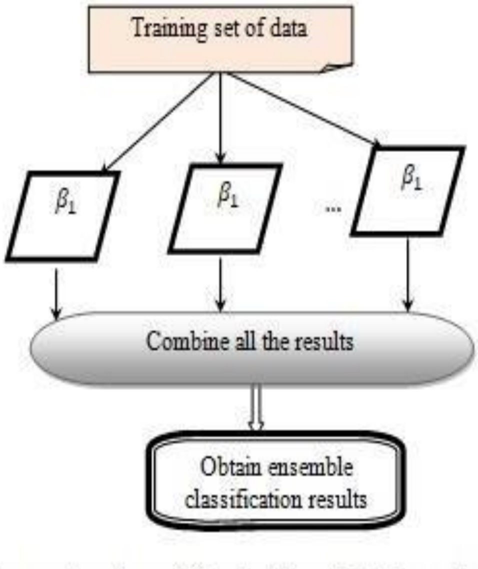


**Figure 4 schematic construction of Quadratic weighted emphasis boost technique**

Figure 4 depicts the schematic construction of the weighted emphasis boost ensemble technique for accurate forecasting by lesser time. Weighted emphasis boost ensemble technique considers the input as a training sample set $\{D_i, Z\}$ where $D_i = D_1, D_2, \dots, D_m$' denotes the sample air data and $Z$ specifies the ensemble classification outcomes. As shown in figure 3, the boost technique initially constructs '$B$' set of weak learners $C_1$, $C_{2,3}$, … . $C_B$ and the outcomes are summed to build strong classification results. The weighted emphasis boost ensemble technique uses the weak learner as a Kernelized support vector classifier for forecasting the air pollution with the selected feature set.

Let us consider the training sample set $\{D_i, Z\}$ that provide results whose value was established. Main aim of Kernelized support vector classifier was used for producing best line or decision boundary separates the n- dimensional within different classes. Extent of hyperplane based on number of output classes. Data points were the closest towards hyperplane as well as influence hyperplane were called as Support Vector.

A separating hyperplane with the input that satisfies equation,

$\boldsymbol{H} \rightarrow \alpha . D_i + d = 0$

$\boldsymbol{H}$ indicates hyperplane or decision boundary, indicates training samples (i.e. input data i.e. air pollutants), '$d$' denotes bias as well as $\alpha$ indicates normal weight vector to hyperplane ($H$). If the training samples were linearly divisible, different parallel support vectors were selected that divide the input into different classes. Hence, input data belongs to decision boundary by,

$K_1 \rightarrow \alpha . D_i + d > 0$

$K_2 \rightarrow \alpha . D_i + d < 0$

Where, $K_1$, $K_2$ denotes support vectors that are positioned above as well as below boundary. Predicted output ($\boldsymbol{y}'$) of support vector using kernel function by,

$\boldsymbol{Z} = \sum \alpha \, \boldsymbol{y_i} \, \boldsymbol{\vartheta} \, ( t, AQI_r )$

Where $\boldsymbol{Z}$ represents predicted classification results, $\boldsymbol{\vartheta}$ ($t$ , $AQI_r$ ) indicates kernel function calculates relationship among Testing Air Quality Index value (i.e. $AQI_t$ ) and Training Air Quality Index (i.e. $AQI_r$ ), $\alpha$ denotes the weights of the training samples.

The air quality index for each sampled data is calculated based on the average of air pollutant concentrationsselected from the feature selection process (i.e. $PM2.5$, $PM10$, $SO2$, $NOx$, $NO2$) and the maximum value of CO (Carbon monoxide) and O3 (Ozone) respectively.

$$AQI = (PM2.5, PM10, SO2, NOx, NO2) + Max\ (CO, O3)$$

From the above equation (11), the air quality index value '$AQI$' is measured based on the average values of PM2.5, PM10, SO2, NOx, NH3, and the maximum values of CO and O3 respectively. Here, the Laplace RBF kernel is applied to measure the relationship between the Testing Air Quality Index value and Training Air Quality Index. The Laplace RBF kernel '$\boldsymbol{\vartheta}$ ($AQI_t$, $AQIr$)' is expressed as follows,

$$\boldsymbol{\vartheta}\ (AQI_t\,, AQI_r) = exp\ (-\|AQI_t - AQI_r\|^2 \,/\, v2)$$

Where '$v$' indicates a deviation. The Training Air Quality Index which is more like the Testing Air Quality Index value is classified as a particular class. In other words, the computed Training Air Quality Index is closer to the Testing Air Quality Index value being classified as a particular class. Here six different classes of air quality prediction are considered as good, satisfactory, moderate, poor, very poor, and severe. Laplace RBF kernel function provides the similarity ranges from 0 to 1. If the similarity is high (i.e. 1), then the accurate classification were attained. Thus, different classes of air quality estimation are obtained.

The observed weak learner results have some training errors during the classification. Therefore, the weak learner results are summed to make a strong classification result.

$$Z = \sum Ci\ B\ i{=}1$$

Where, $Z$ indicates ensemble output, $Ci$ indicates weak learner. Weight gets initialized for making strong classification outcomes.

$$Z = \sum Ci\ B\ i{=}1 * \gamma i$$

Where, '$\gamma i$ ' indicates weak learner results. Weight was a random integer. Ensemble technique uses weighted emphasis function to measure the quadratic error of weak classification results,

$$Eq = \exp[p((\sum Ci\ B\ i{=}1\ \gamma i - Z)2 - (1 - p)(\sum Ci\ b\ i{=}1\ )\ 2\ )]$$

$Eq$ indicates weighted emphasis function, $p$ indicates a weighting parameter, $Z$ represents actual ensemble classification results, $\sum C i \gamma i\ B\ i$=1 indicates a predicted classification result of weak learner with the weight $\gamma i$ and the without weight $\sum C i\ B\ i$=1 .

$\vartheta$ denotes a weighting parameter value is set to 1 and obtain the final quadratic error,

$$Eq = \exp[(\sum C i \gamma i\ B\ i=1 - Z)2]$$

Finally, the weak learner weight is efficient on above-calculated error rate. If weak learner is properly categorized, weight is decreases. If not, initial weight value is improved. Weak learner by lesser error was selected by the strong classification outcome using better accuracy.

## 4. Results and Discussion

The performance of BTBSR-QWEBC and two existing IMD-VAE [1] CLS [2] are discussed with respect to different parameters namely air pollution forecasting accuracy, error rate, air pollution forecasting time. These metrics are described as given below.

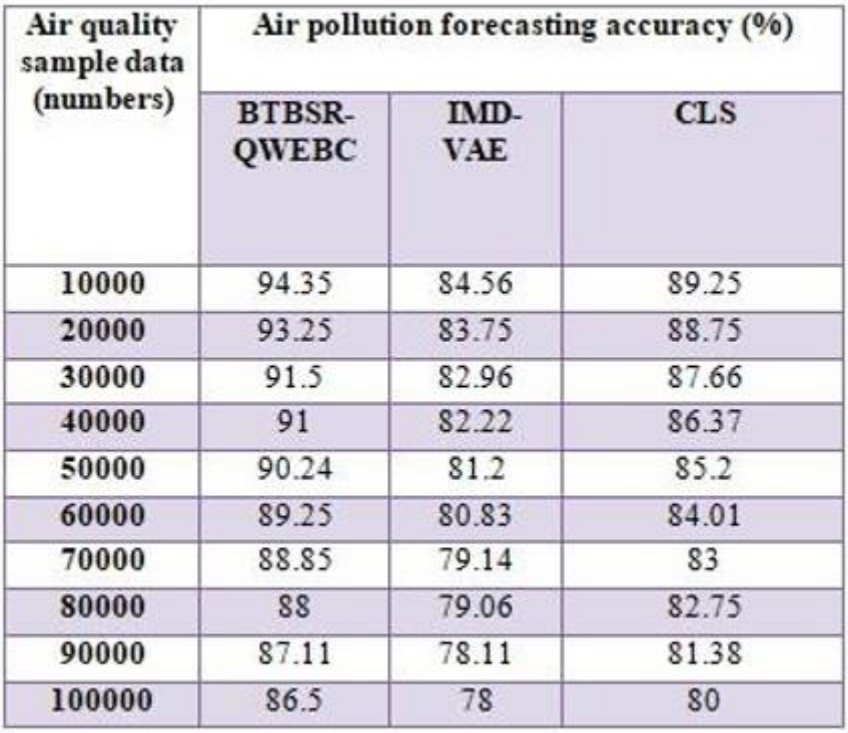
### 4.1 Impact of air pollution forecasting accuracy

Air pollution forecasting accuracy was calculated by proportion of air quality sample data were accurately forecasted to entire number of air quality sample data considered for experiential evaluation. The formulation for air pollution forecasting accuracy is given below.

$APFacc = \sum DFAcc\ Di * 100n\ i$=1

Where, $APFacc$ indicates an air pollution forecasting accuracy, $DFAcc$ denotes an air quality data forecasted accurately, $Di$' denotes the total number of air quality sample data. It was calculated by percentage (%).

**Table 3: Air Pollution Forecasting Accuracy**



| Air quality sample data (numbers) | Air pollution forecasting accuracy (%) | | |
|---|---|---|---|
| | BTBSR-QWEBC | IMD-VAE | CLS |
| 10000 | 94.35 | 84.56 | 89.25 |
| 20000 | 93.25 | 83.75 | 88.75 |
| 30000 | 91.5 | 82.96 | 87.66 |
| 40000 | 91 | 82.22 | 86.37 |
| 50000 | 90.24 | 81.2 | 85.2 |
| 60000 | 89.25 | 80.83 | 84.01 |
| 70000 | 88.85 | 79.14 | 83 |
| 80000 | 88 | 79.06 | 82.75 |
| 90000 | 87.11 | 78.11 | 81.38 |
| 100000 | 86.5 | 78 | 80 |

Table 3 reports the performance analysis of the Air pollution forecasting accuracy versus the number of Air quality sample data taken in the ranges from 10000 to 100000. For each method, ten varieties of performance results were maintained by a different number of inputs. Table 3 shows Air pollution forecasting accuracy by various techniques namely BTBSR-QWEBC as well as two existing IMD-VAE [1] CLS [2]. From the observed results, the BTBSR-QWEBC performs improved accuracy of two methods. Let us consider 10000 Air quality sample data to measure accuracy. BTBSR-QWEBC, 94.35% of accuracy was maintained where accuracy of IMD-VAE [1] and CLS [2] are 84.56% and 89.25% respectively. Likewise, nine different outcomes were observed as well as overall accuracy of BTBSR-QWEBC was compared with conventional techniques. Ten comparison outcomes represent Air pollution forecasting accuracy of BTBSR-QWEBC was improved as 11% and 6% of existing techniques.

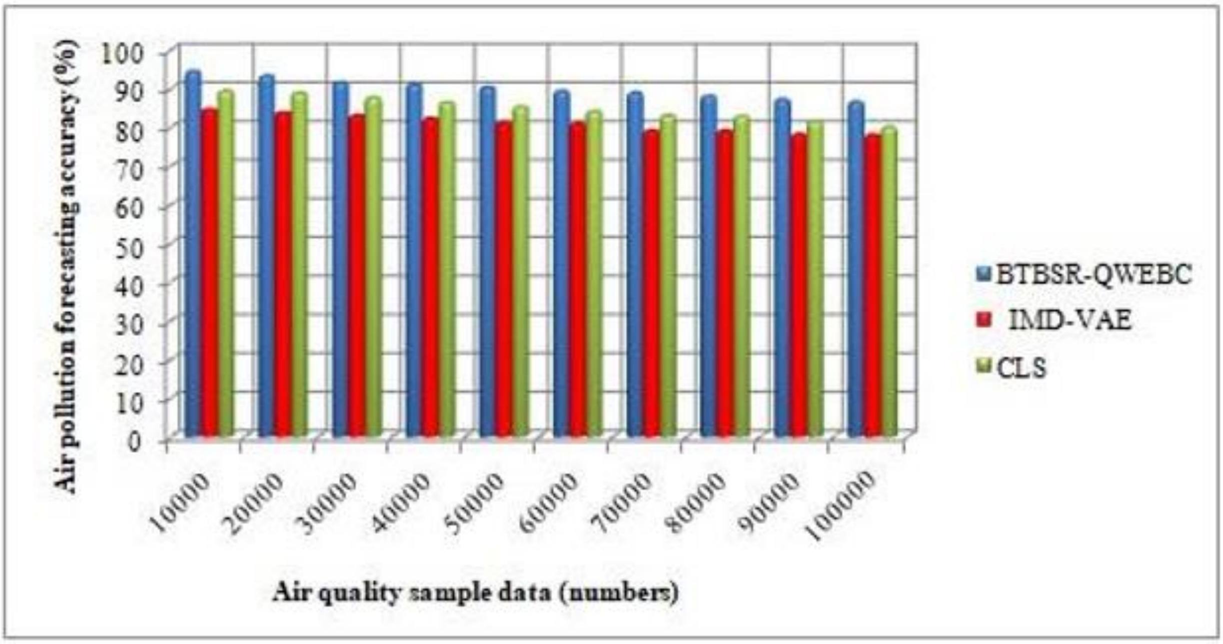


**Figure 6 Air pollution forecasting accuracy versus Air quality sample data**

Figure 6 portrays the comparison of Air pollution forecasting accuracy using three different methods namely BTBSR-QWEBC and two existing IMD-VAE [1] CLS [2]. The graphical chart indicates that the Air pollution forecasting accuracy is observed at the 'y' axis and air quality sample data are given to the 'x-axis. BTBSR-QWEBC enhances accuracy through utilizing Quadratic Weighted emphasis boost ensemble classification technique. Kernelized support vector was employed for analysing the testing and training data. Ensemble method combines weak learners as well as provides strong classification results.

The proposed ensemble technique set weak learners as a Kernelized support vector classifier with the input samples. The weak learner initializes the number of classes and computes the air quality index. Then Laplace RBF kernel is used to compute the relationship between the raining and testing '$AQI$' of a particular class. Weak learner results are summed and assigned the weight. The emphasis was applied for measuring quadratic error. Weak learner using least quadratic error was selected by last strong classification result. In this way, accurate air quality forecasting is performed with minimum time.

## Conclusion

Air quality is observed by challenging issue of early air pollution caution. Identifying different pollutant factorsthat contribute to air pollution plays a fundamental function for achieving efficient scheme for decrease air pollution. BTBSR-QWEBC was effective as well as successful with higher accuracy and minimum time consumption. By designing a Bilateral discretized Z- wavelet transform, the noise data from the dataset is removed. Followed by, Otsuka indexive Broken-stick regression-based feature selection is performed to find the relevant features. Hence it minimizes the time utilization and memory consumption for Air pollution forecasting. Next, the Quadratic Weighted emphasis boost ensemble classification is applied to predict future outcomes by constructing the number of weak learners. The weak learners are combined to make a strong by finding the minimum error. As a result, accurate classification is performed resulting in improves forecasting performance results. The experimentation results of the BTBSR-QWEBC technique and existing classification techniques are estimated with different metrics such as Air pollution forecasting accuracy, error rate, Air pollution forecasting time, and Memory consumption. The experimental results show that the BTBSR-QWEBC technique achieves higher forecasting accuracy and minimum time, Memory consumption as well as error rate.

## References

**1.** Abdelkader Dairi, Fouzi Harrou, Sofiane Khadraoui, and Ying Sun, "Integrated Multiple Directed Attention- based Deep Learning for Improved Air Pollution Forecasting", IEEE Transactions on Instrumentation and Measurement, Volume 70, 2021, Pages. DOI: 10.1109/TIM.2021.3091511.

**2.** K. Krishna Rani Samal, Ankit Kumar Panda, Korra Sathya Babu, Santos Kumar Das, "An improved pollution forecasting model with meteorological impact using multiple imputation and fine-tuning approach", Sustainable Cities and Society, Elsevier, Volume 70, 2021, Pages 1-13. https://doi.org/10.1016/j.scs.2021.102923

**3.** Azim Heydari, Meysam Majidi Nezhad, Davide Astiaso Garcia, Farshid Keynia & Livio De Santoli, "Air pollution forecasting application based on deep learning model and optimization algorithm", Clean Technologiesand Environmental Policy, Springer, Volume 24, 2022, pages 607–621. https://doi.org/10.1007/s10098-021-02080-5

**4.** Gao Huang, Chunjiang Ge, Tianyu Xiong, Shiji Song, Le Yang, Baoxian Liu, Wenjun Yin & Cheng Wu, "Largescale air pollution prediction with deep convolutional networks", Science China Information Sciences, Elsevier, Volume 64, 2021, Pages 1-11. https://doi.org/10.1007/s11432-020-2951-1

**5.** Abdellatif Bekkar, Badr Hssina, Samira Douzi & Khadija Douzi, "Air-pollution prediction in smart city, deep learning approach", Journal of Big Data, Springer, Volume 8, 2021, Pages 1-21

**6.** Bingchun Liu, Xiaogang Yu, Jiali Chen, Qingshan Wang, "Air pollution concentration forecasting based on wavelet transform and combined weighting forecasting model", Atmospheric Pollution Research, Elsevier, Volume 12, 2021, Pages 1-11. https://doi.org/10.1016/j.apr.2021.101144

**7.** Philipp Hähnel, Jakub Mareˇcek, Julien Monteil, FearghalO' Donncha, "Using deep learning to extend the range of air pollution monitoring and forecasting", Journal of Computational Physics, Elsevier, Volume 408, 2020, Pages 1-13, https://doi.org/10.1016/j.jcp.2020.109278

**8.** Doreswamy, Harishkumar K S, Yogesh KM, Ibrahim Gad, "Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models", Procedia Computer Science, Elsevier, Volume 171, 2020, Pages 2057–2066. https://doi.org/10.1016/j.procs.2020.04.221

**9.** Jianzhou Wang, Hongmin Li, Hufang Yang, Ying Wang, "Intelligent multivariable air-quality forecasting system based on feature selection and modified evolving interval type-2 quantum fuzzy neural network", Environmental Pollution, Elsevier, Volume 274 , 2021, Pages 1-17. https://doi.org/10.1016/j.envpol.2021.116429

**10.** Ekta Sharma; Ravinesh C. Deo; Ramendra Prasad; Alfio V. Parisi; Nawin Raj, "Deep Air Quality Forecasts: Suspended Particulate Matter Modeling With Convolutional Neural and Long Short-Term Memory Networks",IEEE Access ,Volume 8, 2020, Pages 209503 – 209516

**11.** Shengdong Du; Tianrui Li; Yan Yang; Shi-Jinn Horng, "Deep Air Quality Forecasting Using Hybrid Deep Learning Framework", IEEE Transactions on Knowledge and Data Engineering , Volume 33, Issue 6, 2021, Pages 2412 – 2424. DOI: 10.1109/TKDE.2019.2954510

**12.** Ning Jin; Yongkang Zeng; Ke Yan; Zhiwei Ji, "Multivariate Air Quality Forecasting With Nested Long Short Term Memory Neural Network", IEEE Transactions on Industrial Informatics , Volume 17, Issue 12, 2021, Pages 8514 – 8522. DOI: 10.1109/TII.2021.3065425

**13.** S Abirami and, P Chitra, "Regional air quality forecasting using spatiotemporal deep learning", Journal of Cleaner Production, Elsevier, Volume 283 , 2021, Pages 1-14. https://doi.org/10.1016/j.jclepro.2020.125341

**14.** Dewen Seng a,b,*, Qiyan Zhang a, Xuefeng Zhang a,b, Guangsen Chen a, Xiyuan Chen, "Spatiotemporal prediction of air quality based on LSTM neural network", Knowledge-Based Systems, Elsevier, Volume 233, 2021,Pages 1-12. https://doi.org/10.1016/j.aej.2020.12.009

**15.** Yu Huang, Josh Jia-Ching Ying, Vincent S. Tseng, "Spatio-attention embedded recurrent neural network for airquality prediction", Knowledge-Based Systems, Elsevier, Volume 233, 2021, Pages 1-14. https://doi.org/10.1016/j.knosys.2021.107416