CHAPTER 6

PERFORMANCE ANALYSIS OF PROPOSED MACHINE LEARNING MODELS

6.1 Introduction

Air pollution leads to critical health issues in urban metropolises. The prediction of air pollution is a significant phase for air quality pollution management that minimizes the pollution's negative influence on the environment and people's health. In addition, Internet of Things (IoT) enabled systems are used to monitor and control air pollution in the cloud computing environment. Forecasting air pollution is difficult due to its dynamic nature, instability and better inconsistency. Air pollution is estimated using particulate matter, chemicals, or biological substances that hurt humans or other living beings in natural habitats and airspace. Several machine learning techniques have been developed, but it is complex to predict accurate air pollution forecasting based on the air quality index. With the identification of accurate air pollution, air quality is monitored with better accuracy and with minimal time consumption.

Predicting air pollution in several natural systems is an important area of research in the air quality dataset. Air pollution forecasting is the process of identifying air pollutant data. The data analysis from the dataset with significant relevant features or attributes is carried out using machine learning and deep learning techniques for IoT networks. Here, air quality data from the dataset is considered among numerous data for accurate air pollution prediction. Due to the presence of relevant and irrelevant features of data, the feature selection technique is significant in predicting air pollution. Thus, selecting similar features helps classify data into particular classes effectively to attain enhanced prediction accuracy and minimum time. The development of machine learning methods forecasts the air pollution quality with monitoring and controlling. Air pollution

forecasting needs to develop effective prediction modeling. The accurate forecasting of air pollution is achieved by classifying air data based on available air datasets. In the forecasting process, feature selection and data classification are necessary for reducing the error rate with efficient data prediction. Identifying forecasted air pollution data from input data is described as a data prediction process. By separating input data into various categories, a prediction process is carried out. Generally, organic and inorganic components with an inappropriate and massive number of data are comprised in the dataset.

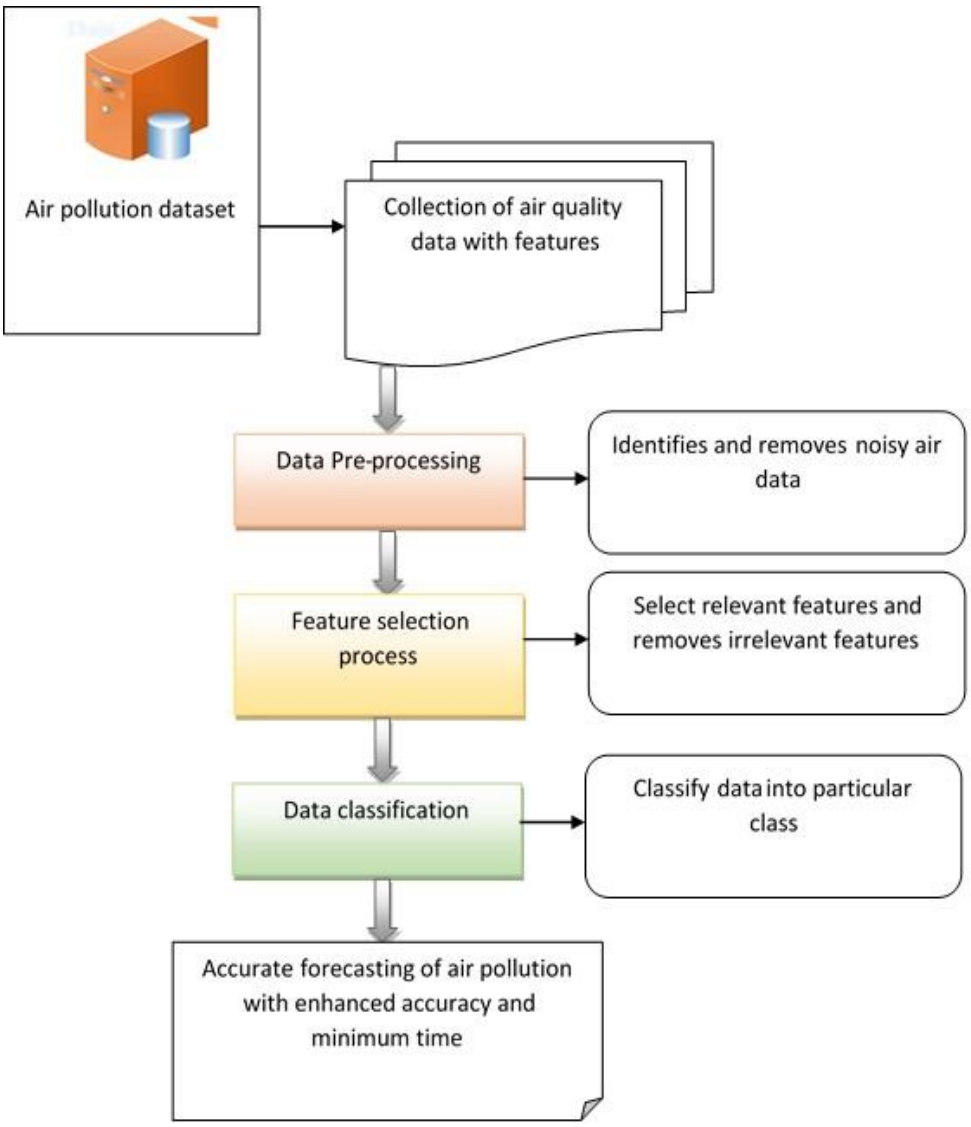

Figure 6.1: Process of accurate forecasting air pollution based on air data

The performance analysis of efficient air pollution prediction is carried out by using machine learning and deep learning techniques. The various research works are focused on predictive analysis. Though, the accurate prediction of data could have been performed more effectively. The overall process of the proposed model using different techniques is illustrated in Figure 6.1.

Currently, several feature selection and classification methods are presented to classify data at an early stage. Still, performance air pollution forecasting faces an issue for an accurate and reliable result. By identifying reasons for air pollution, the risk of pollutants is predicted and an enhanced result is provided. The various approaches were developed to predict data with enhanced performance. While performing the forecasting process, the feature selection is carried out for minimizing the time complexity. With the selected relevant features, classification technique is processed to classify data to predict forecasted air quality data. According to dataset, different air data with different features are taken for monitoring and controlling air pollutant data. The feature selection process analyzes the data to extract similar features from the database. Due to the presence of numerous air data in the environmental area, it is complicated and considered an inefficient process to analyze. Therefore, the data feature is estimated to select a similar feature. Using selected relevant features, data is categorized into classes for better performance prediction. Here, air pollution remains the paramount environmental problem in metropolitan cities. However, the performance error data through the forecasting of air quality was not minimized. Different proposed models are developed using various techniques to address these issues in predicting air pollutant data. Hence, the forecasting air pollution accuracy using reduced error using air quality dataset is achieved by proposing different techniques.

The primary purpose of designing the proposed model is to predict pollutant air data from the dataset. Proposed models are designed by performing data pre-processing, feature selection, and data classification. At first, air quality data with

features is considered from the dataset. For each input data, pre-processing is performed to remove noise data at an hourly and daily level of various stations. A feature selection process is carried out with attained pre-processed data to extract relevant features. From that, significant relevant data features are selected by applying the feature selection technique. In addition, irrelevant features are eliminated to attain efficient data classification to predict pollution forecasting results. This technique helps to improve the performance of data prediction with minimum time. After that, a data classification process is carried out to classify data based on selected features. It classifies air data into classes with minimal time complexity. Based on classified data, pollutant data is predicted with enhanced accuracy and minimum error rate.

Several deep learning and machine learning techniques have been developed to predict air pollution. From that, various classification techniques are most significant for improving data categorization to forecast air pollutant data with reduced error. One of the most critical processes is performance prediction based on similar data. Here, prediction accuracy was not attained at the required level. Thus, forecasting air pollution with increasing rates that assists in identifying accurate pollutants for forecasting air quality index is a significant problem. However, designed methods need to consider the importance of air data processing on an hourly and daily basis from various stations across multiple cities in India. In order to handle such limitations, three proposed models are developed to enhance the prediction performance of air pollution. The proposed models are Linear Regression and Multiclass Support Vector model, the Bilateral Transformative Broken-Stick Regression-based Quadratic Weighted Emphasis Boost Classification model, and Discretized Regression and Least Square Support Vector model. The developed models increase forecasting accuracy on air pollution with minimum time and error rate in an extensive manner.

The Linear Regression and Multiclass Support Vector model is developed first to perform accurate dropout prediction of air pollution by classifying data. The primary purpose of the proposed models is achieved using three different processes: pre-processing, feature selection, and classification technique. For accurate air pollution forecasting, several air quality data with the feature are collected from the considered dataset. After considering input data, pre-processing is performed using a WSW process. Here, noise presented in air data is removed and attains pre-processed data. After that, the linear regression function is applied to extract significant data features. Here, the gradient function is determined for selecting data features. Based on the result of the gradient function, relevant and irrelevant features are identified. The relevant features are selected for predicting pollution forecasting, and irrelevant features are removed. The selected feature is utilized to classify data effectively. With the aid of selected relevant features, multiclass support vector classification is carried out to classify data. By formulating a linear classifier, the air quality index value is estimated. Air data are classified into different classes for accurate pollution forecasting based on AQI value. Hence, the LR-MSV model effectively classifies data that improves pollution forecasting accuracy with minimum time. Here, memory utilization for storing air data is higher while forecasting air pollution. Hence, the next proposed model is developed to achieve minimum memory consumption with better air pollution forecasting.

The bilateral Transformative Broken-Stick Regression-based Quadratic Weighted Emphasis Boost Classification (BTBSR-QWEBC) model is introduced next for predicting air pollution with reduced memory consumption. The objective of the proposed model is to classify data to enhance the performance of air pollution forecasting. Here, pre-processing, feature selection, and classification techniques are carried out to attain an accurate forecasting process. Initially, the number of features and air quality data is considered input from the dataset. For

each input data, pre-processing is carried out using bilateral discretized Z- wavelet transform. The transformation process identifies noise data and removes it to obtain pre-processed data. With pre-processed data, Otsuka Indexive Broken-stick regression process is carried out to determine relevant features from the dataset. In order to select similar relevant features, the Otsuka similarity coefficient is estimated between data features. According to the measured coefficient value, relevant features are selected, and irrelevant features are eliminated. The relevant features help to forecast air data with minimum time. The weighted emphasis boost classification technique is applied to classify data based on selected relevant features. The estimated result of the AQI value classifies data into a particular class with a minimal quadratic error rate. The classified result helps to forecast air pollution with better accuracy and minimum error rate. Thus, the BTBSR-QWEBC model increases the performance of data classification for efficient air pollution forecasting with reduced memory consumption. While reducing memory complexity during air pollution prediction, the time taken to predict air pollution is high. Hence, the following model is proposed with different processes for efficient pollution forecasting with minimum complexity and time.

Finally, the Discretized Regression and Least Square Support Vector model is proposed to predict the performance of air pollution monitoring from a dataset with higher accuracy. The main aim of forecasting is attained with the application of pre-processing, feature selection, and data classification processes. Initially, air quality data with various features are gathered from the database for pollution prediction. The collected air data from the dataset is presented in the input layer. With considered input data, pre-processing is performed at the first hidden layer by applying discretized Hartley transformation. The transformation process converts real input to real output by removing noise air data. After that, pre-processed data is forwarded to the second hidden layer to perform the feature selection process. A constrained maximum likelihood linear regression process is utilized to select

relevant data features. The LoR function is estimated for extracting features from the dataset. Followed by correlative target feature projection matching technique is performed at the third hidden layer. The likelihood estimator is determined to identify relevant and irrelevant features. The relevant features of air quality data are selected to perform data classification. After selecting relevant features, classification is carried out at the output layer. The similarity between selected features is estimated using a concordance correlative least square support vector. The coefficient of concordance value helps to classify data to forecast air pollution with error. Hence, the DR-LSSV model achieves higher air pollution forecasting accuracy with reduced time and error rate.

6.2. Performance Analysis of Proposed Models

The performance analyses of the proposed LR-MSV model, BTBSR-QWEBC model, and DR-LSSV model are carried out by implementing using JAVA JDK 1.8 language. The simulation work is conducted using the air quality India dataset from https://www.kaggle.com/rohanrao/air-quality-data-in-india. The considered dataset consists of 16 attributes and 2,00,000 instances. The air quality dataset includes 16 attributes (i.e., features), namely, City, Date, PM 2.5, PM 10, Nitric Oxide (NO), Nitric dioxide (NO2), Any nitric x-oxide (NOx), Ammonia (NH3), Carbon monoxide (CO), Sulphur dioxide (SO2), Benzene, Toluene, Xylene, air quality indices (AQI) and air quality indices bucket (AQI_Bucket).

The experimental analysis of different proposed models is conducted by comparing them with existing methods. Here, the proposed models are compared to existing methods, namely, the 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), respectively. The results are analyzed based on the considered air quality dataset to attain the efficient result of air pollution forecasting. The results are compared and analyzed with the help of a table and graph given below based on the following parameters:

- Air pollution forecasting accuracy,
- Air pollution forecasting time,
- Error rate, and
- Memory consumption.

6.2.1 Performance Analysis of Air Pollution Forecasting Accuracy

Air pollution forecasting accuracy is defined as the ratio of several air quality data from the dataset that are correctly forecasted pollutants according to the total number of input air quality data. The accurate forecasting of air pollution is attained by utilizing the Air Quality Index value of the Indian government. Accuracy is measured in percentage (%). The proposed models are said to be more efficient when there is higher accuracy in pollution forecasting.

Table 6.1: Forecasting accuracy of existing methods vs proposed models

Number of air quality data	Air pollution forecasting accuracy (%)				
	Existing Deep-AIR framework	Existing IMD-VAE	Proposed LR-MSV	Proposed BTBSR-QWEBC	Proposed DR-LSSV
20,000	82.25	76.25	83.62	85.25	87.96
40,000	83.66	77.52	84.25	86.62	88.76
60,000	83.93	77.62	84.68	87.63	89.96
80,000	84.22	78.56	85.63	88.25	91.01
1,00,000	85.63	79.52	86.11	89.64	91.22
1,20,000	86.89	80.04	87.65	90.25	91.52
1,40,000	87.61	80.25	88.62	90.63	92.04
1,60,000	88.11	80.63	89.52	91.25	92.66
1,80,000	88.24	81.72	89.82	91.36	92.81
2,00,000	89.62	82.46	90.36	91.67	93.66

The experimental result of forecasting accuracy on air pollution is tabulated in Table 6.1 with respect to different number of air quality data. Here, air quality data in the range of 20,000 to 2,00,000 data are considered for conducting the experimental purpose. From the table values, while increasing the number of input data, accuracy on forecasting air pollution is also getting varied in all the models. Here, the table shows comparison of proposed LR-MSV model, BTBSR-QWEBC model and DR-LSSV model 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) respectively. Hence, accuracy on forecasting DR-LSSV model offers comparable values than the other state-of-the-art methods.

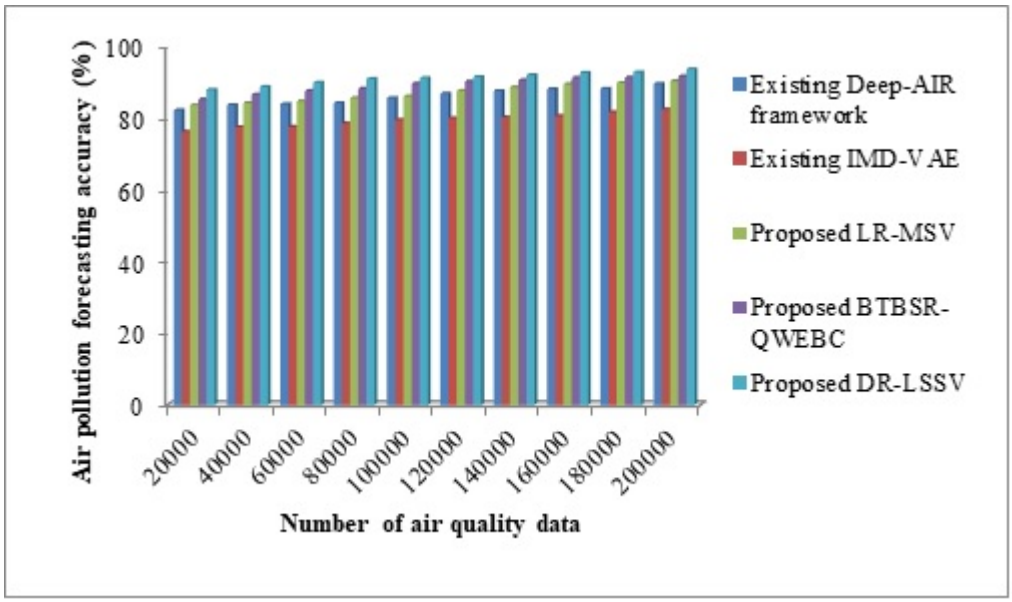

Figure 6.2: Air pollution forecasting accuracy of proposed models

Based on the table values, the graph is drawn as shown in Figure 6.2 for the analysis of proposed models. The number of input air data from dataset in the range 20,000 to 2,00,000 is considered for conducting the experimental purpose. The figure portrays comparison result of different proposed models, namely, LR-MSV model, BTBSR-QWEBC model and DR-LSSV model with various existing methods. The compared existing methods are Deep-AIR framework and IMD-VAE. For example, 20,000 air data are considered for experimental performance.

From the performance analysis, existing Deep-AIR framework and IMD-VAE obtains 82.25% and 76.25% of forecasting accuracy. But 83.62% %, 85.25% and 87.96% of forecasting accuracy is achieved using proposed LR-MSV, BTBSR-QWEBC and DR-LSSV models. From the obtained result, DR-LSSV model achieves better results of pollution forecasting accuracy.

By performing Maximum Likelihood Linear Regression function, DR-LSSV model effectively selects relevant features of data. Based on selected features, correlation coefficient value is measured to classify air data. It helps to forecast pollutant air data correctly with enhanced accuracy. Finally, LR-MSV model improves the forecasting accuracy by 5%, BTBSR-QWEBC model improved by 8% and DR-LSSV model increased by 10% when compared with other state-of-the-art methods. Therefore, proposed DR-LSSV model achieves better result of air pollution forecasting accuracy while comparing with other methods namely Deep-AIR framework developed by Qi Zhang et al. (2022) and IMD-VAE designed by Abdelkader Dairi et al. (2021) respectively.

6.2.2 Performance Analysis of Air Pollution Forecasting Time

The time taken to forecast air quality data is referred as the air pollution forecasting time. Therefore, forecasting time is defined as the product of total number of input air quality data and the time taken to forecast single air data. The time is estimated in terms of milliseconds (ms).

The experimental result of time occurred while forecasting air pollution using proposed and existing method is tabulated in Table 6.2. The table provides the comparison result of proposed and existing methods according to various numbers of input air quality data. For experimental purpose, number of air quality data ranges from 20,000 to 2,00,000 is considered.

Table 6.2: Forecasting time of existing methods vs proposed models

Number of air quality data	Air pollution forecasting time (ms)				
	Existing Deep-AIR framework	Existing IMD-VAE	Proposed LR-MSV	Proposed BTBSR-QWEBC	Proposed DR-LSSV
20,000	2120	2500	2000	1680	1360
40,000	2060	2420	1960	1620	1300
60,000	2000	2350	1860	1550	1250
80,000	1940	2260	1820	1510	1200
1,00,000	1890	2150	1800	1480	1150
1,20,000	1800	2070	1770	1450	1100
1,40,000	1750	2020	1700	1420	1000
1,60,000	1700	1980	1680	1400	950
1,80,000	1680	1800	1630	1350	900
2,00,000	1650	1750	1600	1300	860

From the table value, while increasing input air data, the time taken for forecasting pollution gets varied in all the methods. In the above tabulated values, proposed LR-MSV, BTBSR-QWEBC and DR-LSSV models are compared with other existing methods such as Deep-AIR framework developed by Qi Zhang et al. (2022) and IMD-VAE designed by Abdelkader Dairi et al. (2021) respectively. From the experimental work, the proposed DR-LSSV model resulted with minimum time of pollution forecasting. With help of the taken values, the graph is drawn in Figure 6.3 for enhanced performances of proposed models. Here, the number of air quality data in the range of 20,000 to 2,00,000 is considered from database for experimental purpose. As shown in the figure, time to forecast air

pollution using the proposed model provides better performance. Moreover, while increasing the number of input data, the time taken to forecast data gets varied.

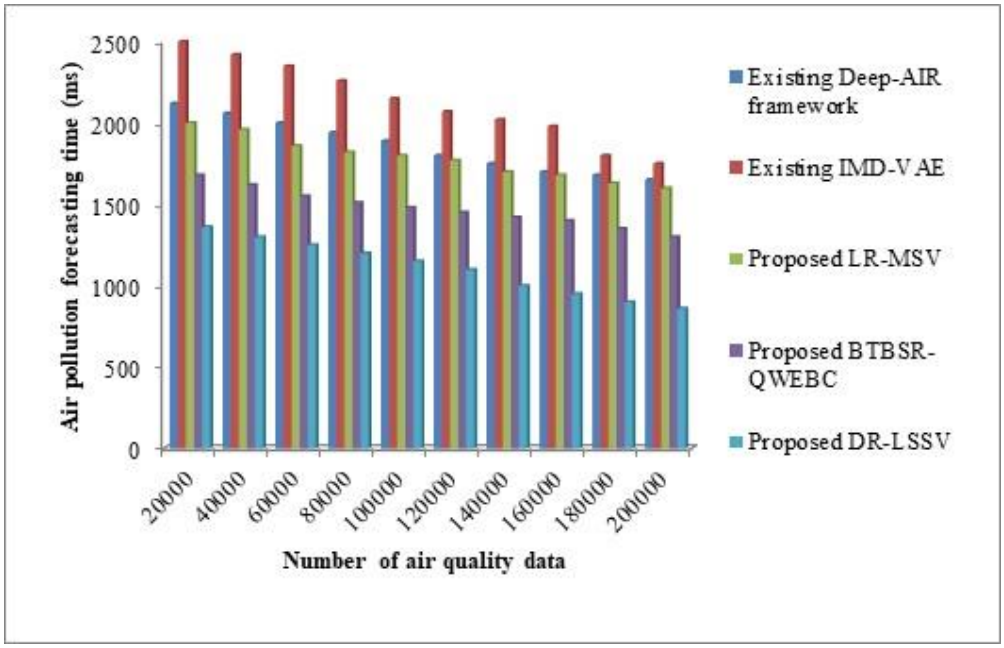

Figure 6.3: Air pollution forecasting time of proposed models

For example, 20,000 different air data are considered from dataset for experimental purpose. From performance analysis, the existing Deep-AIR framework and IMD-VAE obtains 2120 ms and 2500 ms of forecasting time. Whereas 2000ms, 1680ms and 1360 ms of time is achieved in LR-MSV, BTBSR-QWEBC and DR-LSSV models respectively. From the result, DR-LSSV model attains efficient result of minimum air pollution forecasting time than other models. The proposed DR-LSSV model performs maximum semi-parametric likelihood estimator for feature selection. The regression function extracts relevant features to classify air data. Based on selected features, forecasted air pollutant data are correctly detected with minimum time. Consequently, LR-MSV model reduces the time for pollution forecasting by 10%, BTBSR-QWEBC model reduced by 25% and DR-LSSV model reduced by 45% compared with state-of-the-art methods. Therefore, time taken for predicting pollutants using DR-LSSV model obtains better result when compared with existing Deep-AIR framework developed by Qi

Zhang et al. (2022) and IMD-VAE designed by Abdelkader Dairi et al. (2021) respectively.

6.2.3 Performance Analysis of Error Rate

Error Rate is defined as the ratio of incorrectly forecasted air data as for efficient pollution prediction according to the total number of air quality data considered as input from dataset. It is measured in terms of percentage (%).

The Table 6.3 shows the analysis of error rate with respect to different input band image pairs. From air quality dataset, the number of air data in the range of 20,000 to 2,00,000 is considered for predicting air pollution. Here, the proposed LR-MSV, BTBSR-QWEBC and DR-LSSV models are compared with existing Deep-AIR framework developed by Qi Zhang et al. (2022) and IMD-VAE designed by Abdelkader Dairi et al. (2021) respectively. With the assistance of the table values, the graph is plotted as depicted at Figure 6.4 for performances analysis of proposed models.

The figure 6.4 shows the comparison result of proposed models with existing Deep-AIR framework and IMD-VAE. To conduct the experimental purpose, the number of air data in the range of 20,000 to 2,00,000 was considered. Here 20,000 air data are considered for example to show experimental performance result. From the performance analysis, existing Deep-AIR framework and IMD-VAE obtains 17.75 % and 23.75% of error rate, whereas 16.38%, 14.75 % and 12.04 % of error rate is attained using proposed LR-MSV, BTBSR-QWEBC and DR-LSSV models respectively. At Figure 6.4, it is denoted that DR-LSSV model resulted with minimum error rate than the other techniques.

Table 6.3: Error rate of existing methods vs proposed models

Number of air quality data	Error Rate (%)				
	Existing Deep-AIR framework	Existing IMD-VAE	Proposed LR-MSV	Proposed BTBSR-QWEBC	Proposed DR-LSSV
20,000	17.75	23.75	16.38	14.75	12.04
40,000	16.34	22.48	15.75	13.38	11.24
60,000	16.07	22.38	15.32	12.37	10.04
80,000	15.78	21.44	14.37	11.75	8.99
1,00,000	14.37	20.48	13.89	10.36	8.78
1,20,000	13.11	19.96	12.35	9.75	8.48
1,40,000	12.39	19.75	11.38	9.37	7.96
1,60,000	11.89	19.37	10.48	8.75	7.34
1,80,000	11.76	18.28	10.18	8.64	7.19
2,00,000	10.38	17.54	9.64	8.33	6.34

With the application of concordance correlative least square support vector-based classification in DR-LSSV model, air pollutant data is forecasted with minimum error rate. While performing forecasting process, extracted relevant features are considered and coefficient value is estimated. Based on estimated coefficient value, air data is classified into different classes. This helps to correctly forecast the air pollution with minimized error rate. From the experimental result, proposed LR-MSV model reduces the error rate by 22 %, BTBSR-QWEBC model reduced error rate by 36 % and DR-LSSV model reduced error rate by 47% when compared with other state of the art methods. Therefore, the proposed DR-LSSV model achieves minimum error rate while comparing with other existing methods, namely, Deep-AIR framework developed by Qi Zhang et al. (2022) and IMD-VAE designed by Abdelkader Dairi et al. (2021) respectively.

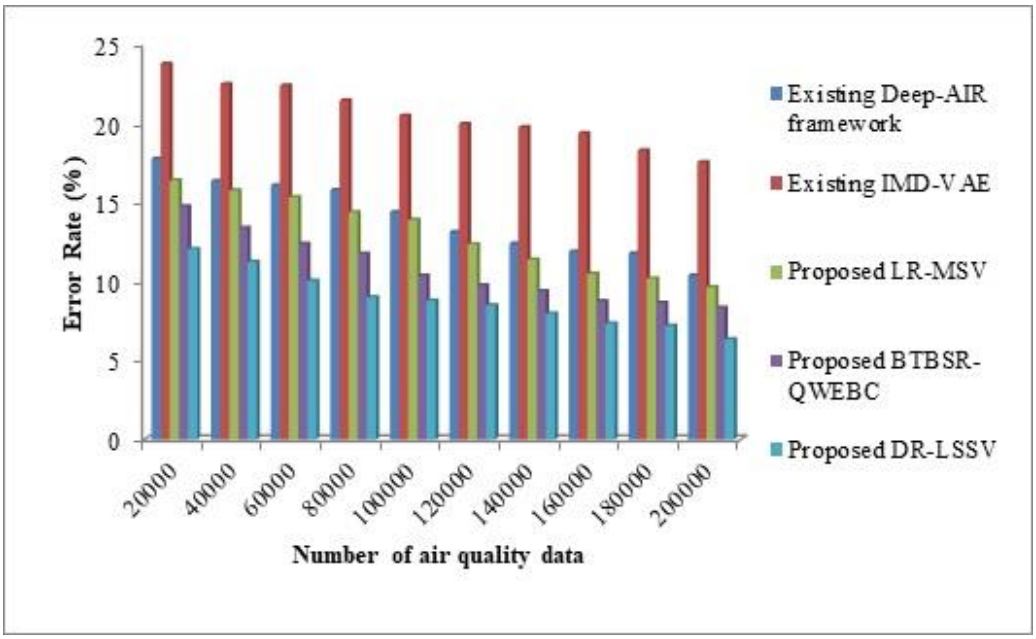

Figure 6.4: Error rate of proposed models

6.2.4 Performance Analysis of Memory Consumption

The memory space acquired for storing the forecasting air data of air pollution is termed as memory consumption. It is described as the product of number of input air quality data and memory utilized for storing single air data. Memory consumption is measured in terms of Megabytes (MB). if the consumption of memory by air data is lower, then the proposed model is said to be more efficient. Table 6.4 presents the experimental values of memory consumption for proposed and existing models. The table shows the comparison of proposed LR-MSV, BTBSR-QWEBC and DR-LSSV models according to input air quality data from air dataset. For experimental purpose, different air data are considered in the range of 20,000 to 2,00,000 for pollution forecasting. From the table value, while increasing the number of input data, the memory utilized to forecast pollution also correspondingly varied in all the models. From the experimental analysis, DR-LSSV model attains minimum memory space utilization for forecasting pollutant air data than the existing methods namely Deep-AIR framework developed by Qi Zhang et al. (2022) and IMD-VAE designed by Abdelkader Dairi et al. (2021)

respectively. Based on the table values, the graph is plotted as shown in Figure 6.5 to analyze the performance of proposed models.

Table 6.4: Memory consumption of existing methods vs proposed models

Number of air quality data	Memory Consumption (MB)				
	Existing Deep-AIR framework	Existing IMD-VAE	Proposed LR-MSV	Proposed BTBSR-QWEBC	Proposed DR-LSSV
20,000	117	112	103	100	97
40,000	112	107	97	95	93
60,000	109	104	95	92	89
80,000	101	96	90	88	86
1,00,000	98	95	89	86	84
1,20,000	96	93	85	83	81
1,40,000	95	92	83	81	78
1,60,000	92	87	81	79	76
1,80,000	89	82	79	77	75
2,00,000	85	77	75	72	70

For the experimental work, various input air quality data are considered in the range of 20,000 to 2,00,000. The figure shows the comparison result of proposed models with existing Deep-AIR framework and IMD-VAE. For example, let us consider 20,000 different air data for experimental purpose. From the experimental result, it is seen that existing Deep-AIR framework and IMD-VAE obtains 117 MB and 112 MB of memory consumption. It is to be noted that, 103 MB, 100 MB and 97 MB is achieved in proposed LR-MSV, BTBSR-QWEBC and DR-LSSV models.

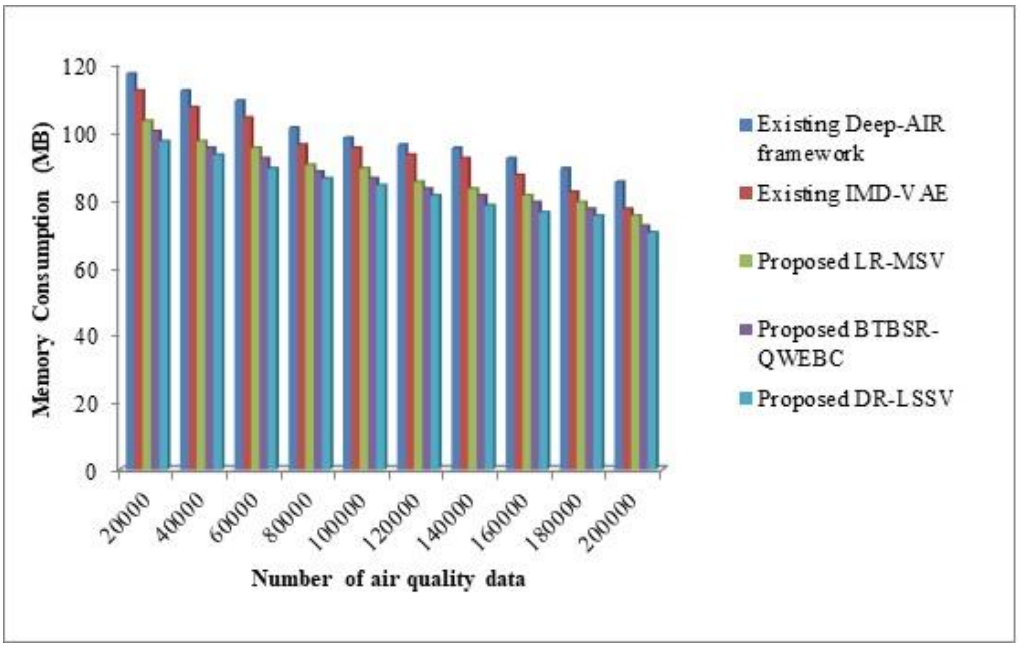

Figure 6.5: Memory consumption of proposed models

The DR-LSSV model achieves improved performance of air pollution forecasting with reduced space complexity. This is because of performing feature selection and classification process. The regression function selects significant relevant features of air data. Based on selected features, data is classified, and space required for forecasting air pollutant data is minimized. As a result, LR-MSV model reduces memory consumption by 9%, BTBSR-QWEBC reduced by 12% and DR-LSSV model minimized by 14% when compared with other state of the art methods. Therefore, the proposed DR-LSSV model achieves efficient result of memory consumption while comparing with other existing methods, namely, Deep-AIR framework developed by Qi Zhang et al. (2022) and IMD-VAE designed by Abdelkader Dairi et al. (2021) respectively.

6.3 Summary

A perfect illustration is discussed on analysis of proposed LR-MSV model, BTBSR-QWEBC model and DR-LSSV model. Theoretical analysis and experimental results show that the proposed models are designed for forecasting air

pollution by extracting features on air data with minimum error rate. The enhancement of air pollution forecasting with higher accuracy and minimum time is achieved by proposing three different models. Thus, it provides accurate air pollution forecasting with minimal time and memory consumption. First, LR-MSV model is introduced for enhancing the performance of air pollution prediction. pre-processing is applied on each input air data to remove noise and obtain enhanced pre-processed data. After that, regression function is used to select relevant features. Based on relevant features, air quality data are classified using a linear classifier process. It correctly forecasts air pollution with higher accuracy and minimum time. Next, BTBSR-QWEBC model is developed to achieve enhanced result of pollution forecasting with minimum error rate than LR-MSV model. There, pre-processing is carried out to eliminate noise air data. From the pre-processed data, significant relevant features are extracted by using similarity coefficient values. Based on selected features, data is classified by applying weighted emphasis boost technique. It effectively forecasts air pollution with higher accuracy, minimum time, and minimum memory consumption than LR-MSV model. Lastly, DR-LSSV model is proposed to forecast air pollution. For each input air data, pre-processing is performed first to remove noise data. Followed by the optimal features are selected using maximum likelihood linear regression process. Then, concordance correlative coefficient is estimated to forecast pollutant air data from dataset. From the simulation results, the proposed models reduce the time taken to forecast air pollutant data and improve accuracy with reduced memory consumption and also with minimized error rate compared with other of the art methods.