CHAPTER 3

FORECASTING AIR POLLUTION USING A MULTICLASS SUPPORT VECTOR MODEL

3.1 Introduction

The biggest threat to people, animals, crops, cities, forests, and aquatic ecosystems is air pollution. Climate variations are linked to all pollutants and have the potential to create air pollution. People's health is impacted by pollution since the atmosphere contains harmful substances. Maintaining air pollution levels benefits human health, development, and the environment by improving air quality. Air pollution comes in a variety of forms, such as gases and biologically based particulate matter. Carbon monoxide, ammonia, sulfur dioxide, methane, nitrous oxides, chlorofluorocarbons, and carbon dioxide are the main contributors to gas pollution. Likewise, when determining particle pollution, both organic and inorganic components are considered. Accordingly, The WHO announced that air pollution is a sky-scraping environmental hazard. Thus, the techniques are developed to predict air pollution quality through Machine Learning (ML) by monitoring it with IoT devices.

Various ML techniques are highly developed to predict air pollution with better quality and provide forecast risk. Additionally, monitoring vehicle emissions, air direction, wind speed and so on are estimated by monitoring and controlling air pollution. This technique supports managing hazards caused to animals, humans and forests. More research still needs to be developed to obtain enhanced air pollution forecasting, though quality maintenance is complex owing to its speed and climatic changes. In order to provide air pollution foresting, many research works have been developed. Still, the prediction is a challenging task to achieve a better result. The pre-processing and feature selection process is carried out for efficient air pollution forecasting. The noisy air is removed to forecast air pollution

effectively. The compelling feature extraction performance helps in providing better results in predicting air pollution with higher accuracy. Feature selection is essential for extracting information by selecting the relevant features of air quality data. In data analysis, feature selection is one of the important ways to select the optimal features. Several techniques have been introduced for handling air quality data, but it resulted in high time. However, achieving an accurate prediction in the least amount of time is quite challenging. Therefore, a novel technique is required to improve the predictive performance of forecasting.

A LR-MSV air pollution forecast model is projected to enhance air pollution prediction performance. The model considers features from the air quality dataset for attaining higher accuracy on air pollution forecast with minimum time and error rate. From the dataset, multiple air quality parameter features are collected with different features. Initially, data pre-processing is performed by applying WSW based Multi-resolute pre-processing model. During pre-processing, noise in air data is removed, and attains valuable data for the classification process. With obtained pre-processed data, the feature extraction process is achieved using LRC-based Feature Selection model. The measurement of linear regression function helps to recognize optimal pertinent features and inappropriate features from the dataset. Followed by loss value is estimated using the gradient function. Here, the gradient function is applied to select relevant features by eliminating irrelevant features with minimal loss. The selected features help to classify air data effectively with reduced time. Lastly, the input data is classified using the MSV based Air Pollution Forecast model. For the purpose of tracking air prediction, linear classifiers efficiently classify air data. Based on the classification outcomes, the air pollution data is anticipated with the least amount of time and error. The LR-MSV model has been proposed to classify data for air pollution forecasting. Variables such as error rate, air pollution forecasting accuracy, and air pollution predicting time are tested with the proposed model. The performance analysis of the proposed model resulted in improved air pollution forecasting with reduced time and error rate.

3.2 Methodology

The air pollution prediction process is achieved by proposing an LR-MSV Air Pollution Forecast model. Depends on the designed model, prediction is carried out and prevents air pollution with a reduced error rate. The proposed model is performed with a combination of pre-processing, feature selection, and classification processes. It effectively classifies air data for efficient forecasting. The architecture diagram for the LR-MSV model is described in the following figure.

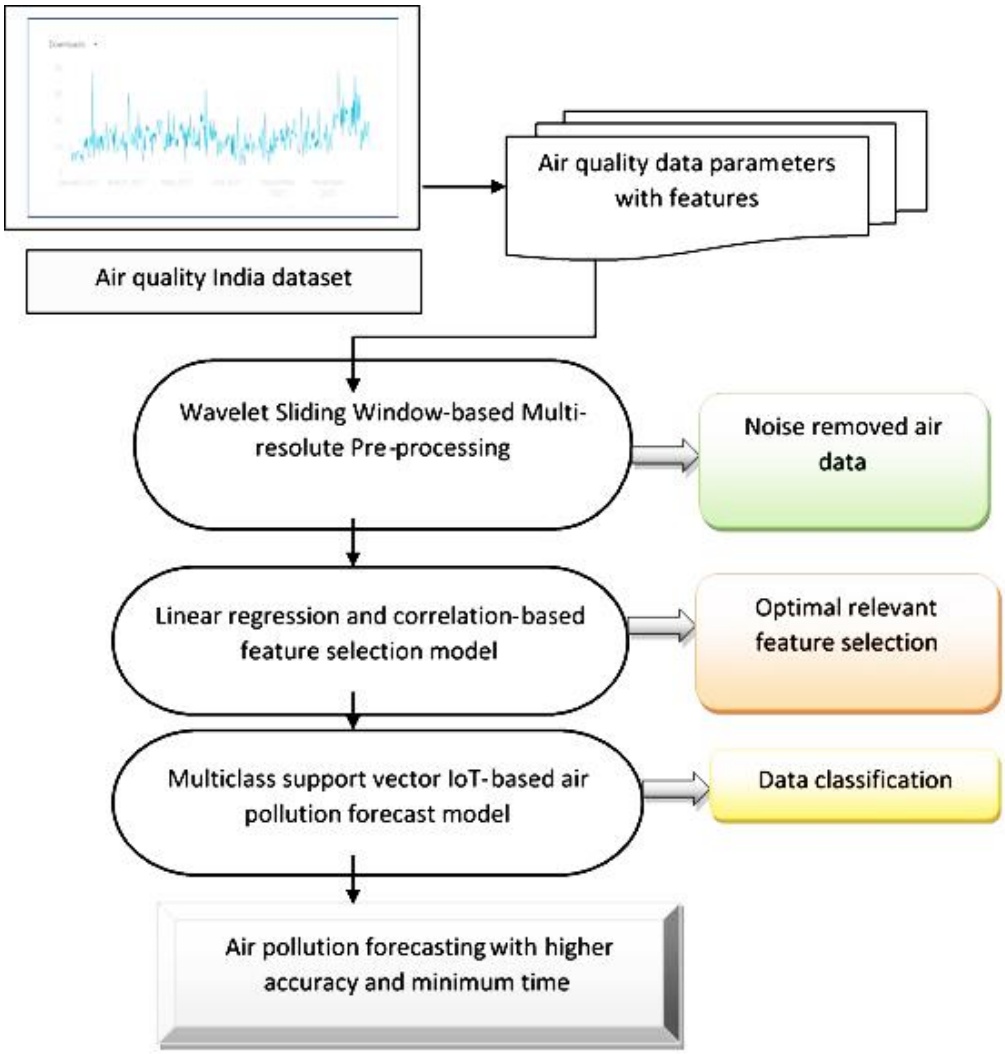

Figure 3.1: Architecture of LR-MSV model for forecasting

The suggested LR-MSV model's procedure for predicting air pollution is shown in Figure 3.1. Several air quality data with various attributes are obtained as input from the considered dataset, and for data to eliminate noise from input data, pre-processing is used. Using the obtained pre-processed data, the LRC model is used to carry out the feature selection. Based on estimated regression coefficient values, optimal relevant features are selected by removing irrelevant features. Here, the gradient function is presented to identify optimum relevant features of air data to perform the air pollution prediction process. Finally, data classification is provided to classify data into different classes based on selected features. Hence, the LR-MSV model efficiently monitors air quality data using multiclass support vectors. The obtained classification result predicts air pollution with improved accuracy and minimum time.

3.2.1 Data pre-processing using Sliding Window

The process of monitoring and controlling air quality data helps to estimate air pollution levels. The estimated pollution level accurately predicts the pollutant concentrations with slight noise. To acquire noise-free air quality data, the multi-resolute pre-processing model based on WSW is utilized. In general, time series units are predicted and wavelet decomposition is carried out via signal processing. By breaking down a non-smooth discrete time series of data into various evolutions according to frequency level, the pollution level is projected. Here, high-frequency feature components "H" and low-frequency coarse-grained components "L" are used to separate the air quality data. The number of layers in the wavelet decomposition determines the components of the high-frequency feature.

The suggested methodology uses a WSW-based Multi-resolute pre-processing model to filter out noise in data according to frequency and time. The pre-processing approach is first used on the unprocessed air data that has been obtained from the relevant sensors. The term "multi-resolute process" describes how often and for how long sliding windows are operated in order to reduce noise.

The sliding window is set up to estimate the window size value for each time instance 'T'. The estimated value of air quality data is specified as '$[T_{n-WS}, T_n]$'. Here, 'T_n' represents the records of values, and 'WS' denotes window size. Based on time and frequency, pre-processed data is obtained. The construction of the Multi-resolute WSW based pre-processing model is exposed in Figure 3.2.

The sliding window size value is presented considering both time series samples and features from the sample dataset. Based on window size, air quality samples are determined as more samples or lesser samples.

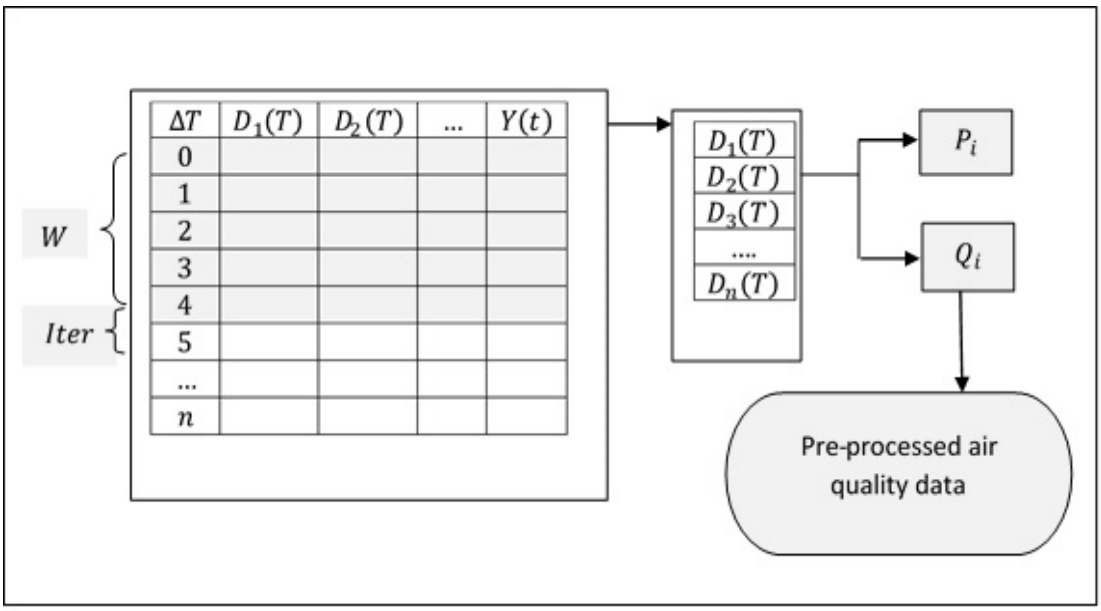

Figure 3.2: WSW-based multi-resolute pre-processing technique

When the sliding window size is small, air samples are higher. A bigger sliding window size, however, also corresponds to fewer examples and more features. As a result, the Equations provide the initial formulation of the sliding window model.

$$y(W + Iter) = f\left(T_0, \dots, T_W, D_{i,0}, \dots, D_{i,W}, T_{W+Iter}, Y_{W+Iter}\right) \quad \text{.....Eqn (3.1)}$$

$$y(W + Iter) = f\left(T_0, \dots, T_W, D_{i,0}, \dots, D_{i,W}, Y_0, \dots, Y_W, T_{W+Iter}, Y_{W+Iter}\right) \dots \text{ Eqn (3.2)}$$

The sliding window model is formulated in Equations (3.1) and (3.2) based on time and frequency series. Additionally, wavelet decomposition 'W' is utilized to determine time series air quality data. The record of air quality on each data is taken from corresponding sensors to produce sliding window. Based on window size, data is identified as high frequency and low frequency data. Here, higher frequency data is considered as feature components, and low frequency data is presented as coarse-grained components. The decomposed high frequency and low frequency data is mathematically formulated as given Equations.

$$P_{i+1} = L(P_i) \quad \text{..... Eqn (3.3)}$$

$$Q_i = H(Q_i) \quad \text{..... Eqn (3.4)}$$

From Equations (3.3) and (3.4), low frequency coarse grained component and high frequency feature component is presented and represented as 'P_i' and 'Q_i'. Here, 'L' specifies the low pass filter and 'H' representing the high pass filter. Each layer of the decomposed signal in a sliding window is created by bi-sectioning the pre-decomposed air quality signal data. The original signal length of the pre-processed air quality data is recovered using dual interpolation reconstructions. Thus, cloud server performs reconstruction of air quality data and given as Equations.

$$P_i = (L_2)^i P_i \quad \text{..... Eqn (3.5)}$$

$$Q_i = (L_2)^{i-1} H_2 Q_i \quad \text{..... Eqn (3.6)}$$

The dual interpolation reconstructions of data with frequency are expressed in Equations (3.5) and (3.6). The dual operators carried by cloud server is denoted as 'L_2' and 'H_2'. The interpolation reconstruction helps to attain noise removed data. The server determines data between noise and significant information at hourly and daily level. Thus, the air pollution is estimated with noise free data. The

process of data pre-processing model for noise removed air quality data is described in Algorithm 3.1.

From air quality dataset, a number of air data and features are gathered by respective sensors for initializing time instances. After initialization, sliding window is presented to determine data as high frequency and low frequency data. Additionally, wavelet decomposition is utilized on each layer of data for providing low frequency course grained component and high frequency feature component. Then, dual interpolation reconstruction is performed to retrieve original noise reduced data.

Input: Dataset 'DS', Cloud Server 'CS', IoT Devices or Sensors '$S = S_1, S_2, \dots, S_n$', Features '$F = F_1, F_2, \dots, F_n$', Air Quality data '$D = D_1, D_2, \dots, D_n$'

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

Step 1: **Initialize** time instance 'T'

Step 2: **Begin**

Step 3: **For** each Dataset 'DS' (Air Quality data '$D = D_1, D_2, \dots, D_n$') with Cloud Server 'CS' and IoT Devices or Sensors '$S = S_1, S_2, \dots, S_n$'

Step 4: Formulate sliding window using equations (3.1) and (3.2)

Step 5: **For** each Features 'F'

Step 6: Perform decomposition using equations (3.3) and (3.4)

Step 7: Model dual interpolation reconstructions to retrieve pre-processed air quality data using equations (3.5) and (3.6)

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

Step 9: **End for**

Step 10: **End for**

Step 11: **End**

Algorithm 3.1: Process of WSW pre-processing

3.2.2 Linear Regression for Feature Selection

After obtaining noise removed data, optimum relevant features are selected by applying LRC model. Here, correlation between features is estimated to obtain optimal features. The flow of LRC-based feature selection model is given in Figure 3.3. The feature selection process is performed next based on obtained pre-processed air quality data. By applying, linear regression correlation function, data features are extracted. The designed regression correlation process is a set of ML techniques that utilized to estimate the relationships between features.

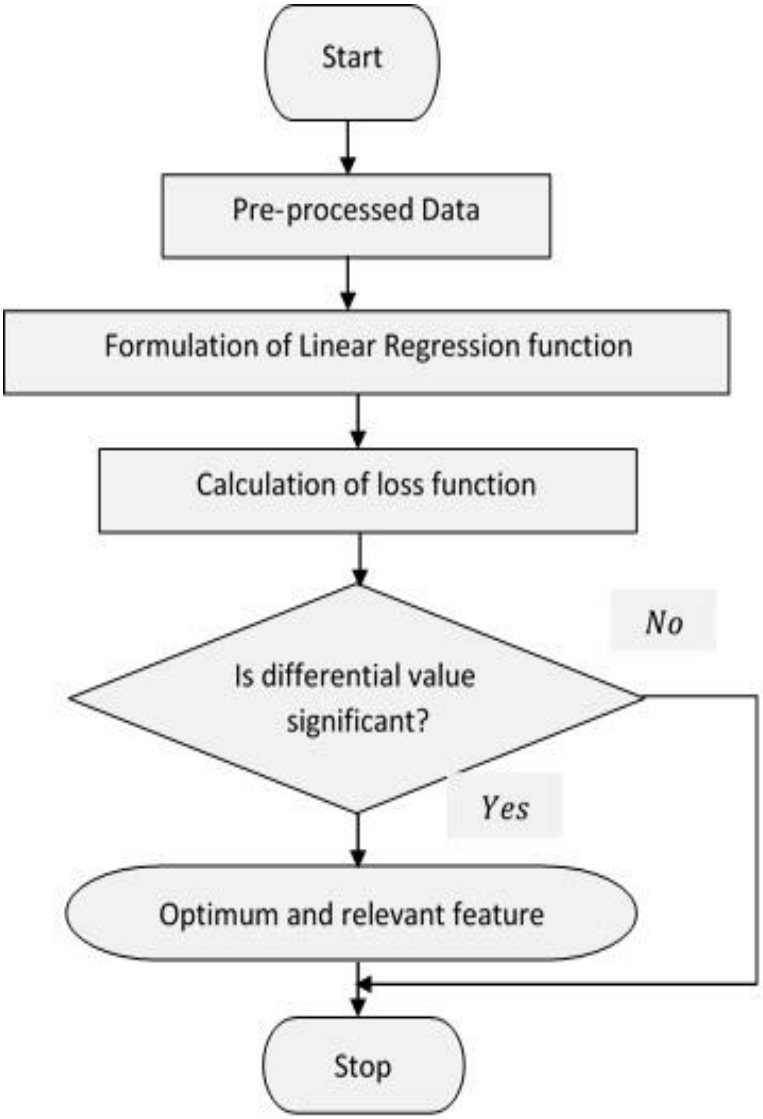

Figure 3.3: LRC-based optimal relevant feature selection

Here, the associations between two features are determined using regression function. After that, loss function is estimated to select optimal relevant features with minimum loss. Let us consider the independent data that denoted as '$PD = (PD_1, PD_2, \dots, PD_n)$'. For each pre-processed data, regression coefficients is estimated and specified as '$\beta = (\beta_1, \beta_2, \dots, \beta_n)$'. Then, the feature selection is formulated as given.

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_j + PD_j^i \qquad \dots\dots \text{Eqn (3.7)}$$

The feature selection prediction 'y_i' is obtained using equation (3.7). Here, 'β_0' specifies initially regression coefficient value of air data, 'β_j' denotes regression coefficients of number of pre-processed data and 'PD_j^i' denotes number of obtained independent data. Depends on linear regression 'y_i', optimal and appropriate features are selected with the use of pre-processed data. With obtained optimum and appropriate features, the sum of mean squared loss is minimized. The loss function is mathematically formulated as given.

$$\beta = argmin L(Dis, \beta) = argmin \sum_{i=1}^{n} (\beta \,.\, PD_i - y_i)^2 \qquad \dots\dots \text{Eqn (3.8)}$$

By using Equation (3.8), the sum of mean squared loss 'L' is estimated. Based on the distance 'Dis' among two pre-processed data 'PD' through regression coefficients, loss function is estimated for selecting relevant feature. After that, gradient function is utilized to attain optimal appropriate features. Consequently, the gradient function can be written as follows in the equation.

$$\frac{\partial L(Dis,\beta)}{\partial \beta} = \frac{\partial (Y^T Y - Y^T PD\beta - \beta^T PD^T PD\beta)}{\partial \beta} \qquad \dots\dots \text{Eqn (3.9)}$$

The optimal appropriate features with least loss are attained by using Equation (3.9). Here, '$\partial L(Dis, \beta)$' denotes the discriminating correlation between

features based on distance and '$\partial\beta$' specifies regression coefficient. The algorithmic process of Gradient LR-based feature selection is labelled in Algorithm 3.2.

At first, the noise removed data is considered in the dataset to select relevant features. For each input data, LR function is applied for feature selection from dataset. It measures correlation coefficient value between the features of air quality data. Based on calculated correlation coefficient value, optimal relevant features are determined. Then, mean squared loss is estimated, and gradient function is performed to obtain optimum relevant features of data. Thus, significant relevant features of air quality data are extracted from considered dataset.

Input: Dataset 'DS', Cloud Server 'CS', IoT Devices or Sensors '$S = S_1, S_2, \dots, S_n$', Features '$F = F_1, F_2, \dots, F_n$', Air Quality data '$D = D_1, D_2, \dots, D_n$'
Output: Optimal and relevant feature selection
Step 1: **Initialize** pre-processed air quality data 'PD', regression coefficients '$\beta = (\beta_1, \beta_2, \dots, \beta_n)$'
Step 2: Begin
Step 3: **For** each pre-processed air quality data 'PD' with Cloud Server 'CS' and IoT Devices or Sensors '$S = S_1, S_2, \dots, S_n$'
Step 4: Formulate linear regression function as given in equation (3.7)
Step 5: Evaluate sum of mean squared loss as given in equation (3.8)
Step 6: Obtain optimum relevant features using gradient function as given in equation (3.9)
Step 7: **Return** relevant features (RF)
Step 8: **End for**
Step 9: **End**

Algorithm 3.2: Procedure of Gradient Linear Regression-based Feature Selection

3.2.3 Data Classification using MSV Technique

Finally, MSV based data classification process is performed to classify data for predicting the air quality. An ensemble classification technique is presented with the use of selected significant features. Here, the selected appropriate features (i.e., PM2.5, PM10, SO2, NOx, NH3, CO and O3) are considered for air quality data classification. The Air Quality Index (AQI) is utilized to classify data by monitoring and protective measures for controlling air data. Hence, the proposed model utilizes MSV based Forecast model for better classification. The construction of data sorting is shown in Figure 3.4.

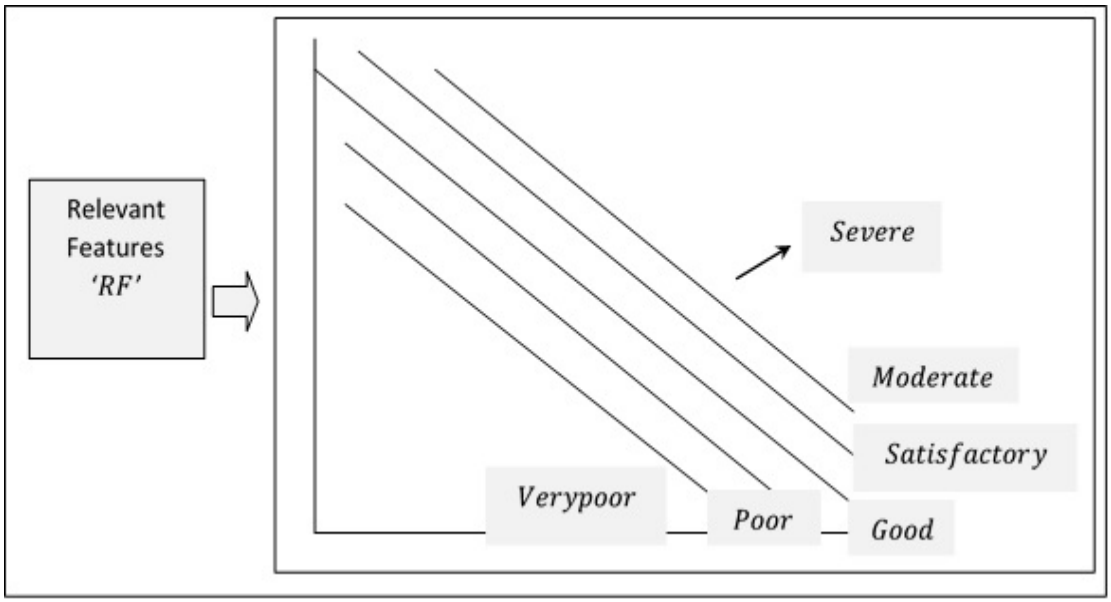

Figure 3.4: Air pollution forecasting using MSV

By considering selected relevant features, classification of air quality data is achieved. At first, air quality dataset 'DS' is considered forecasting. From considered dataset, relevant features are selected and represented as '$RF = \{RF_1, RF_2, \dots, RF_n\}$'. The number of relevant features '$RF_i \in R^n$'' are belonged to different number of classes 'C' (i.e., '$C = 6$'). Based on features and classes, the result is provided as '(RF_i, Y_i)'. Here, 'RF_i' specifies number of selected relevant features and 'RF_i' denotes the range of air data classification. The range is

described as '$Y_i \in \{0, 0-0.25, 0.25-0.50, 0.50-0.75, 0.75-1, >1\}$' that represents six dissimilar classes namely good, satisfactory, moderate, poor, very poor and severe respectively. The classified classes are specified as class labels. Thus, the linear classifier of data through recommendation confinements is formulated as below.

$$W^T RF + B \qquad \ldots\ldots \text{Eqn (3.10)}$$

From Equation (3.10), recommendation confinements are illustrated. Here, 'W^T' denotes the weight vector, 'RF' specifies selected relevant features and 'B' symbolizes the bias value. Among various classifications, SVM classification approach is presented to perform efficient data classification with the determination of decision boundary. When the distance from AQI value is nearer to the selected relevant features, the margin of data classifier is determined. The estimation of AQI for each sampled air data is determined by using following formulation as given below.

$$AQI = Avg(PM2.5, PM10, SO2, NOx, NH3) + Max(CO, O3) \qquad \ldots\ldots \text{Eqn (3.11)}$$

The 'AQI' is estimated by using Equation (3.11). Depending on the average values of PM2.5, PM10, SO2, NOx, NH3 and the extreme values of CO and O3, index value is calculated. These points are denoted as the support vectors in the model. The resulting of AQI_Bucket is measured as the quality of air for forecasting the pollution level. By measuring and controlling air quality values, air pollution is determined with reduced time.

The process of ensemble data classification using MSV based prediction is described in the Algorithm 3.3. The selected relevant features from dataset are considered as input to classify data into different classes. By using selected features, linear classifier is formulated at first. A linear classifier is applied to classify air data into various classes. After that, air quality index value is estimated for obtaining efficient data classification process. Thus, the classified air data helps

to forecast pollution with minimum error. As a result, proposed LR-MSV model is performed for forecasting with enhanced accuracy and least time.

3.3 Simulation Settings

The proposed LR-MSV model is simulated by using Java JDK 1.8 version interfaces and CloudSim simulator. The standard tools utilized in java language are appletviewer, extcheck, jar, java, javac, javadoc, javah, javap, jdb and jdeps. For conducting simulation work, Air Quality India dataset is considered. The dataset comprises of 2,18,640 data and 16 different features. The AQI value is calculated in both hourly and daily basis for several

Input: Dataset 'DS', Cloud Server 'CS', IoT Devices or Sensors '$S = S_1, S_2, \dots, S_n$', Features '$F = F_1, F_2, \dots, F_n$', Air Quality data '$D = D_1, D_2, \dots, D_n$'
Output: Robust classification
Step 1: **Initialize** relevant features (RF)
Step 2: **Begin**
Step 3: **For** each relevant feature 'RF' with Cloud Server 'CS' and IoT Devices or Sensors '$S = S_1, S_2, \dots, S_n$'
Step 4: Formulate recommendation confinements of linear classifier as in equation (3.10)
Step 5: Measure Air Quality Index as in equation (3.11)
Step 6: **If** '$W^T RF_i(AQI) + b = 0$'
Step 7: **Then** AQI_Bucket '$AQI_B \rightarrow Good$'
Step 8: **End if**
Step 9: **If** '$W^T RF_i + b(AQI) > 0 \; and W^T RF_i + b(AQI) < 0.50$'
Step 10: **Then** AQI_Bucket '$AQI_B \rightarrow Satisfactory$'
Step 11: **End if**
Step 12: **If** '$W^T RF_i + b(AQI) > 0.50 \; and W^T RF_i + b(AQI) < 1$'
Step 13: **Then** AQI_Bucket '$AQI_B \rightarrow Moderate$'
Step 14: **End if**
Step 15: **If** '$W^T RF_i + b(AQI) > 0 \; and W^T RF_i + b(AQI) < -0.50$'
Step 16: **Then** AQI_Bucket '$AQI_B \rightarrow Poor$'
Step 17: **End if**
Step 18: **If** '$W^T RF_i + b(AQI) > -0.50 \; and W^T RF_i + b(AQI) < -1$'
Step 19: **Then** AQI_Bucket '$AQI_B \rightarrow VeryPoor$'
Step 20: **End if**
Step 21: **If** '$W^T RF_i + b(AQI) > 1$'
Step 22: **Then** AQI_Bucket '$AQI_B \rightarrow Severe$'
Step 23: **End if**
Step 24: **End for**
Step 25: **End**

Algorithm 3.3: Process of MSV based Air Pollution Forecast

places across India. The measured AQI and air quality data is presented in considered Air Quality India dataset. The features on dataset are described in Table 3.1.

During the experimental consideration, the different quantity of air quality data ranges from 20,000 to 2,00,000 is considered as input from dataset. With the use above dataset description, result analysis is carried out using proposed LR-MSV model. The experimental analysis is performed by using the following parameters:

- Air pollution forecasting time,
- Air pollution forecasting accuracy, and
- Error rate

3.4 Performance Analysis

The result analysis of proposed LR-MSV model is conducted by comparing with different exiting methods. The compared existing methods are IMD-VAE designed by Abdelkader Dairi et al. (2021) and bidirectional RNN which is introduced by D. Saravanan and K. Santhosh Kumar (2021) respectively.

3.4.1 Performance Analysis of Forecasting Time

The term forecasting time refers to the amount of time spent monitoring and predicting air pollution. It is displayed as the product of one data point's predicting time and the amount of data from air quality samples. The time is expressed mathematically as follows and is measured in milliseconds (ms).

$$APF_{Time} = D_i * Time(forecastsingledata) \quad \text{... Eqn (3.12)}$$

By using Equation (3.12), forecasting time 'APF_{Time}' is estimated depends on 'D_i' amount of sample data. When there is a lower time, the proposed model is said to be more efficient.

Table 3.1: Dataset description

S. No.	Features	Description
1	City	Name of City
2	Date	Date of occurrence
3	PM 2.5	Particulate Matter 2.5
4	PM 10	Particulate Matter 10
5	NO	Nitric Oxide
6	NO2	Nitric dioxide
7	NOx	Any nitric x-oxide
8	NH3	Ammonia
9	CO	Carbon monoxide
10	SO2	Sulphur dioxide
11	O3	Ozone
12	C6H6	Benzene
13	C7H8	Toluene
14	C8H10	Xylene
15	AQI	Air quality indices
16	AQI_Bucket	Air quality indices bucket

Sample calculation:

Existing IMD-VAE: Time taken to forecast single air quality data is 0.1215 ms and total quantity of input air quality sample data is 20,000. Then, the air pollution forecasting time is estimated as $APF_{Time} = 20{,}000 * 0.1215\, ms = 2430\, ms$.

Existing Bidirectional RNN: Time taken to forecast single air quality data is 0.139 ms and total amount of input air quality sample data is 20,000.

Then, the air pollution forecasting time is estimated as $APF_{Time} = 20{,}000 * 0.139 = 2780 ms$.

Proposed LR-MSV Model: Time taken to forecast single air quality data is 0.095 ms and total amount of input air quality sample data is 20,000. Then, the forecasting time is estimated as $APF_{Time} = 20{,}000 * 0.095 = 1900ms$.

Table 3.2: Forecasting time of existing methods vs LR-MSV model

Number of air quality data	Air pollution forecasting time (ms)		
	Existing IMD-VAE	Existing bidirectional RNN	Proposed LR-MSV model
20,000	2430	2780	1900
40,000	2360	2750	1840
60,000	2300	2700	1800
80,000	2200	2660	1780
1,00,000	2100	2550	1750
1,20,000	2020	2520	1720
1,40,000	1980	2480	1650
1,60,000	1920	2440	1620
1,80,000	1850	2380	1590
2,00,000	1700	2320	1560

The experimental values of forecasting time are tabulated in the Table 3.2 based on quantities of data from dataset. Measurements of air data in the range of 20,000 to 2,00,000 data are made in order to conduct experimental activities. Table 3.2 provides the comparison result of proposed LR-MSV model with existing IMD-VAE designed by Abdelkader Dairi et al. (2021) and bidirectional RNN introduced by D. Saravanan and K. Santhosh Kumar (2021) respectively. From the experimental analysis, proposed model is resulted with minimum air pollution forecasting time by conducting investigational works. Thus, the time taken for

forecasting air pollution data using proposed LR-MSV model is lower than other methods. Based on the above table value, graph is drawn for analyzing the performance as shown in Figure 3.5.

The suggested LR-MSV model outperforms the current IMDA-VAE and Bidirectional RNN in terms of minimum time, as seen in the figure. Besides, while increasing the amount of input data, the time taken for forecasting air data gets varied. But comparatively, proposed model attains minimum time for air pollution forecasting.

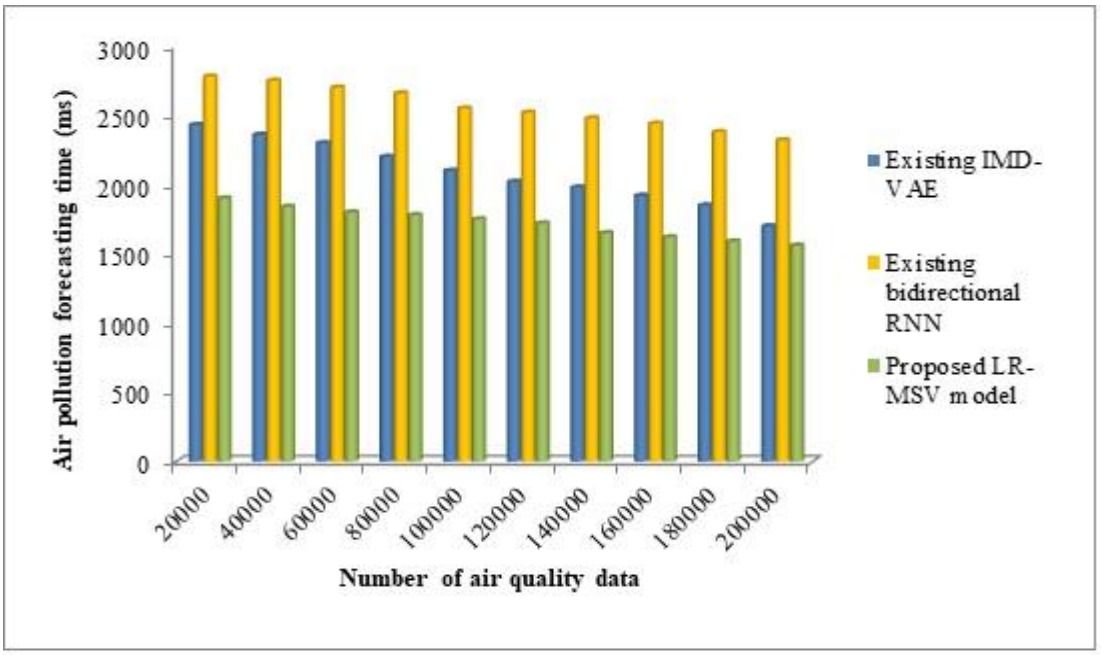

Figure 3.5: Air pollution forecasting time of LR-MSV model

The proposed model initially preprocesses the dataset by using a Time and Frequency-based Sliding Window method to eliminate noisy data. The process of selecting the most pertinent and ideal attributes from the dataset is then shown. By using gradient function, air data is classified into different classes for accurate forecasting. Thus, ensemble classification technique is performed using relevant features to group data into various classes for better forecasting. Therefore, proposed LR-MSV model reduces the forecasting time by 17% and 33% when

compared to existing methods such as IMD-VAE designed by Abdelkader Dairi et al. (2021) and bidirectional RNN introduced by D. Saravanan and K. Santhosh Kumar (2021) respectively.

3.4.2 Evaluation of Forecasting Accuracy

The forecasting accuracy is defined as the ratio of multiple correctly projected data to the entire number of air quality sample data considered from the dataset. The percentage (%) is used as the measurement unit. Equation (3.13) is the mathematical statement for calculating the data forecast.

$$APF_{Accuracy} = \frac{D_{\text{accurately forecasted}}}{Numberof\ air\ quality\ data} * 100 \qquad \text{….. Eqn (3.13)}$$

By using Equation (3.13), air pollution forecasting accuracy '$APF_{Accuracy}$' is determined. Here, '$D_{\text{accurately forecasted}}$' denotes number air data that are accurately forecasted for air pollution. A high level of forecast accuracy indicates that the suggested model is performing more efficiently.

Sample calculation:

Existing IMD-VAE: Amount of air quality data that are accurately forecasted is 15,486 and total quantity of input air quality sample data is 20,000. Then, air pollution forecasting accuracy is determined as$APF_{Accuracy} = \frac{15,486}{20,000} * 100 = 77.43\%$.

Existing Bidirectional RNN: Amount of air quality data that are accurately predicted is 14,326 and total amount of input sample data is 20,000. Then, the accuracy is determined as$APF_{Accuracy} = \frac{14,326}{20,000} * 100 = 71.63\%$.

Proposed LR-MSV Model: Number of air quality data that are correctly predicted is 16,986 and total amount of input air quality sample data is 20,000.

Then, the air pollution forecasting accuracy is determined as $APF_{Accuracy} = \frac{16,986}{20,000} * 100 = 84.93\%$.

The experimental value of forecasting accuracy for the suggested and current approaches is shown in Table 3.3. The current IMD-VAE, created by Abdelkader Dairi et al. (2021), and the bidirectional RNN, presented by D. Saravanan and K. Santhosh Kumar (2021), are compared to the suggested LR-MSV model in this instance. According to the result analysis, compared to other current methods, the suggested LR-MSV model effectively increases the accuracy. As demonstrated in Figure 3.6, the numbers shown in the table serve as the basis for plotting the graph.

Table 3.3: Forecasting accuracy of existing methods vs LR-MSV model

Number of air quality data	Air pollution forecasting accuracy (%)		
	Existing IMD-VAE	Existing bidirectional RNN	Proposed LR-MSV model
20,000	77.43	71.63	84.93
40,000	77.82	71.85	85.14
60,000	78.62	72.33	85.23
80,000	79.54	72.64	86.65
1,00,000	80.25	73.52	87.25
1,20,000	80.61	74.84	88.14
1,40.000	80.94	75.25	89.62
1,60,000	81.25	76.52	90.25
1,80,000	82.62	77.25	90.64
2,00,000	83.64	78.62	91.72

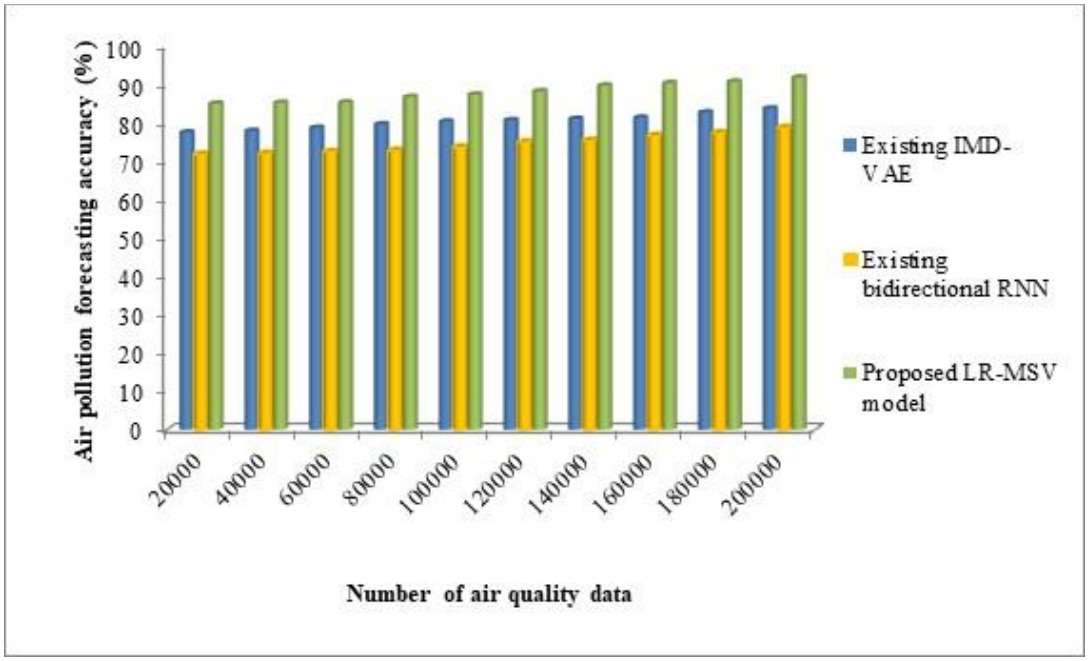

Figure 3.6: Forecasting accuracy of LR-MSV model

The evaluation study of forecasting accuracy for both suggested and current methodologies is displayed in Figure 3.6. The amount of data used as input for the experiment varies from 20,000 to 2,00,000. The simulation is performed using appropriately forecasted data in order to predict air pollution. The suggested LR-MSV model is contrasted with the current IMD-VAE and Bidirectional RNN for simulation purposes. From experiment evaluation results, the accuracy to forecast data is significantly enhanced using proposed model.

With the application of Gradient LR-based Feature Selection algorithm, relevant air quality data is selected for forecasting process. Based on selected features, air quality data is accurately forecasted with minimum sum of mean squared loss. Followed by gradient regression analysis function is performed to achieve better air data forecasting with higher accuracy. Hence, the accuracy is improved using proposed LR-MSV model by 10% and 18% when compared to existing IMD-VAE designed by Abdelkader Dairi et al. (2021) and bidirectional RNN introduced by D. Saravanan and K. Santhosh Kumar (2021).

3.4.3 Performance Analysis of Error Rate

The error rate is defined as the proportion of inaccurately anticipated air quality data to the total amount of input air quality data. The unit of measurement is the percentage (%). The following is the mathematical formula for calculating the error rate.

$$ER = \frac{D_{forecastedwrongly}}{D_i} * 100 \qquad \text{…. .Eqn (3.14)}$$

From Equation (3.14), Error Rate 'ER' is calculated. Based on 'D_i' input sample air quality data and '$D_{forecastedwrongly}$' wrongly forecasted air quality data, error rate is calculated. If the error rate during air pollution monitoring and controlling is lower, then the performance of the projected model is said to be more effective.

Sample calculation:

Existing IMD-VAE: Quantity of wrongly forecasted air quality data is 4514 and total amount of input air quality sample data is 20,000. Then, the error rate is measured as $ER = \frac{4514}{20{,}000} * 100 = 22.57\%$.

Existing bidirectional RNN: Quantity of wrongly forecasted air quality data is 5674 and total amount of input air quality sample data is 20,000. Then, the error rate is measured as $ER = \frac{5674}{20{,}000} * 100 = 28.37\%$.

Proposed LR-MSV model: Number of wrongly forecasted air quality data is 3014 and total amount of input sample data is 20,000. Then, error rate is measured as $ER = \frac{3014}{20{,}000} * 100 = 15.07\%$.

Table 3.4: Error rate of existing methods vs LR-MSV model

Number of air quality data	Error rate (%)		
	Existing IMD-VAE	Existing bidirectional RNN	Proposed LR-MSV model
20,000	22.57	28.37	15.07
40,000	22.18	28.15	14.86
60,000	21.38	27.67	14.77
80,000	20.46	27.36	13.35
1,00,000	19.75	26.48	12.75
1,20,000	19.39	25.16	11.86
1,40,000	19.06	24.75	10.38
1,60,000	18.75	23.48	9.75
1,80,000	17.38	22.75	9.36
2,00,000	16.36	21.38	8.28

The experimental values for error rate during air data monitoring and controlling are tabulated in Table 3.4 for both proposed and existing methods. The proposed LR-MSV model is compared with existing IMD-VAE designed by Abdelkader Dairi et al. (2021) and bidirectional RNN introduced by D. Saravanan and K. Santhosh Kumar (2021) respectively. A number of air quality data points from datasets ranging from 20,000 to 2,00,000 are taken into consideration for the experimental work. The suggested LR-MSV model, depends on the values in the above table, lowers inaccurate air data predictions in order to improve air pollution

forecast performance. With respect to the tabulated values, graph is plotted as given figure for presenting performance analysis result.

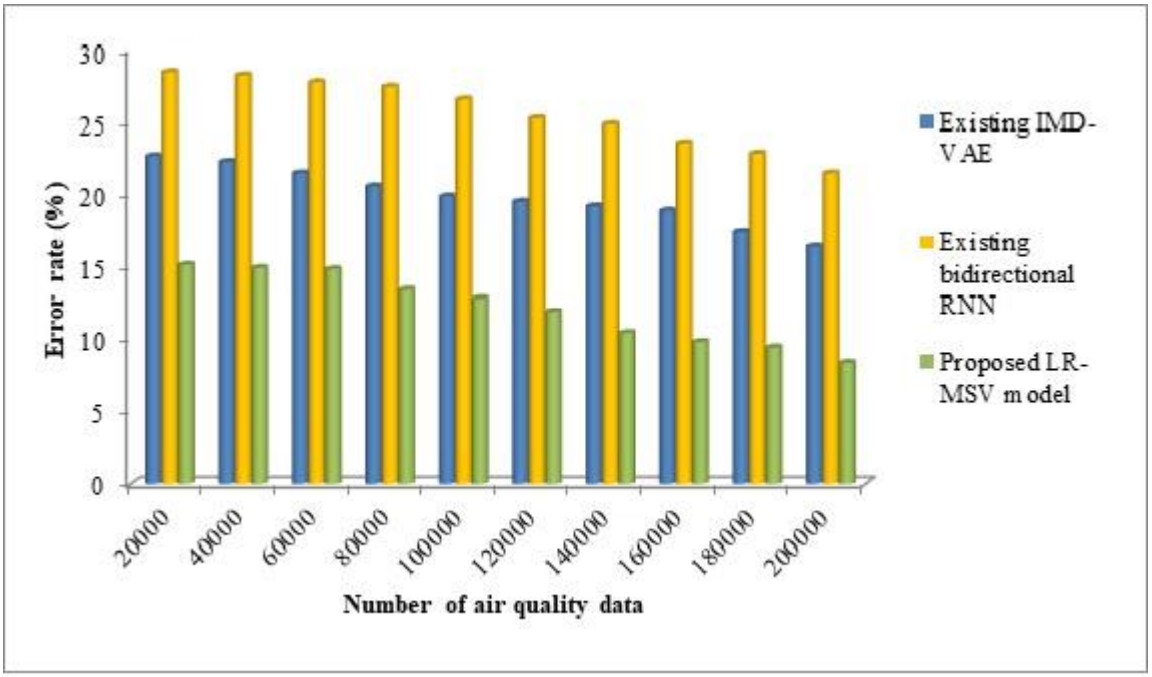

Figure 3.7: Error rate of LR-MSV model

Figure 3.7 illustrates the error rate description for the suggested model and the current approach. With the help of the suggested model and current techniques, the error rate is determined with regard to a range of air quality data. The suggested LR-MSV model offers a lower error rate for precise forecasting and control than the other approaches, according to the findings of the experimental investigation. The two approaches that are being compared are IMD-VAE and Bidirectional RNN. When the quantity of input data rises, the forecasting accuracy rate for air data fluctuates. The forecast model estimates AQI values for classifying data into different classes. Therefore, proposed LR-MSV model minimizes error rate by 40 % when compared to existing IMD-VAE designed by Abdelkader Dairi et al. (2021) and by 53% when compared to bidirectional RNN introduced by D. Saravanan and K. Santhosh Kumar (2021) respectively.

3.5 Summary

An efficient LR-MSV model is proposed for accurate forecasting and control measures. The main objective of air pollution predicting is obtained by classifying data into different classes. The suggested model incorporates several procedures, including feature selection, classification, and pre-processing. Numerous air quality data points are gathered from the examined air quality dataset and are taken into consideration as input. For each input data, data pre-processing is performed using WSW-based Multi-resolute pre-processing model. During pre-processing, noise data is removed and attain pre-processed data. After that, feature selection is carried by applying LRC-based Feature Selection on pre-processed data. Based on linear regression correlation coefficient value, significant relevant features are extracted from dataset. Here, irrelevant air quality data features are removed, and it helps to minimize time taken to forecast air data. Lastly, the MSV IoT-based Air Pollution Forecast model is run using a few chosen characteristics. The ensemble classification technique classifies data into various classes with minimum error, based on estimation of air quality index value. Hence, it effectively classifies the data and provides efficient pollution forecasting. Consequently, the performance results demonstrate that compared to state-of-the-art techniques, the suggested LR-MSV model enhances air pollution prediction accuracy with the least amount of time and error rate. While forecasting air pollution, estimation of space utilization is not considered. Thus, the estimation of memory consumption to forecast air data is attained by proposing a next model.