CHAPTER 5

DISCRETIZED REGRESSION AND LEAST SQUARE SUPPORT VECTOR FOR AIR POLLUTION FORECASTING

5.1 Introduction

The accurate prediction of air quality in natural habitat and airspace is achieved based on available environmental confronts. Air pollution happens in the environment when particle matter, chemicals, or biological materials accumulate. Humans and other living things are affected by the air pollution that is produced. The accurate performance of pollution monitoring is carried through prevailing data-driven methods among spatio temporal characteristics of air pollution. The utilization of learning process provides early prediction on air pollution data with higher quality. There are several ensemble classification techniques presented with various classifier process. By classifying air quality data, pollution present in the environment at specific location and time is effectively predicted. For efficient air pollution forecasting, ML techniques are considered. In the previous section, BTBSR-QWEBC model was explained for predicting air data with higher accuracy and minimal time. Initially, data pre-processing carries wavelet transform to remove noise data. After that, significant features from dataset are determined using Otsuka Indexive Broken-stick regression process. Here, similarity coefficient value is measured to select relevant features of data. Based on relevant features, emphasis boost classification technique is used to categorize data into diverse classes. Thus, classified data with minimized error is obtained for better air pollution prediction.

Similarly, deep learning and classification techniques were developed to obtain higher air pollution monitoring and forecasting. Though, it fails to identify different air quality data in environmental area with reduced time. As air pollution is highly harmful and requires complex relationships between spatial and temporal

features, it is challenging to accurately monitor air data. An existing classification algorithm was presented to perform classification on air data. It classifies data to identify the polluted air with increased accuracy, but error occurrence is higher. However, the classification of data for pollution monitoring was not enhanced. Due to the air pollution, living beings are affected by respiratory related issues, lung diseases, cardiovascular distress and mental related affairs affecting prevailing health situation. Therefore, essential classification techniques are required to identify polluted air data by classifying data. But early identification of data with efficient monitoring and controlling is significant.

The DR-LSSV method is introduced for achieving enhanced result of prediction. The classification of data to determine AQI for air pollution prediction is the key objective of the suggested DR-LSSV technique. The volume of air quality data with features is first processed at the input layer from the dataset in order to conduct forecasting performance. Discretized Hartley Transformation-based pre-processing at the first hidden layer is used to handle the data after considering the input air data. Actual inputs are converted into actual outputs during pre-processing by removing noise from the data. The feature selection procedure is then passed out in the second hidden layer using a CMLLR-based feature selection model. In this case, Logistic Regression (LoR) function is used to identify and choose pertinent features for the prediction procedure.

The third hidden layer then uses CCLSSV as a classification model. Data is categorized for precise air pollution predictions based on specific important features that are chosen. To categorize data, the Kendall's Rank Correlation Coefficient is estimated. The categorized data contributes to more accurate and timely air pollution forecasting. The output layer displays the end-product of the forecasting process. As a result, the suggested DR-LSSV technique performs experimental research on variables including forecasting accuracy, error rate and time. Experiments are carried out based on a number of air quality data from the

dataset. The experimental result demonstrates that, compared to existing methods, the suggested strategy enhances predicting accuracy in less time.

5.2 Architecture of Proposed DR-LSSV Model

For more accurate environmental air pollution forecasting, the DR-LSSV model is presented. To categorize the data for pollution forecasting, a substantial selection of features and classification method is used. Air pollution is predicted early on by the relevant features that have been chosen with maximum accuracy and minimum time. The neural learning is utilized in proposed method with various layers. The considered neural learning is a class of deep neural networks. By performing data classification through various layers, false detection on pollution forecasting gets minimized. Thus, it resulted with enhanced accuracy on forecasting of air pollution with minimum time.

At first, multiple numbers of features is considered from dataset to forecast air pollution. With collected input data, discretized Hartley transformation is applied for data pre-processing. The pre-processing stage detects and removes noisy data from the dataset. The CMLLR technique is used to carry out the feature selection procedure with the pre-processed data that has been collected. The process of selecting important and pertinent features is aided by the likelihood regression function estimate. Lastly, CLSSV technique is performed to carry data classification process with relevant features. Here, coefficient of concordance determines correlation coefficient value. Based on estimated coefficient value, input data is classified with higher accuracy. With data classified result, air pollution is forecasted with minimum time and error rate.

The architecture of the suggested DR-LSSV model for air pollution data prediction is shown in Figure 5.1. The suggested model's primary goal is to increase pollution forecasting accuracy. Here, dataset along with IoT devices is considered with numerous numbers of air quality data with both relevant and

irrelevant features. After that, data classification is carried out for accurate pollution forecasting based on selected significant relevant features. For more accurate pollution forecasting, various steps such pre-processing of data, major relevant feature selection, and classification are carried out in the suggested method.

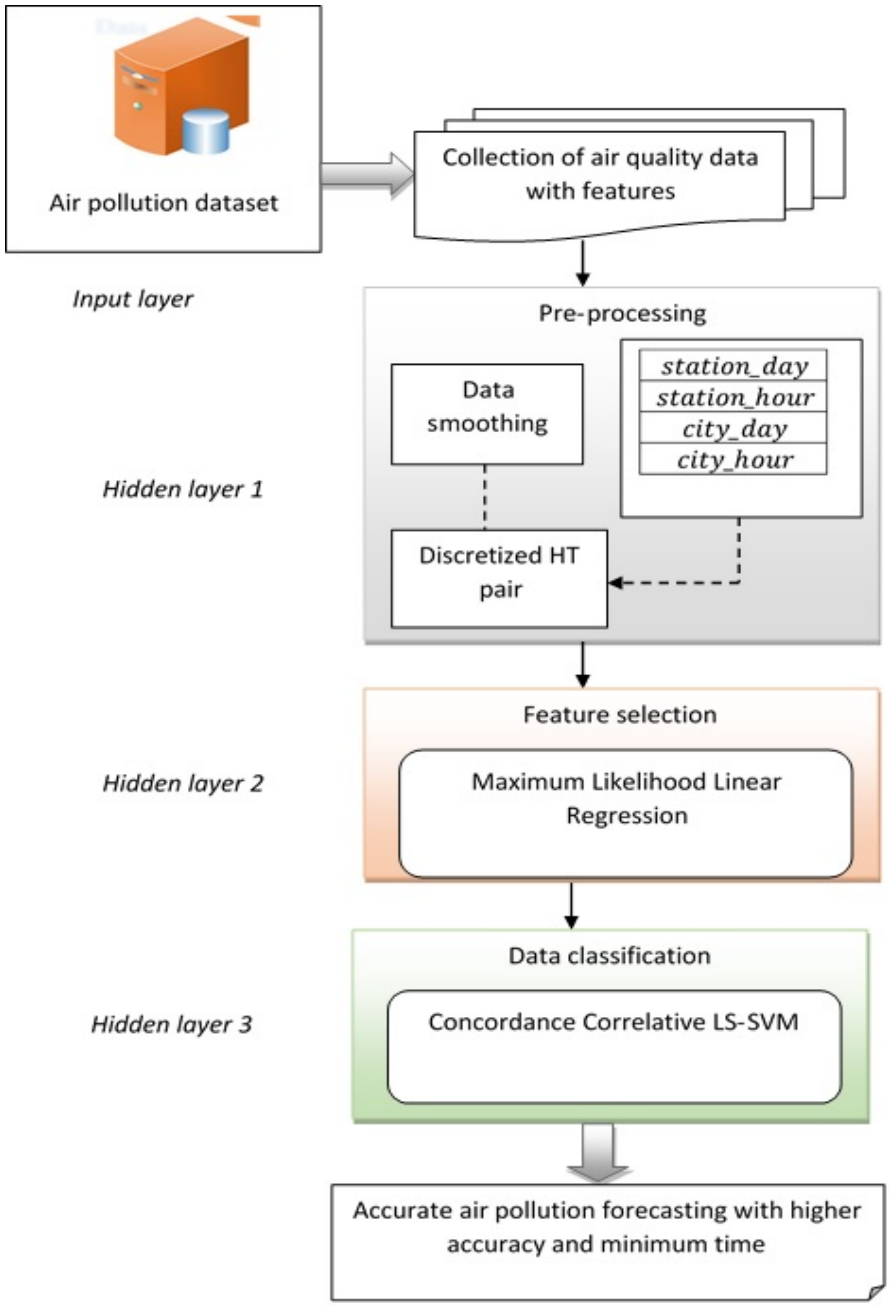

Figure 5.1: Architecture of DR-LSSV Model

Initially, air quality data from dataset is taken as input at input layer. After that, input data is transferred into hidden layer. In the first hidden layer, data pre-processing is applied to remove noisy data from dataset. With pre-processed data, feature selection is performed at second hidden layer to determine relevant features of data. Lastly, data classification technique is presented by estimating concordance correlated coefficient value. Based on measured coefficient value, air data is effectively classified for forecasting air pollution with enhanced accuracy. The air dataset is gathered from Air Quality Data in India to enable precise air pollution predictions. It collects air data at the hourly levels and daily levels from multiple Indian sites for the time frame of 2015 to 2020. The dataset's air data is taken into consideration based on the AQI value. Features such as PM2.5, PM10, SO2, NOx, NH3, CO, and O3 are considered when calculating the AQI result. That pertains to PM2.5, PM10, SO2, NOx, and NH3. The qualities having at least 16 values are therefore selected based on their average value during the previous 24 hours.

Over the past eight hours, the remaining CO and O3 function has been used to its fullest extent. Furthermore, every measure was converted into a Sub-Index using pre-established groups. Due to lack of data point measurement, some measures are not able access. The AQI value of data is estimated for everyday level at the hours of day by various city levels. The averaged value of AQI over stations of city is determined for selecting data features. The predefined buckets of AQI values are as given below. Figure 5.1 illustrates the data description based on calculation of AQI values. With the selection of air data, air pollution forecasting is accurate with minimum time and error rate.

5.2.1 Transformation based pre-processing Technique

The DR-LSSV model initially applies Discretized Hartley Transformation (DHT) process to perform data pre-processing. The process of converting real input air quality data to real outputs is described as data pre-processing.

Table 5.1: Pre-defined AQI values

S. No.	AQI values	Description
1	Good (0 – 50)	Minimal impact
2	Satisfactory (51 – 100)	Minor breathing discomfort to sensitive people
3	Moderate (101 – 200)	Breathing discomfort to the people with lung, heart disease, children, and other adults
4	Poor (201 – 300)	Breathing discomfort to people on prolonged exposure
5	Very poor (301 – 400)	Respiratory illness to the people on prolonged exposure
6	Sever (>401)	Respiratory effects even on healthy people

The presence of raw data in dataset provides various data formats, error, and incomplete data. The error occurrences are resolved by performing data pre-processing on input data from dataset. In environmental data, pre-processing is the most significant process in many machine learning techniques to provide accurate air pollution forecasting. Here, DHT is utilized to eliminate noise presented in the air data. With transformation of real inputs to real outputs, noisy data is eliminated for better pollution prediction.

Figure 5.2 demonstrates the development of data pre-processing by applying Discretized Hartley Transformation. Here, the noisy air data is identified through pre-processing at the first hidden layer. pre-processing is used to identify and resolve noise in the input air quality data. Real trends and patterns are retrieved using the station details, corresponding ID, hourly basis and daily basis in order to eliminate noise in the data. Therefore, noise data is eliminated when actual inputs are converted to real outputs using a transformation function.

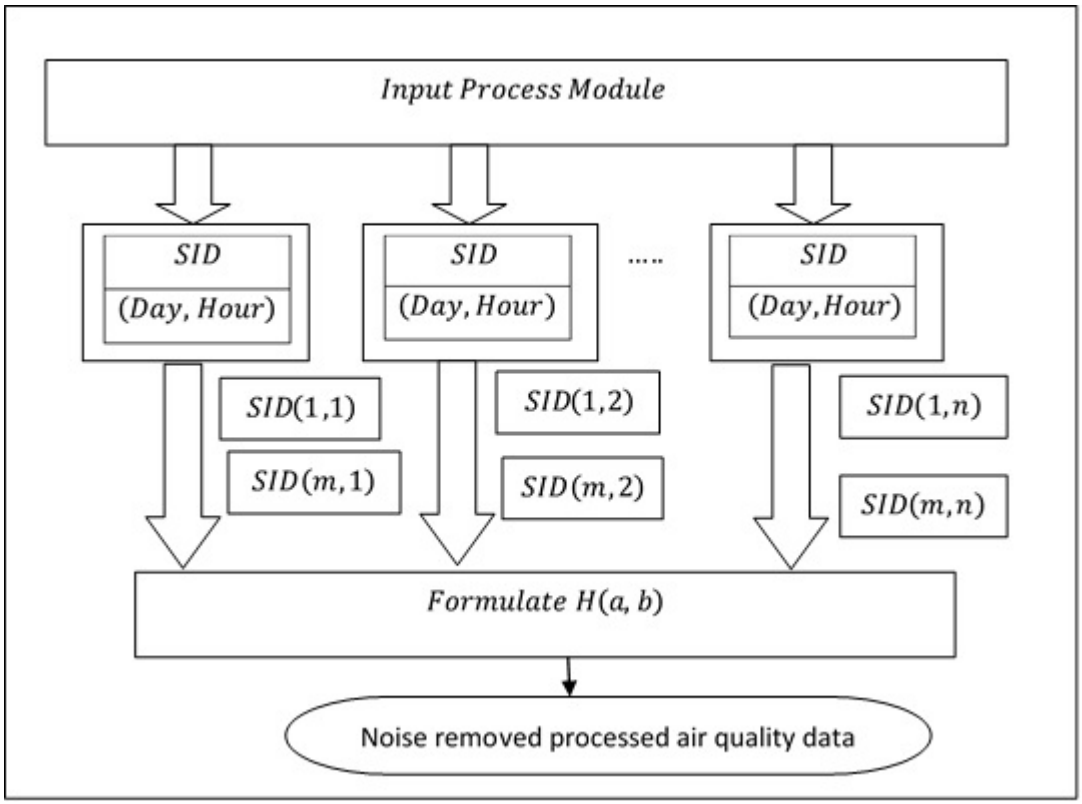

Figure 5.2: Structure of Discretized Hartley Transformation-based pre-processing

Let us consider the data from dataset denoted as '$D = D_1, D_2, \dots, D_n$' with the number of sensors '$S = S_1, S_2, \dots, S_n$'. Next, the data on air quality that has been taken into consideration is kept independently for each station in the method of a vector matrix based on the day and hour. Thus, the data in station ID with day and hour is formulated as below.

$$SID[Day] = Day[D_1, D_2, \dots, D_n] \quad \text{..... Eqn (5.1)}$$

$$SID[Hour] = Hour[D_1, D_2, \dots, D_n] \quad \text{..... Eqn (5.2)}$$

From the Equations (5.1) and (5.2), air data with day and hour is expressed. By using above formulation, the formula for a two-dimensional DHT pair is as follows,

$$H(a,b) = \sum_{Day=1}^{m} \sum_{Hour=1}^{n} SID(Day, Hour) cas\left[2\pi\left(\frac{aDay}{P} + \frac{bHour}{Q}\right)\right] \quad \text{.... Eqn (5.3)}$$

$$SID(Day, Hour) = \sum_{Day=1}^{m} \sum_{Hour=1}^{n} H(a,b) cas\left[2\pi\left(\frac{aDay}{P} + \frac{bHour}{Q}\right)\right] \quad \text{..... Eqn (5.4)}$$

Input: Dataset 'DS', 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: Processed air quality data 'PD'

Step 1: **Initialize** rows and columns of the air quality data 'P' and 'Q'

Step 2: **Begin**

Step 3: **For** each Dataset 'DS' with Sensors 'S' and air quality data '$D = D_1, D_2, \dots, D_n$'

Step 4: Obtain vector matrix separately for corresponding station based on day 'Day' and hour '$Hour$' as in equations (5.1) and (5.2)

Step 5: Formulate two-dimensional HT pair as in equations (5.3) and (5.4)

Step 6: Obtain discretized HT pair as in equation (5) to transform real inputs to real outputs.

Step 7: **Return** processed air quality data 'PD'

Step 8: **End for**

Step 9: **End**

Algorithm 5.1: Discretized Hartley Transformation-based pre-processing

In Equation (5.3), Hartley spectrum coefficient of data is determined and symbolized as '$H(a,b)$'. From Equation (5.4), average value of DHT pair is portrayed and denoted as '$SID(Day, Hour)$'. In vector matrix, considered the data is represented in the procedure of rows and columns. Here, the rows are represented as 'P' and columns as 'Q'. Based on function '$cas(\theta) = \cos(\theta) + \sin(\theta)$', pre-processed data is calculated by using following mathematical formulation.

$$PD = H(a,b) = \sum_{Day=1}^{m} \sum_{Hour=1}^{n} DIS(Day, Hour) \left(Cos \left[2\pi \left(\frac{aDay}{P} + \frac{bHour}{Q} \right) \right] + Sin \left[2\pi \left(\frac{aDay}{P} + \frac{bHour}{Q} \right) \right] \right) \quad \text{..... Eqn (5.5)}$$

The result of pre-processed data 'PD' is obtained by using Equation (5.5). Here, real input air data is converted into real output with the elimination of noise air data. The process of DHT-based pre-processing is described in the Algorithm 5.1. Data on air quality is initially regarded as input at the input layer. Subsequently, noise data is eliminated by applying the Discretized Hartley Transformation function at the first hidden layer. In this case, day and hourly vector matrices and two-dimensional HT pairs are arranged on air data. With obtained result, pre-processed air quality data is attained effectively.

5.2.2 Feature Selection Technique

The proposed model performs feature selection process using CMLLR technique with obtained pre-processed data. Here, feature selection is carried on second hidden layer to abstract the relevant data from dataset. The extracted features help towards reducing the time complexity while predicting air pollution data. Generally, air quality data comprised with numerous features that causes difficulty of monitoring and controlling air pollution.

Therefore, feature selection is most significant process for analysis of air quality data in efficient manner. To select relevant features, CMLLR is performed in the projected model.

The correlative data are considered to select the most relevant features for AQI prediction. With estimation of maximum likelihood function, relevant features are selected for pollution forecasting. The construction of CMLLR feature selection is revealed in Figure 5.3. Pre-processed data is measured as input in order to identify the most important features from the dataset. Here, homogenized air

sample data is gathered for hour and daily basis. With collected air data, CMLLR technique is applied to identify maximum likelihood regression. Additionally, Maximum SPLE is considered to choose the pertinent features.

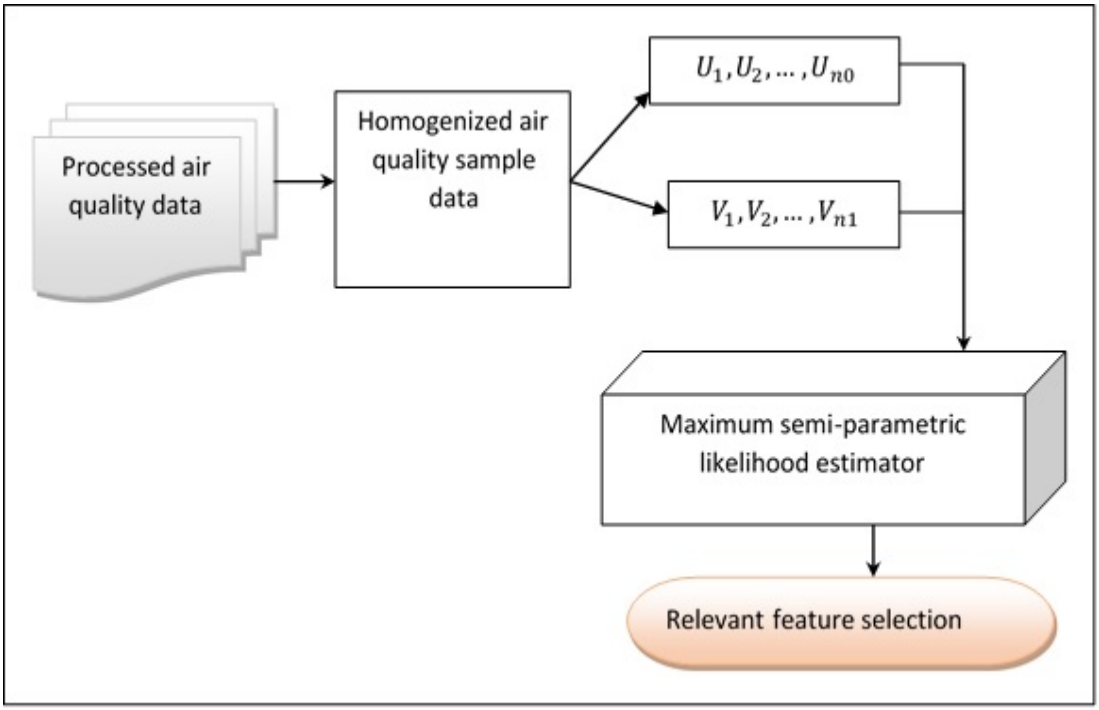

Figure 5.3: Structure of CMLLR - based Feature Selection

Here, the subjective air quality sample data from '$Prob(u|V = 0)$' at hour basis is denoted as '$(U_1, U_2, \dots, U_{n0})$' and subjective air quality trial data from '$Prob(u|V = 1)$' at daily basis is denoted as '$(V_1, V_2, \dots, V_{n1})$'. Then the normalized air quality trial data '$(U_1, U_2, \dots, U_{n0}; V_1, V_2, \dots, V_{n1})$' is represented as '$(T_1, T_2, \dots, T_n)$'. After considering air sample data, the logistic regression (LoR) function is processed with respect to the constrain '(α, β)'. The initial value of '(α, β)' is described as '(α_0, β_0)'. The result of logistic regression function using constrained is given below.

$$Prob(V = 1|U = u) = \frac{\exp(\alpha+\beta^{\tau}T)}{1+\exp(\alpha+\beta^{\tau}T)} \qquad \text{….. Eqn (5.6)}$$

The LoR function of arbitrary trial data '$Prob(V = 1|U = u)$' is obtained using Equation (5.6). Here, the scale values for hourly basis and daily basis on several stations in India is denoted as 'α' and 'β'The following formula is the highest semi-parametric likelihood estimator of '(α_0, β_0)'.

$$\frac{\partial l(\alpha,\beta)}{\partial \alpha} = \sum_{i=1}^{n} \frac{\exp(U_1,U_2,\ldots,U_{n0})(\alpha+\beta^{\tau}T_i)}{\exp(\alpha+\beta^{\tau}T_i)} \quad \text{….. Eqn (5.7)}$$

$$\frac{\partial l(\alpha,\beta)}{\partial \beta} = \sum_{i=1}^{n} \frac{\exp(V_1,V_2,\ldots,V_{n1})(\alpha+\beta^{\tau}T_i)}{\exp(\alpha+\beta^{\tau}T_i)} \quad \text{….. Eqn (5.8)}$$

From Equation (5.7) and (5.8), maximum semi-parametric likelihood estimator for air data at hour and daily basis is expressed. With estimated result of likelihood estimator, relevant features are selected and expressed as given below.

$$FS = \frac{\partial l(\alpha,\beta)}{\partial \alpha} \cup \frac{\partial l(\alpha,\beta)}{\partial \beta} \quad \text{….. Eqn (5.9)}$$

The result of feature selection 'FS' from air quality data is illustrated in Equation (5.9). Using the maximal SPLE in conjunction with double constraints '$\partial\alpha$' and '$\partial\beta$', relevant features are selected.

The procedure for feature selection using CMLLR is explained in Algorithm 5.2. For better feature selection process, pre-processed air quality data is considered. With each input pre-processed data, air data is processed at hour and daily basis. After initializing air data, LoR function is applied to homogenized air quality sample data. Then, the determined SPLE is evaluated to determine significant related features of air data. With the result of hidden layer output, data classification process is performed to classify data to obtain enhanced result of air pollution forecasting.

Input: Dataset 'DS', 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: Computationally efficient feature selection 'FS'

Step 1: **Initialize** processed air quality data 'PD'

Step 2: **Initialize** arbitrary air quality sample data '$(U_1, U_2, \dots, U_{n0})$' (i.e., processed station data athourly basis

Step 3: **Initialize** arbitrary air quality sample data '$(V_1, V_2, \dots, V_{n0})$' (i.e., processed station data at daily basis)

Step 4: **Begin**

Step 5: **For** each Dataset 'DS' with Sensors 'S' and processed air quality data 'PD'

Step 6: Evaluate logistic regression function for the homogenized air quality sample data as in equation (5.6)

Step 7: Evaluate maximum semi-parametric likelihood estimator as in equations (5.7) and (5.8)

Step 8: **Return** features selected 'FS'

Step 9: **End for**

Step 10: **End**

Algorithm 5.2: CMLLR -based Feature Selection

5.2.3 Concordance Correlative based Classification Technique

The proposed method finally performs CCLSSV-based data classification with the use of designated relevant features. The data classification process is carried at output layer to classify data for predicting air pollution. The significant features help to perform data classification to forecast air pollution with minimum error rate. The determination of AQL value, air pollutant levels is estimated. Here, AQL value ranges between '0' to '500' where higher value specifies lower air

quality. Thus, accurate forecasting of air pollution is achieved by using CCLSSV-based Classification model.

The classification of air data yields the least error in the nonlinear relationship between input and output variables. To improve forecasting, precise air contaminants are examined based on the link between the measured variables. First, a selection of characteristics is used to formulate the LSSV technique. The value of the coefficient between variables is then measured using the concordance correlation function. To achieve precise air pollution forecasting, data is characterized into classes based on the value of the correlative coefficient.

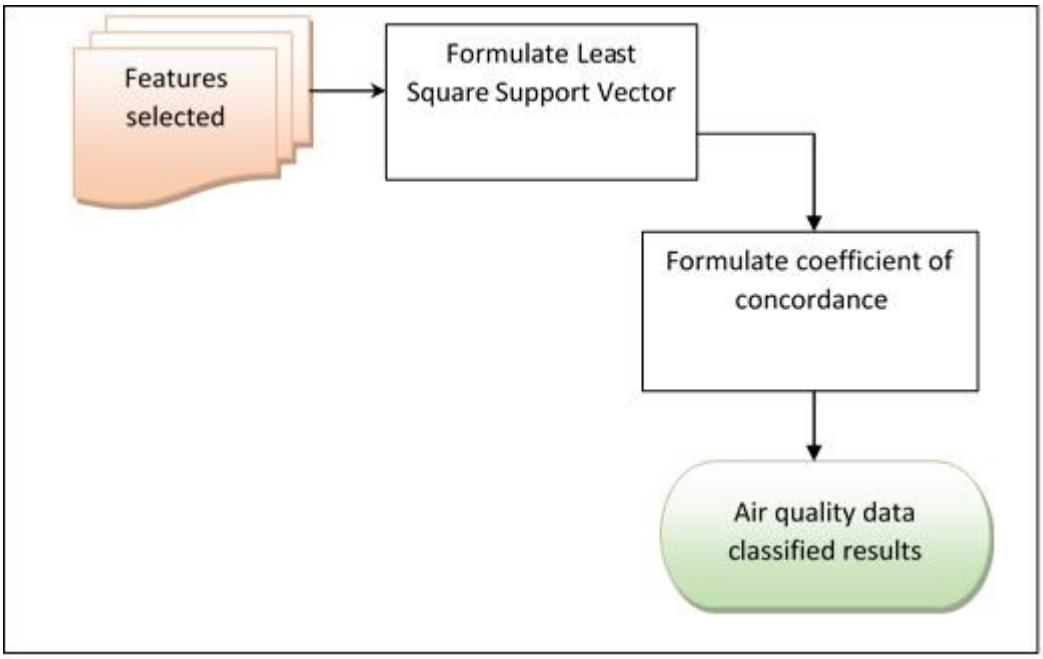

Figure 5.4: CCLSSV-based Classification

The classification model for forecasting air pollution is illustrated in Figure 5.4. From dataset, significant features are selected with maximum likelihood estimator with semi-parameter. After that, the relevant features are applied with LSSV. Then, concordance correlative coefficient value is estimated on input variables. Based on measured coefficient value, air quality data is effectively classified for pollution prediction. Here, the LSSV is measured as,

$$y(FS) = \omega^T \varphi(FS) + B \qquad \text{..... Eqn (5.10)}$$

From Equation (5.10), the result of least square support value with features selected is achieved and symbolized as '$y(FS)$'. Here, the non-linear mapping function for selected features is represented as '$\varphi(FS)$' and the weight is symbolized as 'ω' and 'B' represents bias. Then, coefficient value is estimated for data variables. The output of the classification is provided below and respect to error, it is considered as a minimum correlation.

$$min(\omega, Err) = \frac{1}{2}\omega^T\omega + \frac{1}{2}\gamma \sum_{k=1}^{n} Err_k^2 \qquad \text{..... Eqn (5.11)}$$

$$Subjecttoy_k = \omega^T \varphi(FS_k) + B + Err_k \qquad \text{..... Eqn (5.12)}$$

The concordance correlative function with minimum error is described in Equation (5.11) and (5.12). In the above formulation, the regularization constraint is 'γ' (i.e., setting index value) and 'Err' is denoted as the error. Then the concordance coefficient is estimated by using Correlation Coefficient of Kendall's Rank. The measurement of coefficient value is formulated below.

$$AQI = \tau = \frac{(N_{CP})\varphi(FS_k) - (N_{DP})\varphi(FS_k)}{\frac{n(n-1)}{2}} \qquad \text{..... Eqn (5.13)}$$

By using Equation (5.13), the coefficient of concordance 'τ' is estimated that represents the air quality index value. With estimated coefficient value, pair combinations of data are predicted. Here, any pair of observations '(Tr_i, Ts_i)' and '(Tr_j, Ts_j)' are said to be concordant. Otherwise, observations '(Tr_i, Tr_j)' and '(Ts_i, Ts_j)' is said to be discordant. The binomial coefficient for classify data with minimum error is denoted as '$\frac{n(n-1)}{2}$'. The estimated coefficient value ranges between '$-1 \leq \tau \leq 1$'.

According to measured coefficient value, air quality data is classified. As a result, the proposed model attains accurate data classification result with minimum error at output layer. Initially, the number of pre-processed data and selected features of data from dataset is considered as input. For each selected feature, least support vector is formulated to classify data with minimum error. After that, correlation coefficient value is estimated by applying Kendall's Rank Correlation function. Based on the coefficient value, data is classified into class for attaining accurate prediction. Based on the classified data, air pollution is forecasted with higher accuracy and minimum error.

5.3 Simulation Setting

The simulation analysis of proposed Discretized Regression and Least Square Support Vector (DR-LSSV) model is carried out by implementing Java JDK 1.8 language. For experimental purpose, Air Quality India dataset is considered from https://www.kaggle.com/rohanrao/air-quality-data-in-india. The input air quality dataset comprises 16 features with 2, 00,000 air quality data. For conducting simulation process, different number of air quality data in the range of 20,000 to 2,00,000 data is considered from dataset. Hence, the performance of proposed DR-LSSV model is estimated for efficient air pollution monitoring and control for IoT network. The simulation result is conducted by using the following parameters:

1) Air pollution forecasting accuracy,

2) Air pollution forecasting time, and

3) Error rate.

Input: Dataset 'DS', 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: classification of air pollutant data

Step 1: **Initialize** processed air quality data 'PD', features selected 'FS'

Step 2: **Begin**

Step 3: **For** each Dataset 'DS' with Sensors 'S', processed air quality data 'PD' and features selected 'FS'

Step 4: Formulate Least Square Support Vector-based Classification as in equation (5.10)

Step 5: Evaluate minimum correlation with respect to error as in equations (5.11) and (5.12)

Step 6: Measure Kendall's Rank Correlation Coefficient as in Equation (5.13)

Step 7: **If** '$\tau liesbetween -1 \; and \; -0.5$'

Step 8: **Then** air quality is very poor

Step 9: **End if**

Step 10: **If** '$\tau liesbetween -5 \; and \; 0$'

Step 11: **Then** air quality is poor

Step 12: **End if**

Step 13: **If** '$\tau liesequals \; 0$'

Step 14: **Then** air quality is good

Step 15: **End if**

Step 16: **If** '$\tau liesbetween \; 0 \; and \; +0.5$'

Step 17: **Then** air quality is satisfactory

Step 18: **End if**

Step 19: **If** '$\tau liesbetween +0.5 \; and \; 1$'

Step 20: **Then** air quality is moderate

Step 21: **End if**

Step 22: **Else**

Step 23: Air quality is very severe

Step 24: **End if**

Step 25: **End for**

Step 26: **End**

Algorithm 5.3: Process of Concordance Correlative Least Square Support Vector-based Classification

5.4 Performance Analysis

The experimental analysis of proposed Discretized Regression and Least Square Support Vector (DR-LSSV) model is compared with different exiting methods. Compared existing methods are, namely, Deep-AIR framework developed by Qi Zhang et al. (2022) and Integrated Multiple Directed Attention and Variational Auto Encoder (IMD-VAE) designed by Abdelkader Dairi et al. (2021) correspondingly. To evaluate the proposed DR-LSSV model, the following metrics are used. Performance is evaluated based on the following metrics with the help of table and graph given below.

5.4.1 Performance Analysis of Air Pollution Forecasting Accuracy

The ratio of number of air quality data that are correctly classified for accurate air pollution forecasting according to total number of input air quality data taken as input from database is described as air pollution forecasting accuracy. It is measured in percentage (%) and it is mathematically formulated as given:

$$Accuracy_{APF} = \frac{D_{AF}}{D_i} * 100 \qquad \text{…. Eqn (5.14)}$$

By using Equation (5.14), air pollution forecasting accuracy '$Accuracy_{APF}$' is estimated. Here, 'D_i' denotes number of input air data and 'D_{AF}' represents the number of data accurately forecasted for air pollution.

Sample calculation:

Existing Deep-AIR framework: Number of air data accurately forecasted for pollution is 16,904 and the number of input air quality data is 20,000. Thus, the air pollution forecasting accuracy is calculated as $Accuracy_{APF} = \frac{16{,}904}{20{,}000} * 100 = 84.52\%$.

Existing IMD-VAE: Number of air data accurately forecasted for pollution is 15,486 and the number of input air quality data is 20,000. Thus, the air pollution forecasting accuracy is calculated as $Accuracy_{APF} = \frac{15,486}{20,000} * 100 = 77.43\%$.

Proposed DR-LSSV model: Number of air data accurately forecasted for pollution is 17,672 and the number of input air quality data is 20,000. Thus, the air pollution forecasting accuracy is calculated as $Accuracy_{APF} = \frac{17,672}{20,000} * 100 = 88.36\%$.

Table 5.2: Forecasting accuracy of existing methods vs DR-MSV model

Number of air quality data	Air pollution forecasting accuracy (%)		
	Existing Deep-AIR framework	Existing IMD-VAE	Proposed DR-LSSV method
20,000	84.52	77.43	88.36
40,000	85.63	77.82	89.14
60,000	86.25	78.62	90.12
80,000	86.74	79.54	91.25
1,00,000	87.14	80.25	91.65
1,20,000	87.67	80.61	92.73
1,40,000	88.05	80.94	92.98
1,60,000	88.36	81.25	93.65
1,80,000	88.76	82.62	93.82
2,00,000	89.11	83.64	95.11

Table 5.2 demonstrates the comparative result analysis of air pollution forecasting accuracy with respect to different number of input air quality data from dataset. Here, the different numbers of air data in the range of 20,000 to 2,00,000 data are considered to conduct the experimental purpose. The table shows the comparison result of proposed DR-LSSV model with existing methods named as Deep-AIR framework developed by Qi Zhang et al. (2022) and IMD-VAE designed by Abdelkader Dairi et al. (2021) correspondingly. Hence, air pollution forecasting accuracy using DR-LSSV model is higher when compared to other existing methods.

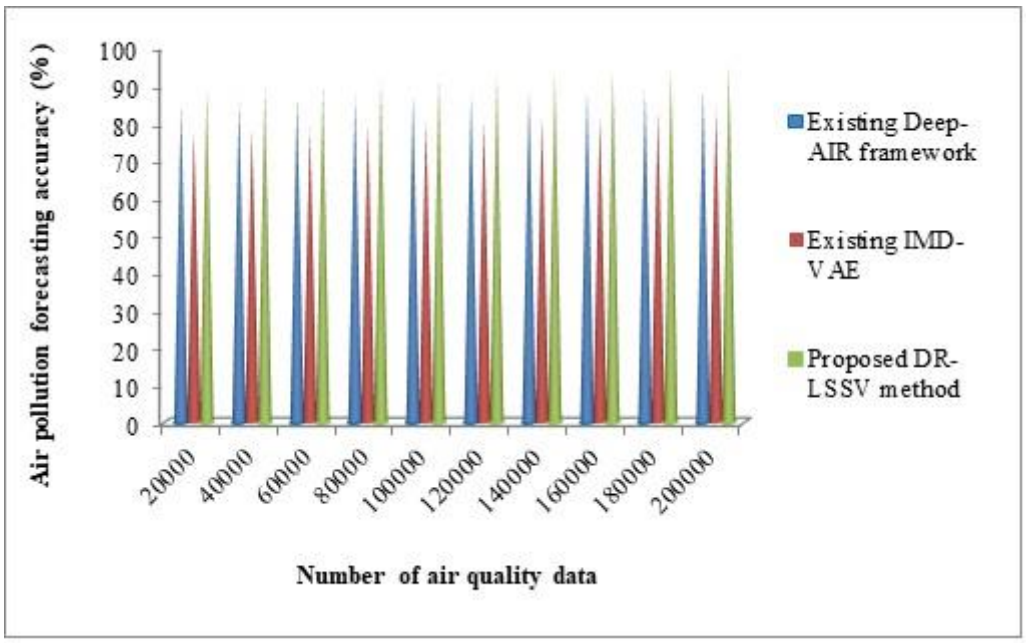

Figure 5.5: Forecasting accuracy of DR-LSSV model

The experimental description of air pollution forecasting accuracy for both proposed model and existing method is presented in Figure 5.5. With respect to different number of input air quality data in the range of 20,000 to 2,00,000 data, forecasting accuracy is obtained. From the result of experimental analysis, it is seen that the proposed DR-LSSV model provides accuracy to forecast air data effectively than the existing methods. Hence, the air pollution forecasting accuracy

using the proposed DR-LSSV model is higher when compared to other existing methods namely, Deep-AIR framework and Integrated Multiple Directed Attention and Variational Auto Encoder.

The regression-based feature selection process in proposed technique correctly identifying relevant features. Then Kendall's Rank Correlation Coefficient is estimated to classify data into different classes. Thus, input data are correctly classified for enhanced air pollution forecasting accuracy. From the result analysis, proposed DR-LSSV model increases air pollution forecasting accuracy by 5% and 14 % when compared with existing methods such as Deep-AIR framework developed by Qi Zhang et al. (2022) and IMD-VAE designed by Abdelkader Dairi et al. (2021).

5.4.2 Performance Analysis of Air Pollution Forecasting Time

Air pollution forecasting time is described as the estimation of amount of time involved to forecasting process of air pollution from air data according to total number of input air quality data from database. It is measured in terms of milliseconds (ms) and is evaluated by using the following mathematical equation.

$$Time_{APF} = D_i * Time(FS) \quad \text{…. Eqn (5.15)}$$

From above equation (5.15), air pollution forecasting time '$Time_{APF}$' is measured based on 'D_i' number of input air quality data. Here, '$Time(FS)$' specifies time taken to forecast single air data.

Sample calculation:

Existing Deep-AIR framework: Time taken by single air data for air pollution forecasting is 0.1025ms and the number of input air quality data is 20,000. Thus, the air pollution forecasting time is measured as $Time_{APF} = 20{,}000 * 0.1025 = 2050ms$.

Existing IMD-VAE: Time taken by single air data for air pollution forecasting is 0.1215 ms and the number of input air quality data is 20,000. Thus, the air pollution forecasting time is measured as $Time_{APF} = 20{,}000 * 0.1215 = 2430\, ms$.

Proposed DR-LSSV model: Time taken by single air data for air pollution forecasting is 0.065 ms and the number of input air quality data is 20,000. Thus, the air pollution forecasting time is measured as $Time_{APF} = 20{,}000 * 0.065 = 1300 ms$.

Table 5.3: Forecasting time of existing methods vs DR-MSV model

Number of air quality data	Air pollution forecasting time (ms)		
	Existing Deep-AIR framework	Existing IMD-VAE	Proposed DR-LSSV method
20,000	2050	2430	1300
40,000	2000	2360	1250
60,000	1950	2300	1200
80,000	1920	2200	1160
1,00,000	1860	2100	1070
1,20,000	1810	2020	1020
1,40,000	1760	1980	980
1,60,000	1710	1920	900
1,80,000	1660	1850	870
2,00,000	1600	1700	840

Table 5.3 illustrates the experimental result of air pollution forecasting time that obtained during pollution prediction with classified data. For experimental purpose, number of input air quality data is considered in the range of 20,000 to 2,00,000 data. Based on considered input data, comparison of proposed DR-LSSV

model is made with existing methods namely Deep-AIR framework and Integrated Multiple Directed Attention and Variational Auto Encoder respectively. From the result of simulation, proposed DR-LSSV model achieves minimum time to forecast air pollution from dataset when compared to other works.

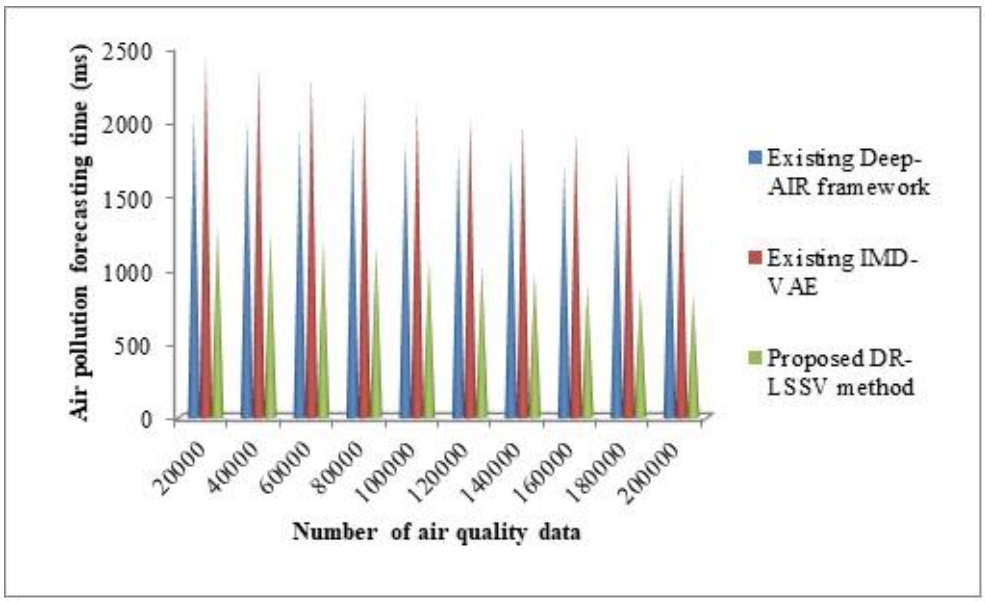

Figure 5.6: Forecasting time of DR-LSSV model

The experimental result analysis on air pollution forecasting time is described in Figure 5.6 according to different number of input air quality data. For the experimental purpose, different numbers of input data between 20,000 to 2,00,000 data are considered for accurate air pollution forecasting by classifying data. Figure shows the experimental results of forecasting time for proposed DR-LSSV model when compared with existing Deep-AIR framework and Integrated Multiple Directed Attention and Variational Auto Encoder. From the result analysis, DR-LSSV model attains minimum result of time for attaining effective pollution forecasting. Based on maximum semi-parametric likelihood estimator-based feature selection process, similar features for pollution forecasting are selected. Then, classifier is performed using selected features to effectively classify the homogenized air quality sample data. It helps to minimize the time taken to air pollution forecasting from input data. Hence, proposed DR-LSSV model minimizes

the air pollution forecasting time by 43% when compared to Deep-AIR framework developed by Qi Zhang et al. (2022) and 49% when compared to IMD-VAE designed by Abdelkader Dairi et al. (2021) correspondingly.

5.4.3 Performance Analysis of Error Rate

The error rate is defined as the measure of ratio of number of input air quality data that are incorrectly forecasted for air pollution according to the total number of input data taken from database. It is measured in percentage (%) and formulated as given below.

$$FPR = \frac{D_{IF}}{D_i} * 100 \qquad \text{…. Eqn (5.16)}$$

From Equation (5.16), error rate is estimated and represented as 'FPR'. In above expression, 'D_{IF}' denotes number of incorrectly forecasted air data and 'D_i' specifies total number of input air quality data.

Sample calculation:

Existing Deep-AIR framework: Number of air data incorrectly forecasted air pollution is 3,096 and the number of input air quality data is 20,000. Thus, the error rate is estimated as $FPR = \frac{3,096}{20,000} * 100 = 15.48\%$.

Existing IMD-VAE: Number of air data incorrectly forecasted air pollution is 4,514 and the number of input air quality data is 20,000. Thus, the error rate is estimated as $FPR = \frac{4,514}{20,000} * 100 = 22.57\%$.

Proposed DR-LSSV model: Number of air data incorrectly forecasted air pollution is 2,328 and the number of input air quality data is 20,000. Thus, the error rate is estimated as $FPR = \frac{2,328}{20,000} * 100 = 11.64\%$.

Table 5.4: Error rate of existing methods vs DR-MSV model

Number of air quality data	Error rate (%)		
	Existing Deep-AIR framework	Existing IMD-VAE	Proposed DR-LSSV model
20,000	15.48	22.57	11.64
40,000	14.37	22.18	10.86
60,000	13.75	21.38	9.88
80,000	13.26	20.46	8.75
1,00,000	12.86	19.75	8.35
1,20,000	12.33	19.39	7.27
1,40,000	11.95	19.06	7.02
1,60,000	11.64	18.75	6.35
1,80,000	11.24	17.38	6.18
2,00,000	10.89	16.36	4.89

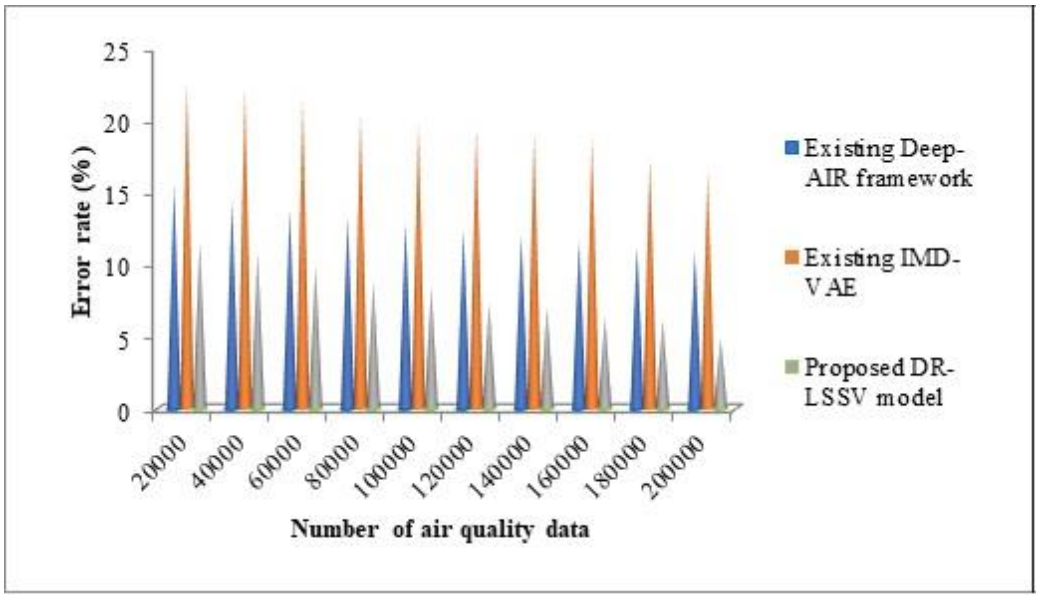

Figure 5.7: Error rate of DR-LSSV model

Table 5.4 describes the comparative result analysis of error rate with respect to different number of input air quality data in the range of 20,000 to 2,00,000 data using three methods. Above table values provide the result analysis of error rate

using proposed DR-LSSV model with existing Deep-AIR framework and IMD-VAE respectively. As a result, the proposed model achieves minimum error rate compared to other existing works.

The measure of error rate is demonstrated in the Figure 5.7 with respect to different data from dataset. Here, a number of air quality data are considered to conduct experimental purposes. From the Figure, DR-LSSV model provides enhanced result of air pollution forecasting with minimum error rate. It is obtained when compared to existing Deep-AIR framework and Integrated Multiple Directed Attention and Variational Auto Encoder respectively. Additionally, while increasing the number of input data for forecasting, the error rate gets varied using all methods. Hence, the proposed DR-LSSV model attains reduced error rate than the other methods.

With the application of Concordance Correlative Least Square Support Vector-based Classification, air data is forecasted with minimum error rate. The measured correlative coefficient value classifies air data for forecasting pollution. Thus, minimized error rate is provided to achieve better air pollution forecasting. Therefore, proposed DR-LSSV model attain reduced error rate by 37% and 59% when compared to existing Deep-AIR framework developed by Qi Zhang et al. (2022) and IMD-VAE designed by Abdelkader Dairi et al. (2021) respectively.

5.5 Summary

An efficient DR-LSSV model is proposed for enhancing the performance of forecast air pollution with higher accuracy and minimum time and error rate. The main aim of accurate air pollution forecasting is obtained by performing pre-processing, feature selection and data classification process. Initially, several air quality data are collected as input from dataset. For each input data, Discretized Hartley Transformation is applied to perform data pre-processing. In pre-processing, noise data is determined and eliminated. With obtained pre-processed

data, feature selection is carried out using Constrained Maximum Likelihood Linear Regression function. Here, regression function process is utilized to select relevant and irrelevant features of air data. With the removal of irrelevant air pollutant data, time taken for forecasting air pollution is reduced. Finally, Concordance Correlative Least Square Support Vector-based Classification is performed by analyzing the selected features to classify data. Based on Correlative coefficient value, data classified results are attained for efficient air pollution monitoring and controlling. Therefore, the proposed DR-LSSV model gives better performance of air pollution prediction with reduced time and error rate compared to existing state-of-the-art methods.