CHAPTER 1

INTRODUCTION

One of the primary problems resulting from the unrestrained use of conventional energy sources is air pollution. Using the Air Quality Index (AQI) to forecast air pollution in the environment at various times and locations is the most important application of science and technology. The measurement of AQI value is used for the forecast of air pollutant concentrations. The pollutant causes major harm to humans and other living organisms, namely, lung diseases, respiratory problems and cardiovascular problems. Due to the present issues, fitness conditions change, and mental problems crop up. Air pollution prediction is essential for reducing health issues for humans and living organisms. Thus, accurate air pollution forecasting helps to increase the safety level of climatic changes. Additionally, it gives instructions for simple daily planning, avoiding regions with high alert levels, and putting in place efficient pollution management measures.

The air pollution forecasting is carried out in highly populated areas as it directly affects the cost-effective activity of countries and the health of their inhabitants. The World Health Organization (WHO) has compiled a list of the worldwide causes of air pollution in both rural and urban areas. The components and compounds in air pollution have an impact on the health of people, animals and plants. Numerous health problems affect all living creatures in the environment and are based on air quality. Smog, aerosol production, reduced visibility, global warming, acid rain, early deaths, and impaired visibility are all effects of poor air quality. As a result, monitoring and managing air quality in metropolitan areas is a crucial and difficult task. To improve both human health and governmental decision-making, an effective method must be used to monitor polluted air data.

Air pollution is an important issue in every developed and developing country due to fast industrialization. Anticipating the occurrence of pollution is crucial for mitigating its deleterious impact on the environment and public health. When it comes to air pollution forecasting, data-driven prediction techniques usually work well, but they struggle when it comes to actually executing the prediction modeling. For current data-driven strategies to achieve complex relationships between geographical and temporal air pollution variables, monitoring and forecasting air pollution is essential. Additionally, a greater understanding of the specific environmental issues that our planet faces has spread throughout the world due to concerns for the environment, human health, and welfare.

The challenging issue of air quality forecasting necessitates the application of the temporal and spatial features of the pollutants. Even with the development of several machine learning approaches, it remains difficult to identify specific pollutants for AQI predictions. Numerous air pollution levels are harming human health. Also, improper identification of air pollution outcomes is more danger to humans and living creatures. Therefore, predicting air pollution techniques is necessary to reduce the risk.

1.1 Air Pollution Forecasting

According to the WHO, air pollution increases death rates, is detrimental to health, and is extremely susceptible to skyscraper conservation. For the purpose of predicting air pollution quality, machine learning techniques are being developed. The detection of air pollution allows for the accurate monitoring of air quality. Gases, liquid droplets, and solid materials all contribute to air pollution. The atmosphere's usual properties are altered by the polluted air. Automobiles, factories, and forest fires are some common home items that contribute significantly to air pollution. Public health is seriously endangered by changes in particulate matter, carbon monoxide, ozone, nitrogen dioxide, and sulfur dioxide levels in the air.

Thus, The WHO demonstrates air data which shows how the level of pollutants affects health. Generally, air quality is directly associated with the earth's weather conditions and ecosystems globally. Burning fossil fuels in the greenhouse produces a gas discharge and causes pollutant air data. The pollution occurring in the air is overcome through various rules, namely, a win-win approach for climate and health, minimizing disease-attributable weight, and improving climate change.

The WHO provides the following guidelines for air pollution in the environment: -

- WHO encourages healthy planes to address indoor and outdoor air pollution issues. In addition, health benefits are achieved by using climate change mitigation policies,
- An improvement of normative management tools and the condition of authoritative advice supports WHO Member States in overcoming health issues caused by sources of air pollution,
- The organization manages air pollution from national, regional and global levels by monitoring the actions on global development, and
- Air pollution for the health and environment sector is illustrated through digital outreach and partnerships.

The environment is polluted by the interaction of solid materials, liquid droplets, and gases. Air pollution can be caused by a range of factors, including burning fuel in homes, industry emissions, car exhaust, power generation, burning waste, agricultural activities, dry soil, and more. Numerous variables lead to the creation of air pollution at various sites and stations in both rural and urban areas. Urban areas close to the ocean are polluted by factors like road dust, sea salt, and diesel engine smoke. Similar to urban pollution, rural pollution results from using particulate matter made up of soil, cook stove smoke, and forest fire smoke.

1.2 Sources of Air Pollution

The atmosphere of the world is contaminated by numerous technologies with solid and liquid particles. Air pollution in the environment of the planet is caused by industrial units, vehicles, hydroplanes and aerosol cans. People's emissions of cigarette smoke are also a significant contributor to air pollution. The term "anthropogenic sources" describes air pollution that is caused by humans. In addition, wildfires and volcanoes are natural causes of air pollution. The following figure describes the main cause of air pollution.

The main environmental sources of air pollution are depicted in Figure 1.1. Fossil fuel, agricultural, natural, industrial, and vehicle transportation are the sources. The pollution caused by transportation, heating, industrial process and power plants is represented as fossil sources. The agriculture source is described as forming where fertilizers and livestock are produced in cities and particles.

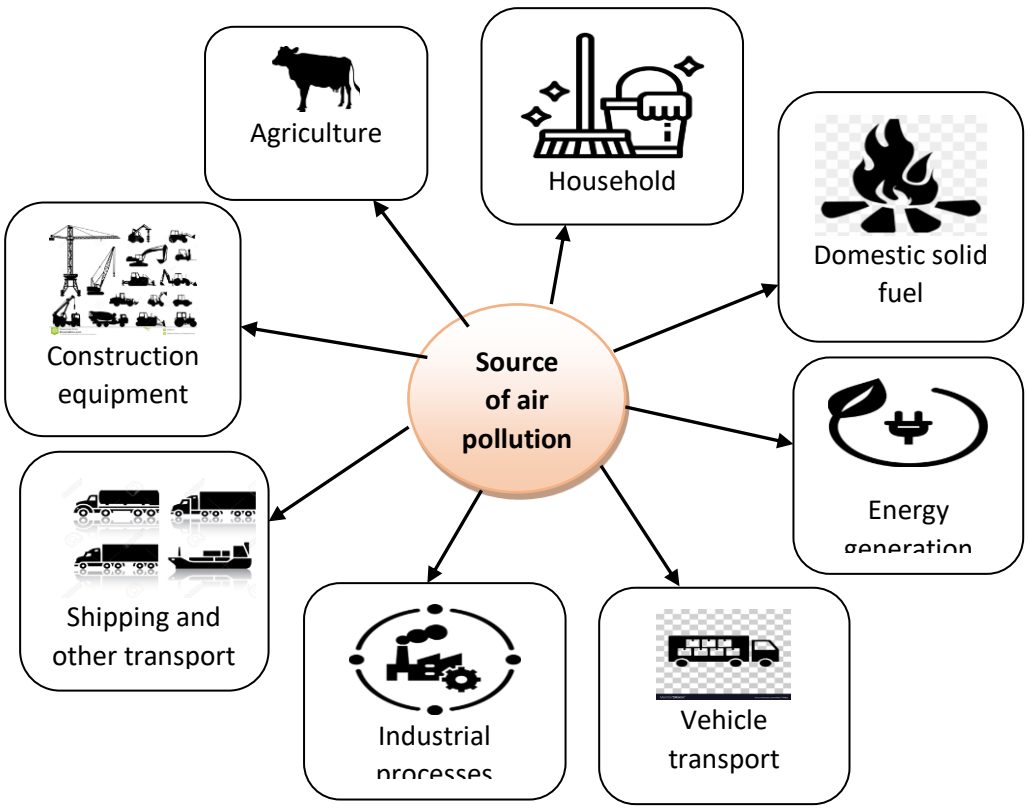

Figure 1.1: Air pollution source

The burning substances in the house are kerosene, wood and coal, which lead to air pollution inside the house. The ash and smoke generated by flaming are attached to walls, food and clothing, making breathing difficult for humans and animals. The generated air pollution is the primary cause of many deaths in the earth's environment.

Air pollution is frequently determined in large cities based on the discharges from many different sources. In contrast, the emission of air or gases from sources like mountains or tall buildings helps prevent air pollution from spreading. There are two different kinds of pollutants: primary and secondary. Primary pollutants directly damage the air and generate pollution, whereas, secondary pollutants are produced when primary pollutants mix and have an impact on the environment, plants, animals and living beings. As a result, early identification by air quality data prediction enhances both the environment and human health.

1.3 AQI Prediction Models

Data on air contaminants from numerous sources are considered when predicting air quality. The AQI is calculated using a variety of machine-learning methods. The standard measure of air pollutant data is represented by the air quality index level for a specific time period. Many factors and air quality data are considered when predicting environmental contamination by air quality. Machine learning approaches include the support vector machine, random forest, decision tree, and linear regression. By predicting air pollution using machine learning algorithms, the air quality index is estimated. The classification of the variables influencing air quality is displayed in the following figure.

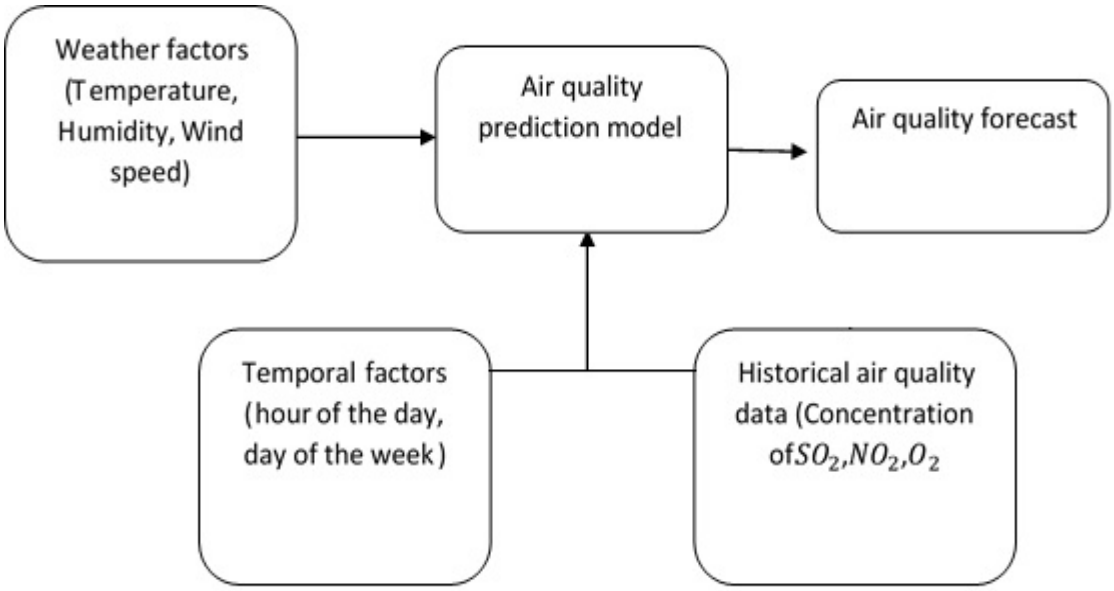

Figure 1.2: Factors of air quality prediction

The features namely, weather factors, temporal factors and historical air data are presented in the figure 1.2 to carry the prediction model. The AQI value aids in the daily and hourly reporting of air data from various locations throughout numerous Indian cities. Based on the index value, pollution affecting health is predicted within a short time. A specific contaminant's average concentration is used to calculate the AQI, which is then used to forecast air quality data. The machine learning approach's regression function is frequently used to forecast the AQI value. Here, the connection between the dependent and independent variables is determined using a linear regression model. Both linear and non-linear regression functions are considered for predicting the AQI of air data. The AQI level based on calculated value predicts air data and is tabulated in the following table.

For effective air pollution prediction, the level of the AQI is shown in Table 1.1. The air quality dataset contains a wide range of factors, such as air quality indexes, City, Date, PM2.5, PM10, NO2, NH3, SO2, OZONE, CO, SO2, NO, Benzene, Toluene and Xylene. All the attributes are considered as the independent variable, whereas the air quality index is designated as a dependent attribute. As a

result, multiple procedures such as feature selection, data pre-processing, and classification are used to predict air pollution.

Table 1.1: Tabulation of AQI level

AQI Level	AQI Range
Good	0-50
Moderate	51-100
Unhealthy	101-150
Unhealthy for strong people	151-200
Hazardous	201+

- **Data pre-processing**

air quality prediction model initially carries data pre-processing to reduce the noise in air data. Here, noisy data and unwanted air data are removed from the dataset by performing various operations. Removing noisy air data helps minimize space storage, attribute subset selection, and dimensionality reduction of the dataset. This contributes to improving the predictability and accuracy of air pollution.

- **Feature selection**

The correlation-based feature selection method is used with pre-processed data to choose important air data features for pollution prediction. In this case, the correlation was determined between the input and object variables. The AQI features are examined and appropriate features are chosen for more accurate pollutant prediction based on the estimated correlation value.

- **Data classification**

In order to anticipate air pollution accurately and with the lowest possible error rate, classification is finally done using the relevant features that have been chosen. The air data is determined and classified based on AQL pollutant levels. It effectively classifies higher air quality and lower air quality data. A data classifier is carried out by varying AQI limits for precise air pollution forecasting with an increased rate and minimum time.

1.4 Motivation

Air comprises numerous data and essential factors that provide living creatures on the earth. The pollution in the air data causes a health issue for urban metropolises and its effects on the environment are the worst. During air pollution forecasting, relevant data feature is essential to obtain better pollution prevention performance. Managing air data in the environment improves human health and obtains efficient air pollution forecasting. Additionally, higher forecasting performance with minimum memory consumption is one of the significant issues to be addressed.

Abdelkader Dairi et al. (2021) developed the Integrated Multiple Directed Attention and Variational Auto Encoder (IMD-VAE), which predicts various air contaminants by utilizing attention techniques and traditional VAE. It forecasts air pollution with efficient computationally complexity accurately. Also, temporal dependencies among non-linear approximations focused on the relevant feature extraction were attained. Moreover, authentication was applied to the prediction of air pollution time series data, both univariate and multivariate. In order to get better outcomes with improved forecasting accuracy, early air pollution prediction using a detection system still needs work.

A bidirectional Recurrent Neural Network (RNN) was designed by D. Saravanan and K. Santhosh Kumar (2021) with temporal factors. Here, the

temporal factor specifies an hourly and daily level of air data that monitors pollution with minimized error. The developed neural network technology monitors air quality data to forecast pollutants temporally. The bidirectional RNN controls current and historical input for air pollution quality monitoring. Pollution forecasting is enhanced with higher accuracy by managing air quality data but fails to reduce error.

Mauro Castelli et al. (2020) developed a unique machine-learning method called Support Vector Regression (SVR) to predict air pollution. The air quality index value is predicted by considering the levels of pollutants and particulates in the air. Here, regression is performed using the radial basis function (RBF) as a kernel function to produce more accurate air pollution forecasts. The air dataset presents numerous data features, and significant features are selected for forecasting using Principal Component Analysis. It accurately predicts hourly pollutant air data with higher consumption of memory space.

A method for monitoring indoor air quality was developed by Wen-Tsai Sung and Sung-Jung Hsiao (2021) in order to forecast pollution and improve the quality of the air for living things. Air quality indices, carbon dioxide indices, and an AQI value are used to monitor indoor air quality with the help of the Internet of Things. Air quality is monitored and analyzed considering environmental safety as well as interior and outdoor characteristics. In addition, fuzzy control algorithm was designed to achieve an air quality monitoring system to forecast air pollution. However, it could have decreased the time to forecast environmental air pollution.

Air Quality Monitoring System (AQMS) is introduced by Koel Datta Purkayastha et al. (2021) on a cloud platform for storing and processing air data information. The considered data information from data sources is transmitted and received using sensors included with a microcontroller. Temperature, relative humidity, CO2, CO, NO2, and other characteristics that store value in the database

are monitored by the monitoring system. Thus, monitored features are utilized to forecast air pollution with increased accuracy and minimum time. However, the performance-accurate pollution forecasting metrics still needed to be estimated.

Machine Learning algorithm develops a Non-linear Auto Regression with eXogenous by Ahmed Samy Moursi et al. (2021). The regression function is combined with an IoT-enabled system to monitor air pollution effectively. With the estimation of mean square error, normalized error, and coefficient of data, air pollution prediction is arraigned. But the system should have included or changed air data for pollution prediction with reduced error occurrences.

K. Krishna Rani Samal et al. (2021) introduced the Convolutional Neural Network-Long Short-Term Memory, Sparse Denoising Autoencoder (CNN–LSTM–SDAE) (CLS) model to improve air pollution prediction. At first, the missed value of air quality data is determined by applying the K-nearest neighboring algorithm. Next, significant features are identified using CNN-LSTM and reconstructed for removing noisy air data to obtain enhanced pollution prediction. However, it was unsuccessful in enhancing the multivariate performance of forecasting air pollutant data in the environment.

A hybrid intelligent model based on the long short-term memory and multiverse optimization algorithm processes was created by Azim Heydari et al. (2022). The air quality dataset includes a wide range of air data elements. From that, air data features are predicted for pollution forecasting through the LSTM model. After that, the MVO algorithm minimizes errors while forecasting air pollution. Thus, it helps to forecast air pollution with reduced error effectively and fails to attain accurate prediction of air pollution.

Gao Huang et al. (2021) presented an advanced deep convolutional network for large-scale air pollution prediction. Air quality is tracked to precisely predict air pollution in the shortest amount of time through the dissemination of air data. The

observation points of air data are considered to predict air quality from various locations at various intervals. However, the convolutional network was unable to assess some parameters in order to improve the forecast accuracy of pollution.

A deep learning solution termed CNN-LSTM was performed by Abdellatif Bekkar et al. (2021) to predict air quality concentration. pre-processing is offered for each input air quality data set to fill in the missing value of the air data for more accurate pollution forecasts. The geographical and temporal features are then determined by estimating the correlation coefficient between the data features. Air data is efficiently classified using chosen features to anticipate air pollution data with the least amount of error. Accurate prediction of air data concentrations is essential to shield populations from the damaging impacts of air pollution earlier.

The combined weight forecasting model (CWFM) was projected by Bingchun Liu et al. (2021) to perform an accurate prediction of air pollution. Initially, discrete wavelet transform is applied on each input data to eliminate noise and reduce the dimensionality of data. With wavelet decomposition results of air data, the neural network is carried out for pollution prediction results. Afterward, weight assignment is applied to combine weighted results for accurate prediction. However, accurate pollution evaluation still needs to be developed due to time and individual capacity constraints.

Deep-learning construction was portrayed by Philipp Hahnel et al. (2020) to monitor and predict air pollution with reduced run-time. Here, a different model is designed to forecast pollution from air data with two orders of magnitude. They are such deep-learning and domain-decomposition techniques. This technique supports forecasting air pollutant data with higher accuracy. However, the time complexity and space utilization are still to be identified.

The machine learning predictive method was carried out by Doreswamy et al. (2020) to predict atmospheric air awareness and reduce mean square error. Each input air data collection undergoes data pre-processing in order to locate

missing data from the dataset. With pre-processed data, machine learning is utilized to categorize air data for forecasting air pollution at an enhanced rate. However, data pre-processing of the dataset was not achieved with minimized time consumption.

With the combination of CNN and LSTM construction, Deep-AIR was developed by Qi Zhang et al. (2022). Using domain-specific features to gather spatio-temporal information, the Deep-AIR design bridges the gap for achieving the city-wide pollution evaluation. Convolution layers were also employed to enhance the acquisition of spatial and temporal relationship. Consequently, the prediction accuracy of air pollution was enhanced it, but failed to reduce time consumption.

Baowei Wang et al. (2019) created a two-layer model prediction algorithm based on a gated recurrent unit and a short-term memory neural network. The double-layer network that has been established aids in accurate air pollution prediction. Here, the IoT technique is considered using a neural network to monitor air data. It predicts pollutant air data with increased accuracy, and memory utilization for storing air data is higher in range.

Shengdong Du et al. (2021) presented a hybrid neural learning system for early air pollution forecasting. For hybrid neural networks, one-dimensional CNN and bi-directional LSTM networks are used. Based on network structure, significant features are extracted with spatial correlation and temporal dependencies. It helps to determine multivariate air quality data with reduced time series. However, the goal of making an accurate prediction in less time was not attained.

Suspended Particulate Matter Modeling was performed by Ekta Sharma et al. (2020) to predict air pollutant data from the environment to prevent public health issues. It presents a convolutional neural network and a deep learning hybrid

CLSTM model. A data processor built into the neural network extracts pertinent elements for hourly data forecasting with total suspended particle matter. Thus, the hybrid machine learning algorithm was presented to forecast pollution with better accuracy. However, the time utilization of air pollution prediction remained relatively high.

Raquel Espinosa et al. (2021) presented a time series forecasting model to improve air quality prediction for pollution forecasting. Sliding windows and a multistep forecasting procedure are used in the forecasting process, which is based on deep learning and machine learning models. A sliding window is applied to each input air data to eliminate noisy data. Then, multistep forecasting is carried out to perform pollutant prediction on air. A temporal analysis was to learn early prediction models to recognize air pollutant data at risk. However, higher prediction accuracy still needed to be attained.

With the use of a deep learning strategy, Yue-Shan Chang et al. (2020) presented an Aggregated LSTM model (ALSTM). The designed LSTM model monitors air pollutant data hourly at various stations near industrial areas and external pollution sources. By considering pollution data, air quality is predicted effectively with minimized complexity. However, the system should have included or transformed data to prevent air pollutant data from the dataset.

Yu Huang et al. (2021) implemented the spatial attention embedded RNN to forecast air quality prediction using Spatiotemporal dynamic associations. However, it was unsuccessful in reducing the error and complexity of the neural network algorithm. Several deep and machine learning techniques are developed but it is complex to predict accurate pollutants for forecasting the air quality index.

1.5 Problem Statement

Air pollution prediction is a significant aspect that helps identify pollutant air data and health issues on living organisms. Various machine learning techniques

were designed with an enhanced result of air pollution forecasting. Classification is classifying the air quality data for predicting pollutant data. However, the existing classification techniques could have improved the forecasting rate using lesser time. Many research works have been designed to attain efficient and higher forecasting accuracy with a minimum error rate. Due to the presence of false data, memory utilized for forecasting air pollutant data is higher. However, the algorithm failed to minimize memory consumption. To solve these problems and increase the precision of air pollution predictions, a novel recommended approach is applied. The Problem Statement is thus formulated as,

"Design and implement a hybrid machine learning-based forecasting model that leverages multi-modal data to provide real-time predictions of various air pollutants with spatial and temporal features."

1.6 Research Objectives

In order to find the solution to the above problem statement, the primary objectives of the research work are to improve the performance of air pollution forecasting accuracy, time and error rate in terms of multiple number of features with time and space complexity. Hence the developed research work focuses on designing and implementing a machine-learning algorithm for air pollution forecasting. To increase the performance of air pollution prediction, significant relevant features of air data are selected from the dataset. The specific objectives framed are,

- To Develop Linear Regression and Multiclass Support Vector (LR-MSV) Model,
- To Develop Bilateral Transformative Broken-Stick Regression-based Quadratic Weighted Emphasis Boost Classification (BTBSR-QWEBC) Model, and

- To Develop Discretized Regression and Least Square Support Vector (DR-LSSV) Model.

1.7 Research Contribution

The research methodology of different proposed techniques is developed for improving forecasting accuracy on air pollution with minimum error rate and time. The key contribution of the suggested models is to forecast data on air pollution by choosing pertinent variables and categorizing the data into diverse groups. As a consequence, the contributions of multiple proposed research models are outlined below.

The LR-MSV model was initially created to improve the predictability of air pollution. Prior to processing the data, each air quality data goes through pre-processing for noise removal. The input data is pre-processed using the WSW (WSW) method, which helps shorten the time it takes to classify air data. Following data pre-processing, LRC based feature selection are used to carry out the feature selection procedure. In that, Loss function and linear regression are used to choose the features. It helps to increase forecasting accuracy by choosing the most pertinent air data features with the least amount of loss. Ultimately, the input air quality data is classified using a multiclass support vector model based on the AQI value as a guide. The air pollution forecast is then executed with the least amount of time and error rate possible, depending on the results of data categorization.

The BTBSR-QWEBC model is proposed next to forecast air pollution with improved accuracy and less memory utilization. The number of air data is taken into consideration for forecasting from the air quality dataset. Bilateral Discretized Z-wavelet Transform (BDZWT) is used for each input air data to remove noise and produce pre-processed data. With obtained pre-processed air data, OIBR is performed to select relevant features. The selected significant features help to attain

accurate pollution forecasting with minimum memory consumption. Then, data classification is performed using ensemble classification named as QWEBC technique. A Kernelized support vector is presented to classify air quality data with minimized quadratic error. Therefore, the BTBSR-QWEBC technique attains improved performance in air pollution forecasting with minimum memory utilization.

Lastly, in order to achieve precise air pollution forecasting, a unique DR-LSSV model is presented. The input layer receives the air quality data that has been collected from the dataset. After that, discretized Hartley transformation-based data pre-processing is performed at the first hidden layer. It transforms real input into actual output by removing noise air data. With obtained pre-processed data, feature selection is performed at the second hidden layer using a CMLLR function. Then, classification is performed with selected relevant features at the third hidden layer with the CCLSSV process. Then, the concordance correlative coefficient value is estimated to classify data into various classes to forecast air pollutant data. Finally, classified data is presented at the output layer for forecasting air pollution with enhanced accuracy and minimum error.

1.8 Research Methodology

Proposed research work performed three different Enhanced Models; each model is integrated with each other by means of some enhancements made with initial model. The models are,

- LR-MSV model,
- BTBSR-QWEBC model, and
- DR-LSSV model.

The suggested models are contrasted with the outcomes obtained from the other most recent methods by using the metrics, Accuracy, Error Rate, Time Consumption and Memory Consumption. Finally, the Proposed Models are

compared within the three identified results and then the best technique for air pollution forecasting will be identified.

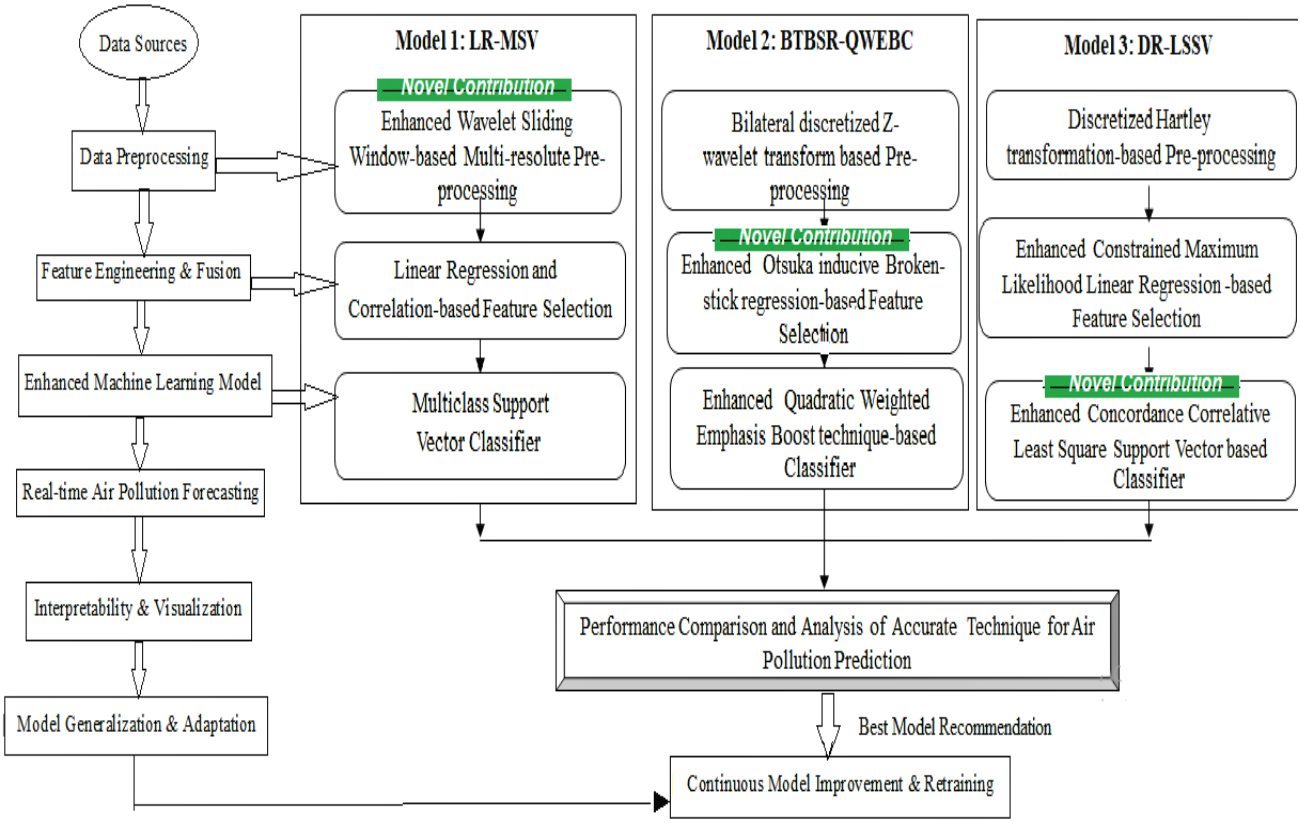

Figure 1.3: Overall methodology

1.9 Dataset Description

https://www.kaggle.com/ rohanrao/ air-quality-data-in-India is the source of the Air Quality India dataset. The Central Pollution Control Board has released the statistics to the public at https://cpcb.nic.in/, the official Government of India website. The dataset under consideration includes 16, distinct features from the year and 2,18,640 air quality data points. The Air Quality Index value is computed hourly and daily for many stations located in various Indian cities. The taken into consideration Air Quality India dataset contains the measured AQI and air quality information. The 26 distinct cities from which the dataset was gathered. These include the following cities: Ahmedabad, Aizawl, Amaravati, Amritsar, Bengaluru,

Bhopal, Brajrajnagar, Chandigarh, Chennai, Coimbatore, Delhi, Ernakulam, Gurugram, Guwahati, Hyderabad, Jaipur, Jorapokhar, Kochi, Kolkata, Lucknow, Mumbai, Patna, Shillong, Talcher, Thiruvananthapuram and Visakhapatnam.

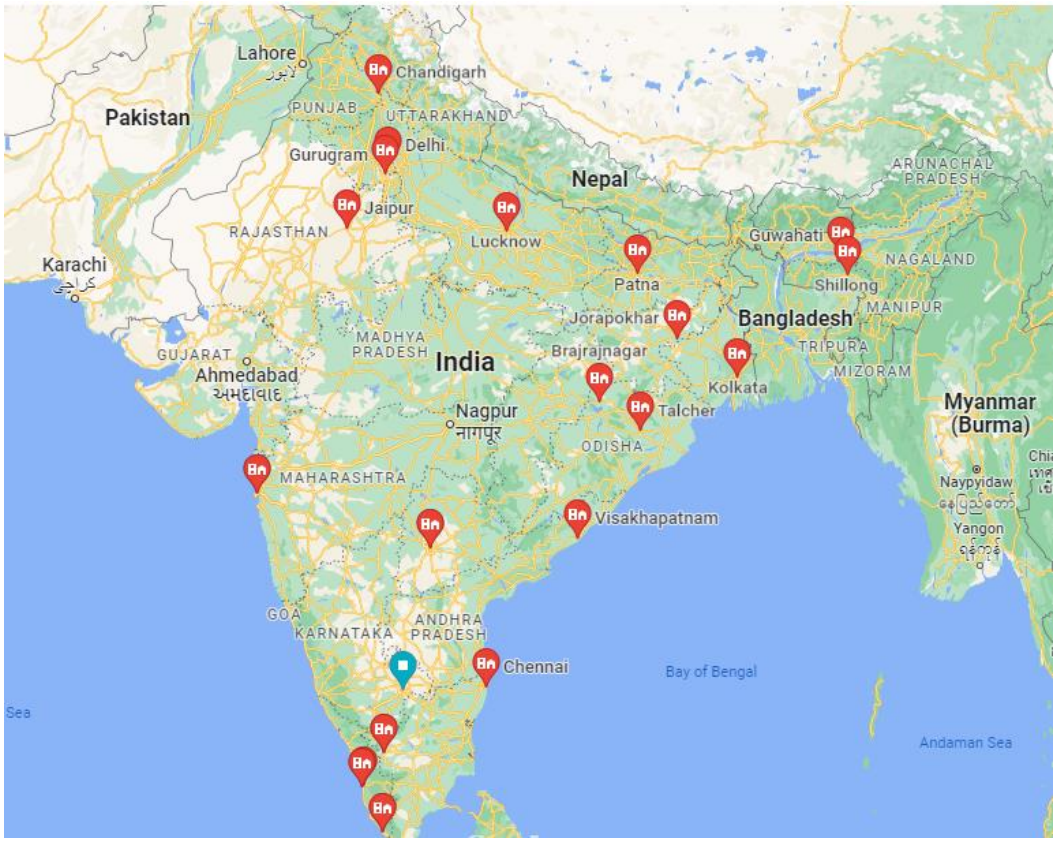

Figure 1.4: Geographical representation for data collected cities

1.10 Organization of Thesis

The thesis is organized in the described format below:

Chapter 1 presents the air pollution and air pollutant data quality introduction. Additionally, an IoT network is presented to select optimal relevant features of air data for achieving enhanced air pollution forecasting in a cloud environment.

Chapter 2 describes the prior work done in air pollution forecasting using ML techniques in order to gather data for forecasting process. There is also discussion of the benefits and drawbacks of monitoring and predicting air pollution quality.

Chapter 3 describes the proposed LR-MSV Air Pollution Prediction model with air quality data. It effectively forecasts air pollution with higher accuracy and with minimum time.

Chapter 4 explains the BTBSR-QWEBC model to attain efficient pollution air prediction with advanced forecasting accuracy and minimized memory consumption.

Chapter 5 portrays DR-LSSV model to address air pollution prediction with IoT networks to improve forecasting accuracy on air polluted data with reduced error rate.

Chapter 6 represents the observations of the proposed methodologies, such as LR-MSV, BTBSR-QWEBC, and DR-LSSV model, with metrics for air pollution forecasting time, error rate, and memory usage.

Chapter 7 describes the overall key outcomes, problems faced in the proposed research work, and the future direction of the proposed techniques.

1.11 Summary

The primary framework for predicting air pollutant data in a cloud environment has been discussed in this chapter. The motivation for research work and objectives are explained briefly. This chapter also includes descriptions of the problem definition, study objectives, contributions of developed proposed models, research methodology, dataset description, and thesis arrangement.