Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Air Quality Analysis of Tamil Nadu State Using Advanced Artificial Intelligence Algorithms

P. Aruna Rani[1], Dr. V. Sampathkumar[2]

[1]Research Scholar, Department of Civil Engineering, Sathyabama Institute of Science and Technology, India, arunaranip76@gmail.com
[2]Professor, Department of Civil Engineering, Sathyabama Institute of Science and Technology, India, svsjpr@gmail.com

Around 90% of the world's population does not get pure air to breathe because of pollution. Air pollution is ranked as South Asia's 6th most fatal killer, causing premature death due to heart attack, stroke, and respiratory diseases. Research conducted so far on determining the pollutant levels in the air has not given satisfactory results due to the complexities and unpredictable nature of climate and industrial emissions. So, understanding the existing issues and challenges for a better solution is paramount. Most of the pollution that threatens the world today is artificial. Pollutants mostly travel through air and water, impacting human lives in various ways. Water pollution can be controlled by adopting preventive measures, whereas limiting air pollution is challenging as it keeps increasing with the increase in chemical factories, automobiles, locomotives, etc. Few recent research works have stated that machine learning data analytics can help predict uncertain data accurately. Thus, this paper aims to implement a Multi-class Support Vector Machine algorithm to accurately predict the air's pollutants level. The implementation is carried out in Python, the prediction results are compared with the other machine learning algorithms, and the performance is evaluated. The comparison shows that the Multi-class Support Vector Machine algorithms outperform the other approaches regarding prediction accuracy.

Keywords: Forecasting Model, Air Quality Examination, Machine Learning Algorithm, Multi-class Support Vector Machine Algorithm, Pollutant Level Estimation.

1. Introduction

Half a century before, people breathed pure and fresh air. But, the drastically growing industries and industrialization emit poisonous gases into the air. So, the air becomes more toxic daily and causes more respiratory diseases. Air pollution contaminates natural environments and surroundings with chemical and biological agents that affect the atmosphere's behavior. Pollution is created in various forms like heat, air, water, etc. The major forms of environmental pollution affecting human health are water and air. Several remedies have been applied for controlling and preventing water pollution. Still, it is not possible in the air since the pollution range increases continuously from various chemical industries, automobiles, locomotives, etc. But some of the earlier research works obtained estimated the level of pollutants.

The World Health Organization (WHO) estimates that air pollution caused 3.7 million early mortalities worldwide in 2012, making it a serious health issue associated with atmospheric pollution [15]. Any substance that alters the environment's natural properties, whether it is chemical, physical, or biological, is considered an air pollutant. Air pollution can occur indoors or outdoors. Frequent air pollution causes include automobile emissions, industrial waste processes, home appliances, and forest fires. Particulate matter, carbon monoxide, ozone, nitrogen dioxide, and sulphur dioxide are among the pollutants that constitute the greatest threat to human health. Interior and exterior air pollution are significant contributors to illness and death. It is known to cause pulmonary as well as other health illnesses. Generally, the climate and biodiversity of the globe are highly related to air quality. Burning fossil fuels is one of the main causes of air pollution and ultimately contributes to greenhouse gas emissions.

Air pollution is a hidden predator on an international basis. India has some of the worst air pollution globally, which poses a serious danger to the welfare and prosperity of our nation. India's outdoor and indoor air pollution will contribute to 1.7 million fatalities in 2019 [16]. Pollution costs the economy heavily in terms of its effects on human health. The primary causes of air pollution, regrettably, have been well-identified and are the same in all Indian cities: automotive emission, large-scale production in huge companies, minuscule businesses like brick kilns, trapped dust on the pavement due to traffic and building construction works, waste burning, burning of different fuels for cooking, lighting and heating and sedentary power production using diesel generator machines.

Moreover, the seasonal climate changes resulting from dust storms, open field fires during harvesting, and sea salt in coastal areas are major causes of air pollution. Ultimately, the other major contributors to air pollution are diesel, petrol, gas, coal, and other waste dust. As humans, we need some basic requirements for leading a healthy life. It includes air, food, shelter, and water; without one of these needs, humans will suffer a lot and also cannot survive. Therefore, air pollution is a major threat that adversely affects human life. Fresh air is an elixir of healthy life that may also heal mild problems present in humans. Even a mild drop in the purity level of the air may increase the percentage of illness in persons already affected with asthma, lung diseases, and other breathing issues, which may, in rare cases, become fatal. A further increase in the pollution level can cause illness in otherwise healthy persons. Measuring the quality of air becomes vital to reduce the mortality rate and reduce health-related issues as an indicator of how pure or dirty the air is its quality.

Air Quality Index (AQI) is used to measure air quality daily. With the help of AQI, one can measure the purity of the air surrounding us. The measurement of AQI is also related to health issues, and the gravity of health risk depends upon the inhalation period of impure air. The air quality measurement and its components or pollutants are shown in Figure-1.

1.1 Contribution of the Paper

The paper's main objective is to analyse the air quality data and predict the pollution level to assess human health risks. The paper's novelty is that the artificial intelligence algorithm is trained with three different datasets and combines the output of the trained data to improve the testing accuracy. The level of pollutants mixed in the air is mapped with an index value called Air Quality Index (AQI), which is used to predict the human health risk level. It helps the government or non-government organizations to control air pollution and save the people. To do that, this paper contributes,

1. A detailed study has been carried out to create the air quality index table based on the pollutants' level.

2. Create a multi-class support vector machine and convolutional neural network model by training with different periods of data and labeling them.

3. Label the data by creating an LSTM model by training with current time-series air quality data.

4. Combine the trained data obtained from MCSVM, CNN, and LSTM, fine-tune it and provide fine-grained trained data.

5. All the models are tested using fine-grained and test data, and their accuracy is verified.

1.2 Review of Literature

Understanding the issues and challenges met by the earlier research works is essential, and obtaining a new methodology to overcome the same. Hence, this section carries out a detailed literature survey. For example, Anikender Kumar et al. (2011) had proposed the Principle Component Regression (PCR) method for forecasting the air pollution level for Delhi in India. The PCR method follows the principles of the Multiple-Linear-Regression (MLR) approach. Initially, the air quality data is analyzed, the pollutants are extracted, AQI was created for the data, and the air quality is forecasted. The dataset used in the experiment was taken for various climate seasons. Huixiang Liu et al. (2019) had presented a study for analyzing and predicting the air pollution level over the data taken from two cities: Beijing and Italian. Two different machine learning algorithms, such as the SVM and RFR methods, were used for computing the AQI value and concentration of NOx, respectively. The experimental results indicate that the SVM method performed better in estimating the AQI value and the RFR method was better for predicting NOx concentration. Ziyue Guan and Richard (2018) proposed various machine learning algorithms for predicting the PM2.5 concentration. The input data were collected from Melbourne's Environment Protection Agency (EPA) official website. Different ML algorithms like ANN, LSTM, RNN, and L.R. were used in the experiment. LSTM had performed better for predicting the PM2.5 concentration with better accuracy. Heidarmaleki et al. (2019) had implemented the ANN technique to predict the ambient air pollutants NO_2, PM_{10}, SO_2, PM2.5, O_3, and CO_2 in Iran's most pollutant city. The AQI and AQHI value of the city mentioned above was calculated using the ANN technique. Metrological parameters, such as air pollutants, time and date were considered the model's input parameters. The proposed model had produced efficient prediction results.

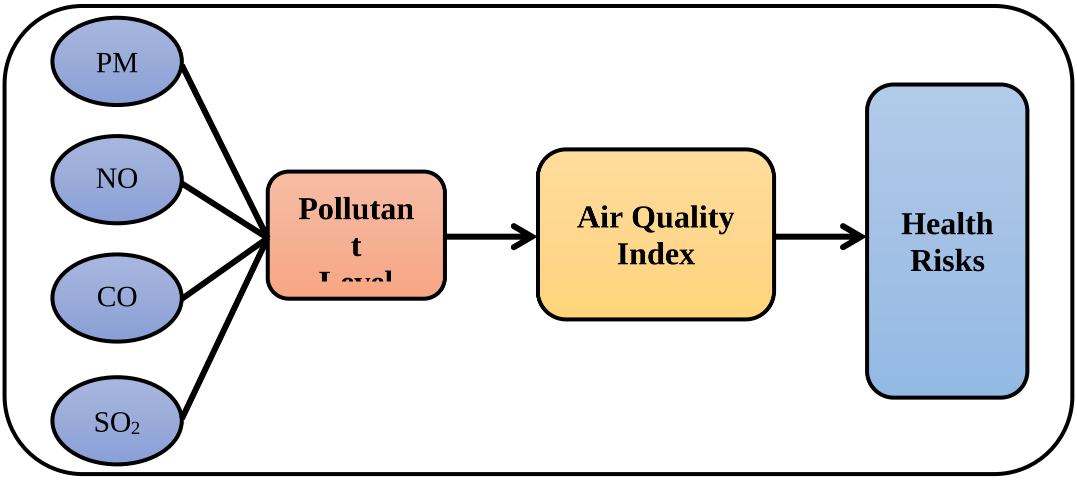

Figure-1. Pollutant and Air Quality Measurement

C.R. Aditya et al. (2018) employed a machine learning algorithm to detect and forecast the PM2.5 concentration in the atmospheric data. At first, the air pollutants' level was calculated using the Logistic regression method. Then, the autoregression algorithm was implemented to predict the future value of PM2.5 concentration. The prediction was accurate since the estimated values were mapped with the previous records and statements. Nidhi Sharma et al. (2018) conducted a study to observe the air pollution level in Delhi, India, from 2009 to 2017. It also predicted the level of SO_2, NO_2, PM, CO, and benzene in the air. The result of the study showed that NO_2 increased to 16.77$\mu g/m^3$, ozone level increased by 6.11 mg/m^3, benzene reduced to 1.33 mg/m^3, and SO_2 increased to 1.24$\mu g/m^3$. Mohamed Shakir and Rakesh (2018) analyzed the level of various air pollutants like NO, NO_2, SO_2, CO, PM10, and PM2.5, along with the time, date, and environmental changes, using the WEKA tool. The WEKA tool used ZeroRalgorithm in the proposed model. The result of the study shows that the air pollutant level during the peak hours of the workingdays has increased significantly. During weekends the pollutants were less. The K-means clustering algorithm measures the environmental changes and the air pollutant level. KazemNaddaf et al. (2012) proposed Airq software to analyze the air pollutant level in Iran's most polluted city. The study showed that PM10 pollutants caused many side effects and sometimes led to death compared to other contaminants.

Yuseofmidikhaniabadi et al. (2016) presented a study to define the relationship between cardiovascular diseases and air pollutantslevel in the air. Using Ariqsoftware was used to predict the pollutants level. The analysis showed that if the air pollutant level increases, the mortality level also increases. The pollutant level of PM10, NO_2, and O_3 is increased to 1.066, 1.012 and 1.020, respectively. R. Gunasekaran et al. (2012) conducted a study to analyze the air pollution level; the data was taken from Salem district. The result of the study showed that the pollutant level around the area was average. But the annual PM10 level was slightly higher than the national standard. S.Tikhe shruti et al. (2013) proposed ANN and G.P. methods for predicting the future air pollutant level in Maharashtra. Archontoula Chaloukou et al. (2003) implemented ANN and MLR techniques to find the PM10 concentration in the air. The methods were applied to extensive data. The input data were divided into training, testing, and validation. The results of these two algorithms were compared and verified, proving that the ANN was the efficient method for predicting the PM10 concentration.

From the above survey, it has been identified that most of the earlier research methods have used manual methods like active and passive sampling. Microsensors have been used for detecting air pollution and pollutant levels. But the measurement got failed to detect the pollution and pollutant levels accurately. The earlier research works have not accurately obtained the individual pollutants bloating the overall air pollution over some time. These limitations have kindled the interest in implementing an automatic prediction and forecasting model for air pollution and pollutant levels. Also, this paper motivates to correlate the human health risks based on the air pollution level.

1.3 Problem Statement

Recent times have seen countries focusing on the health and welfare of their citizens. Thanks to the past two years, the pandemic has baffled the world. Air pollution tops the list of several causes that prove detrimental to people's health worldwide. So, monitoring and checking the pollution and pollutant level have become all the more important in the present. It is essential to understand the creation of the air quality index to monitor and measure pollution and pollutant levels.

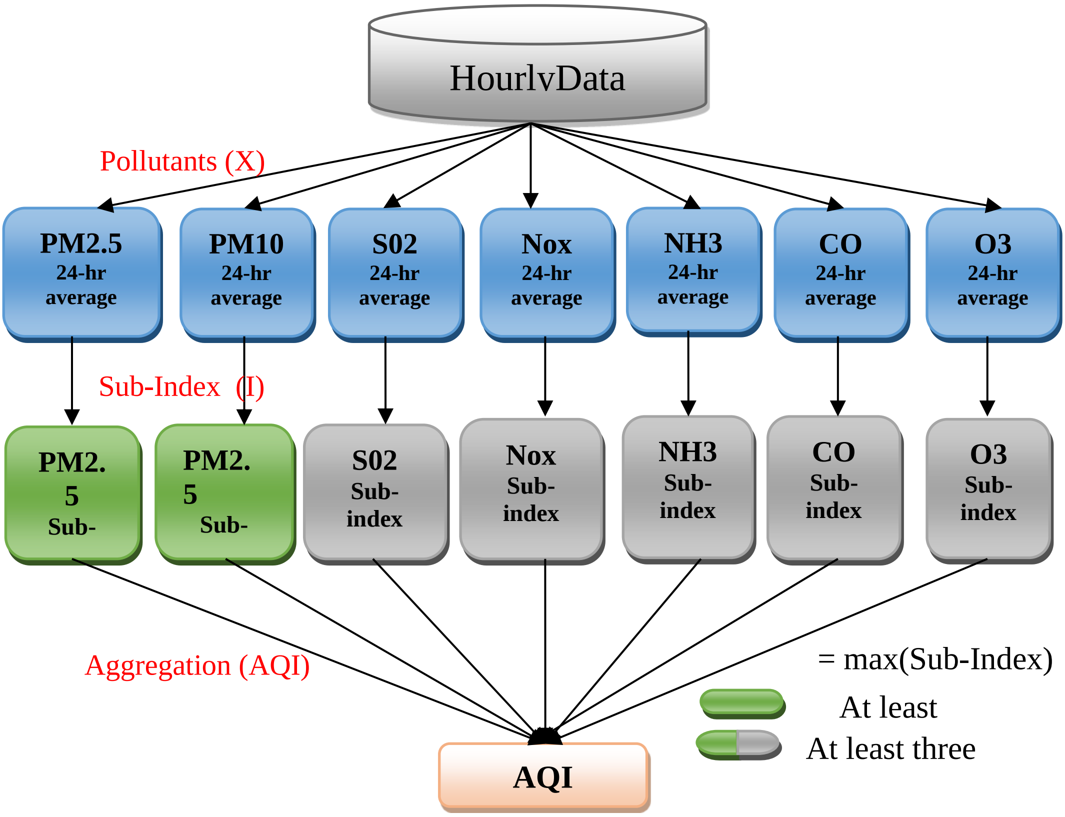

Figure-2. Air Quality Index Calculation

The Air Quality index is a tool that government and private agencies use to educate the people on the standard of the air quality and the correcting measures to be taken depending on it. An increase in the health risk of people is directly proportional to the increase in the AQI. In order to create awareness among the public, this is given as a color, terminology, or index value. The AQI calculation using pollutants concentration is illustrated in Figure-2. The following step of the procedure explains the way of AQI calculations.

1. Each pollutant's concentration is recorded for 24 hours, and their average value is used for calculating the sub-indices and health risks. The worst index is considered the AQI of the particular location.

2. It is not necessary to monitor all the pollutants at a single location. AQI can be calculated using at least three pollutants such as SO_2, CO, and PM levels is sufficient. Similarly, the sub-index is calculated from the data collected from 16 hours of monitoring.

3. Even if the data is inadequate for calculating AQI, the sub-indices for the pollutants are calculated. Each pollutants' sub-index also provides AQI status for their pollutant.

4. One web-based automatic tool is provided to calculate AQI based on the real-time values.

5. AQI can also be obtained by feeding the data manually to the AQI calculator.

Table-1. Air Quality Index For Gaseous Pollutants

AQI Class	Range	SO_2 (24 hrs)
Good	0 – 50	0 – 40
Normal	51 – 100	41 – 80
Moderate	101 – 200	81 – 380
Poor	201 – 300	381 – 800
Very Poor	301 – 400	801 – 1600
Severe	401 - 500	> 1600

Table-2. Air Quality Index Versus Pollutants' Range

AQI Class	Good	Normal	Moderate	Poor	Very Poor	Severe
Range	0 – 50	51 – 100	101 – 200	201 – 300	301 – 400	401 - 500
SO_2 (24 hrs)	0 – 40	41 – 80	81 – 380	381 – 800	801 – 1600	> 1600

The AQI is mapped with the pollutant range (P.R.) to identify the health risk value (HRV). Each pollutant is related to a human health issue. Characterizing the consequences of human exposure to toxic agents is known as human HRA. HRA is usually estimable, which can be accomplished by prediction with previous exposure data. It is widely used to monitor human health regarding exposure to pollutants. The risk assessment based on the AQI value is given in Table-3. Various gases are mixed in the air contents. But only a few of them cause air pollution. This paper mainly focused on predicting SO_2 and calculating the AQI value to determine human health risks to increase efficiency.

1.4 Sulphur Dioxide (SO_2)

Multiple pollutants are present in the air, but only a few cause air pollution. Among them, SO_2 is a dangerous gaseous pollutant that needs to be reduced to reduce the risks of air pollution. Hence, this paper mainly focused on analyzing the air quality data and predicting the SO_2 levels present in the air. SO_2 is one of the significant gaseous pollutants that change the fresh air as pollution air. The combustion of sulfur and fossil fuels releases SO_2 into the atmosphere.

Table-3. Air Quality Index Versus Health Risks

Good (0-50)	Not affecting health
Satisfactory (51-100)	It affects sensitive people like those who have lung problems.
Moderate (101-200)	Breathing uneasiness to people with lung and heart diseases, children, and other adults
Poor (201-300)	Unhealthy people cannot breathe well. It affects healthy people slightly.
Very Poor (101-400)	Increase the severity level to hazardous for unhealthy people, and affects healthy persons too.
Severe (>401)	It affects healthy people immediately and unhealthy people fatal.

Fuels containing Sulphur are burned in the production of coal, metal mining and processing, ship turbines, and massive diesel machinery.The pollutant, SO_2, presented in this study for analysis is in the form of gas. The characteristics of SO_2 are no color, nasty and sharp smell. It can be easily combined with other chemicals and produce harmful substances. One among them is SO_2 which affects the health when it is breathed. It is one of the significant causes of coughing, shortness of breathing, wheezing, and chest problems. It also irritates the nose and throat of humans. SO_2 levels in the environment can affect how suitable it is for animal and plant life. SO_2 exposure is linked to a rise in pulmonary problems and illnesses, respiratory disorders, and early mortality. Acid rain, proven to cause deforestation, is primarily composed of sulfuric acid, created when SO2 reacts with water.

This paper extends the implementation of SVM for modeling the MSVM. Compared to other traditional and machine learning algorithms, SVM performs well over the dataset having an understandable margin among the classes. Its' classification accuracy is high in a high-dimensional dataset. Also, it has more efficient in instances when the dimension increases than the specimens. SVM works well on the data even if it does not have an idea. It can process unstructured data like numeric, alpha-numeric, images, and signals. The logical functionality of its kernel is the major strength of the SVM. So, SVM can solve any multifaceted problems. SVM can perform faster with high accuracy than the Naïve Bayes algorithm. It can increase the number of hyperplanes according to the number of classes and the dimensionality of the data.

2. Support Vector Machine

One of the machine learning algorithms connected with the supervised learning algorithms is SVM. It is primarily used to examine the structured data involved in the process and do classification plus regression assessments [17, 18]. The Support Vector Regression (SVR) a kind of SVM mainly used to solve regression problems, was hypothesized by Vapnik and his co-workers [19]. The training data in the SVR constitutes the predictor variables as well as examined output values. The ultimate aim of this method is to determine the function $\boldsymbol{f(x)}$ which departs from $\boldsymbol{y_n}$. The term $\boldsymbol{y_n}$ denotes the output labels, and its deviations are performed by a value never more than $\boldsymbol{\varepsilon}$, which denotes the bias for every training location $\boldsymbol{x}$. The location becomes flat. Hence, the method SVR is usually termed tube regression. Figure-3 shows the schematic representation of SVR. Based on the literature done on SVM linear regression [19] the response of SVR is denoted by:

$$\boldsymbol{f(x) = \sum_{n=1}^{N} (a_n - a_n^*)\,(x_n^T x) + b}$$

The term $\boldsymbol{x}$ in the above equation denotes the input feature vector and $\boldsymbol{b}$ distance parameter. The term $\boldsymbol{a_n}$ and $\boldsymbol{a_n^*}$, denotes the established Lagrange multipliers. Certain regression issues could hardly be explained frequently with the linear model. In those circumstances, one may get a nonlinear model of SVR using the nonlinear kernel function $\boldsymbol{K(x_1, x_2)} =< \gamma\varphi(\boldsymbol{x_1}), \boldsymbol{\varphi(x_2)} >$which is the replacement of the dot product $\boldsymbol{x_n^T x}$. The term $\boldsymbol{\varphi(x)}$denotes the transformation which maps $\boldsymbol{x}$ into a greater dimensional area. Hence, the ultimate output for a nonlinear SVR issue could be determined by:

$$\boldsymbol{f(x) = \sum_{n=1}^{N} (a_n - a_n^*)\,K(x, x_n) + b}$$

In the SVR and SVM, a single hyperplane is created for classifying the overall data based on the $\boldsymbol{b}$ and $\boldsymbol{\varepsilon}$. The overall data is compared and mapped with the hyperplane by changing the $\boldsymbol{b}$ and $\boldsymbol{\varepsilon}$. Since the SVM uses a single hyperplane, it can provide only two different classes from the entire data as +ve and -ve (true or false).

$$\boldsymbol{y_i = \{x_i * \ \varepsilon_i + b\} \forall\ i = -n\ to + n}$$

Where $\boldsymbol{n}$ is defined as the target margin, the classifier is forced to use more hyperplanes in MSVM to increase the number of output classes. The kernel function of the SVM can be expressed as:

$$\boldsymbol{K(\overline{x}) = 1, \qquad if\ \ \|\overline{x}\| \leq 1}$$

$$\boldsymbol{K(\overline{x}) = 0, \qquad otherwise}$$

To understand its functionality easily.

2.1 Multi-class Support Vector Machine

The problem is divided into sub-problems and solved using MSVM by adapting the principle of SVM. Instead of using a single hyperplane in SVM, the MSVM uses multiple hyperplanes (Figure-4) according to the number of classes required to be classified on the data.

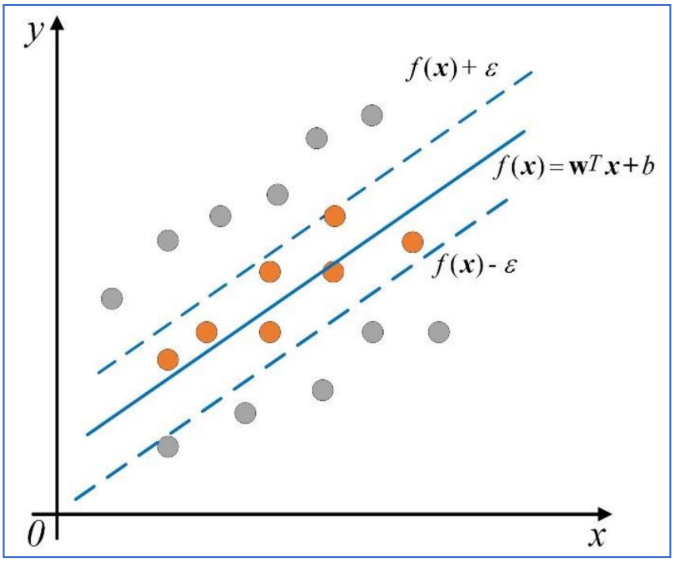

Figure-3. Support Vector Regression Analysis

Generally, SVM is a binary classifier that provides only two classes in the output. The structure of MSVM is obtained by stacking layers of SVM. The response of SVM in multi-class is denoted by

$$\boldsymbol{f(x) = \sum_{n=-N}^{n=+N} (a_n - a_n^*)\,(x_n^T x) + b}$$

The structure of the MSVM is illustrated in Figure-2.

2.2 MSVM Model Implementation

The proposed MSVM is developed by training the model. It learns a significant portion of the data and the object of the model created during the training process. It is well known that the training data has correct answers called target attributes. The learning model searches the required pattern in the training data to map the input data's attributes to the target attributes. Finally, it results in a trained MSVM model, which can capture the patterns. Here the MSVM obtains the SO_2 pattern for estimating the air pollution.

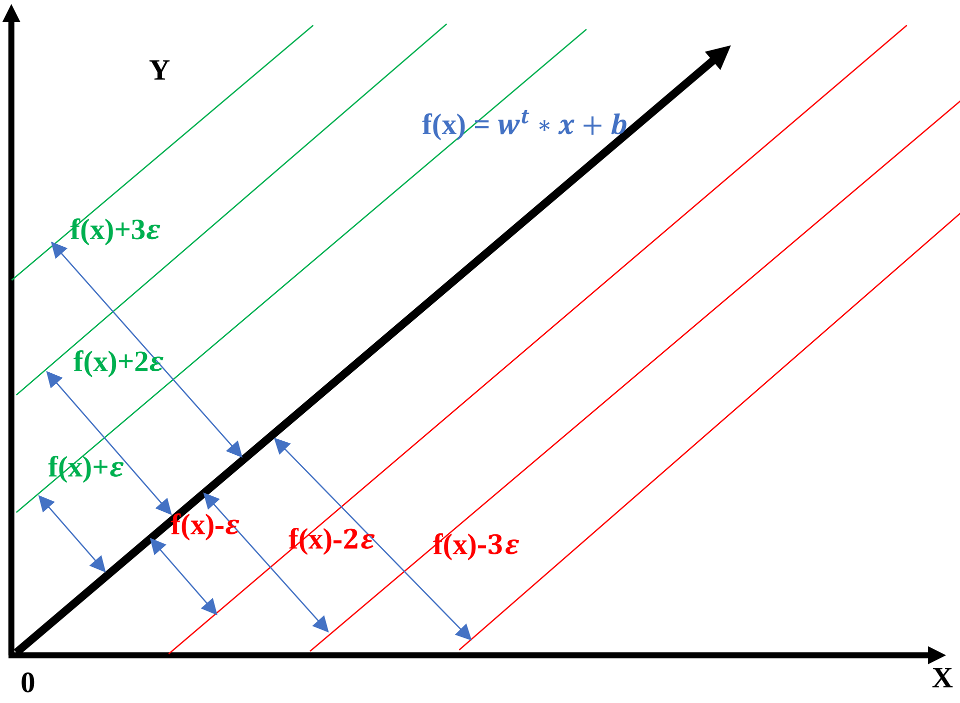

Figure-4. Multi-class Support Vector Machine

After successful training, the MSVM model is used to predict similar patterns from the new input dataset to which we don't know the target. The overall training process of MSVM is given as a stepwise procedure below.

Step-1: Input the air quality data (training data).

Step-2: Label the attributes which have the target will be predicted.

Step-3: Apply data-transformation procedures.

Step-4: Obtain the training parameters for controlling the MSVM.

For training the MSVM model, some parameters need to be defined to control the properties of the overall training process [20]. For example:

➢ Maximum

○ Model size

○ Number of passes

➢ Shuffle type

➢ Regularization type and amount

Once the data source is created successfully, the MSVM model is created. Initially, the data source is divided into two portions as training data (80%), used for training the model. Testing data (20%) is used for evaluating the model. Choose the set of all existing parameters that can operate the model as training parameters. Provide the custom recipe (mathematical methods) to manipulate the parameters. Then the feature transformation is applied to each entity of the data. The data fields are considered as features that are essential to integrate with the input dataset. These features are extracted, learned from the actual data, and create a random I.D. for the data. Based on the random I.D., the data is classified by naming it. The main aim of training the model is to make the model to make the best predictions. By using a different dataset for evaluation, issues like overfitting are avoided.

3. Experimental Results and Discussion

The experiment is conducted on Manali, Chennai, India dataset. The machine learning model (MSVM) has been implemented in Python, verifying the results. The dataset comprises day, week, month, and year-wise air quality information. The air quality data is recorded for 24 hours of each day every month from 2019 to till date. Analyzing SO_2 is carried out at different times like morning, afternoon, and evening. Two different observation locations are used. Python is installed on Windows-10 OS, where the system has a 2.64GHz processor, Intel Core i7, 7th gen processors, 1TB HDD, and 12 GB RAM. The result of the dataset is cross-verified by executing it on COLAB and Kaggle-NoteBook.

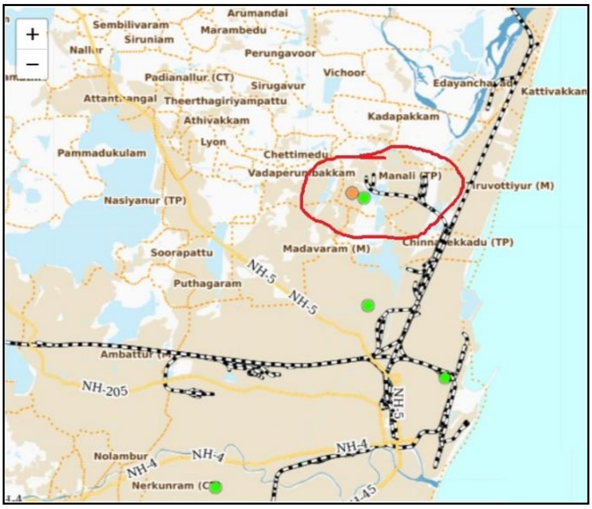

Figure-5. Geo-Location of the Research Area

Table-4. One Day Data From Each Month Of 2021- Till Date

Date	Average	Minimum	Maximum	Overall AQI	Description
06/01/2021	10	10	12	78	Satisfactory
06/02/2021	17	8	35	61	Satisfactory
06/03/2021	21	6	76	27	Satisfactory
06/04/2021	-	-	-	29	Satisfactory
06/05/2021	9	8	10	33	Satisfactory
06/06/2021	36	1	93	174	Moderate
06/07/2021	10	6	15	37	Good
06/08/2021	-	-	-	Insufficient data	
06/09/2021	5	2	8	45	Good

06/10/2021	10	10	12	78	Satisfactory
06/11/2021	21	16	28	47	Good
06/12/2021	24	21	29	73	Satisfactory
06/01/2022	39	30	57	60	Satisfactory
06/02/2022	8	1	14	46	Good
06/03/2022	4	4	4	46	Good
06/04/2022	9	1	23	38	Good
06/05/2022	6	1	18	54	Satisfactory
06/06/2022	10	4	22	70	Satisfactory
06/07/2022	6	3	11	72	Satisfactory

The data is collected from Manali, Chennai, India. It is a refinery station located on the coast of Bay-of-Bengal, with a high population. It is monitored by CPCB, residential cum industrial region. The geo-location of the research area is shown in Figure-5.

The amount of SO_2 present in the air for 1 hour, 24 hours, and 1 year are measured and found to be 29.63, 21.48, and 18.68 µg/m^3, respectively, in Table-5. The corresponding values are small compared to the mean values (350, 125, and 50 µg/m3) of Manali-Chennai (column-2 of Table-5). The prediction from the exposure to SO_2 shows that the Hazard Quotient (H.Q.) value is less than 1.0 for infants, kids, and adults. From Table-3, it is identified that the risk is negligible even for an unhealthy individual. When the exposure to SO_2 is acute, infants and kids (2.0×10^{-3}) are more prone to the same exposure as adults (1.4×10^{-3}). If the hazard quotient is more than one, then there are certain health risks for unhealthy individuals, which differ depending on the age. The worst-case exposure of the SO_2 is given in Table-6. The worst-cast exposure of SO_2 helps to take preventive actions to the public before meeting the risks. Table-4 shows that the concentration of SO_2 is considerably good compared to other pollutants. In this paper, the data set is obtained from a data source publicly available in [20] which has the air quality data for various states of India. Extracting the different data information from other data sources is difficult since many missing data values exist. For this reason, the air quality data from 2017 to 2022 is taken from [21] to experiment with the MSVM model.MSVM model is trained to extract the information about SO_2 contamination in the air year-wise (Figure-6) and the particular date of each month, 2021 (Table-4).

Table-5. Ambient SO_2 Concentration

Averaging Period	SO_2($\mu g\ /m^3$) mean$\pm$ STD	The exposure level of SO_2
1hr	31.48 ± 29.43	350*
8hrs	31.48 ± 29.43	-
24hrs	24.17 ± 23.71	125*
Yearly	15.32 ± 23.16	50*

Table-6. Worst-Case Exposure of SO_2

Group	Intermediate Worst-Case	Chronic Worst-Case
Infant	$1.1\ \times 10^{-1}$	$7.55\ \times 10^{-1}$
Children	$1.1\ \times 10^{-1}$	$4.49\ \times 10^{2}$
Adult	$6.76\ \times 10^{-2}$	$6.98\ \times 10^{2}$

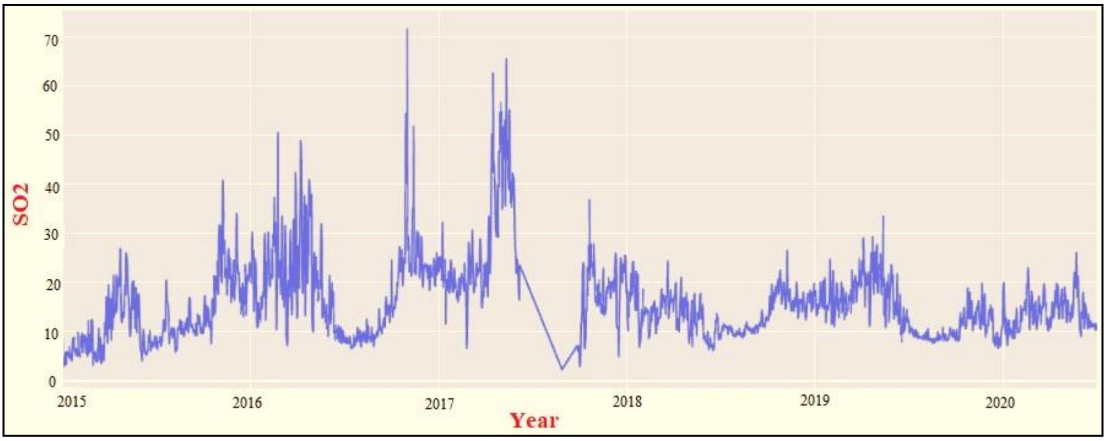

Figure-6. SO_2 Prediction Yearwise (Manali-Chennai)

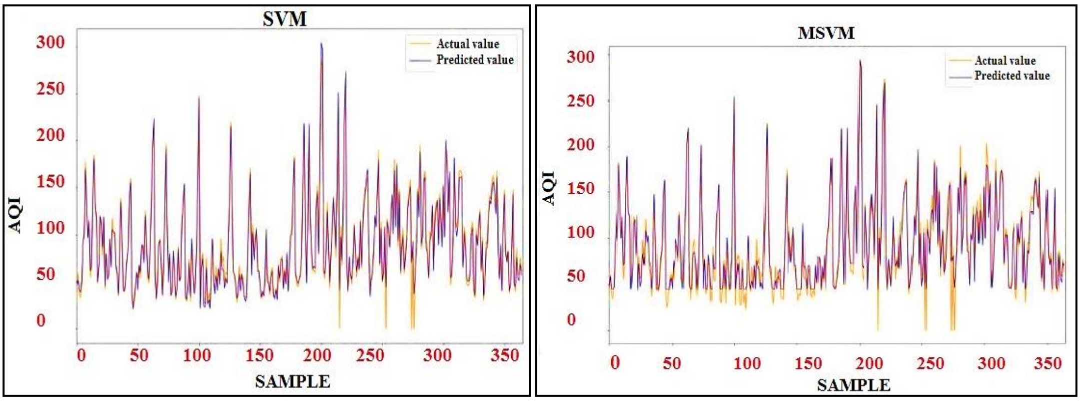

Figure-7. AQI Prediction of SO_2 Using SVM and MSVM

Similarly, the data for every day, week, month from 2015 to today is obtained from https://aqicn.org/city/chennai/, fed into MSVM for analysis. Of the entire dataset, 80% is used for training the model and evaluated with 20% of the data. From the experiment, the predicted SO_2 value is compared with the actual value shown in Figure-7 for SVM and MSVM algorithms. To avoid high computational time complexity, the comparison was carried out for a set of sample data. The sample data is taken from 50 to 350. Also, the performance evaluation is obtained by changing the data size and duration of the data monitored in the experiment.

This paper analyzed the daily air quality data over Manali, Chennai. Though the paper's objective is to extract and determine the air quality from SO_2 contamination, the few gaseous pollution levels in the dataset are also verified. In addition to SO_2, it is also identified that the humidity, temperature, and speed increased by 10%, 2℃, and 2m/s from 2019 to 2020, respectively. The amount of SO_2 is reduced by 71.3% in Manali and reduced by 40.x% in Teynampet, and 71% in Velachery, Chennai. The amount of PM.x is also reduced from 24% to 65%, respectively. The Multi-class Support Vector Machine analysis reveals the variation in the data (SO_2) at Manali, Chennai predicted for different periods. This paper helps to predict the health risk level based on the air pollution level month-wise and forecasted to the public.

People can take necessary actions to save themselves based on the severity level. The prediction accuracy is cross-verified manually to evaluate the performance.

Though uncertainties are present in risk analysis, it is identified that an application is required for risk analysis to provide a decision support system for health-related decision-making frameworks. Human health risk prediction is a safety factor adopted in air quality data analysis. This paper has found that the ecological and environmental information needs to be correlated with the population information. The ecological data analysis considered that individuals in the research region are exposed to the concentration of various air pollutants that trigger multiple diseases. Also, the earlier systems could not predict the increased health risks due to the combination of different gaseous pollutants.

The merits of this work are that it is potential research work. Though one of the research works focused on Malani, Chennai, it describes the health risks related to the % SO_2 contaminated in the air, not by the other pollutants. It is focused on predicting the exposure of SO_2 in terms of hours, days, weeks, months, and years on ambient air pollution data. This research follows a quantitative and qualitative analysis of the prediction of SO_2 contamination used to predict health risks equally.

4. Conclusion

Ambient air pollution is mixing unwanted particulates and gaseous pollutants from NO_2, CO, O_2, O_3, and SO_2. The normal, mild, moderate, and severe ambient contamination of gaseous pollutants. It is monitored using geological sensors, and their mixing level is obtained to forecast the health risks. This paper focused only on SO_2 pollutants and associated health risks to avoid major health problems. Since the data is time-series and continuous, this paper used one of the machine learning algorithms, Multi-class Support Vector Machine, for air quality data analytics. The analytics results show that only in some seasons is SO_2 exposure high and affects infants and children. Usually, in Manali, Chennai area, the contamination of SO_2 is normal and will not affect human health. Thus, it is concluded that the MSVM model can accurately predict the contamination of air pollutants. It is also suggested that MSVM is good for single pollutant prediction in single region data.

4.1 Future Work

In the future, the work will be extended to predict the contamination of gaseous pollutants from various locations in Chennai. For the prediction process, deep learning algorithms can be used to increase efficiency.

References

1. Kumar, A., & Goyal, P. (2011). Forecasting of air quality in Delhi using principal component regression technique. Atmospheric Pollution Research, 2(4), 436-444.
2. Liu, H., Li, Q., Yu, D., & Gu, Y. (2019). Air quality index and air pollutant concentration prediction based on machine learning algorithms. Applied Sciences, 9(19), 4069.
3. Sonu, S. B., & Suyampulingam, A. (2021, August). Linear Regression Based Air Quality Data Analysis and Prediction using Python. In 2021 IEEE Madras Section Conference

(MASCON) (pp. 1-7). IEEE.

4. Maleki, H., Sorooshian, A., Goudarzi, G., Baboli, Z., Tahmasebi Birgani, Y., & Rahmati, M. (2019). Air pollution prediction by using an artificial neural network model, Clean Technol. Envir., 21, 1341–1352.
5. Aditya, C. R., Deshmukh, C. R., Nayana, D. K., & Vidyavastu, P. G. (2018). Detection and prediction of air pollution using machine learning models. International Journal of Engineering Trends and Technology (IJETT), 59(4), 204-207.
6. Pant, P., Lal, R. M., Guttikunda, S. K., Russell, A. G., Nagpure, A. S., Ramaswami, A., & Peltier, R. E. (2019). Monitoring particulate matter in India: recent trends and future outlook. Air Quality, Atmosphere & Health, 12(1), 45-58.
7. Çelik, M. B., & İbrahim, K. A. D. I. (2007). The relation between meteorological factors and pollutants concentrations in Karabük city. Gazi University Journal of Science, 20(4), 87-95.
8. Sharma, N., Taneja, S., Sagar, V., & Bhatt, A. (2018). Forecasting air pollution load in Delhi using data analysis tools. Procedia computer science, 132, 1077-1085.
9. Shakir, M., & Rakesh, N. (2018, August). Investigation on air pollutant data sets using data mining tool. In 2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), 2018 2nd International Conference on (pp. 480-485). IEEE.
10. Naddafi, K., Hassanvand, M. S., Yunesian, M., Momeniha, F., Nabizadeh, R., Faridi, S., & Gholampour, A. (2012). Health impact assessment of air pollution in megacity of Tehran, Iran. Iranian journal of environmental health science & engineering, 9(1), 1-7.
11. Khaniabadi, Y. O., Goudarzi, G., Daryanoosh, S. M., Borgini, A., Tittarelli, A., & De Marco, A. (2017). Exposure to PM10, NO2, and O3 and impacts on human health. Environmental science and pollution research, 24(3), 2781-2789.
12. Gunasekaran, R., Kumaraswamy, K., Chandrasekaran, P. P., & Elanchezhian, R. (2012). Monitoring of ambient air quality in Salem city, Tamil Nadu. Int J Curr Res, 4(3), 275-280.
13. Tikhe Shruti, S., Khare, K. C., & Londhe, S. N. (2013). Forecasting criteria air pollutants using data driven approaches; An Indian case study. Journal Of Environmental Science, Toxicology And Food Technology (IOSR-JESTFT), 3(5), 1-8.
14. Chaloulakou, A., Grivas, G., & Spyrellis, N. (2003). A comparative assessment is a neural network and multiple regression models for PM10 prediction in Athens. Journal of the Air & Waste Management Association, 53(10), 1183-1190.
15. A. J. Cohen, M. Brauer, R. Burnett et al., "Estimates and 25 year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the global burden of diseases study 2015," 2e Lancet, vol. 389, no. 10082, pp. 1907–1918, 2017.
16. Pandey, A., Brauer, M., Cropper, M. L., Balakrishnan, K., Mathur, P., Dey, S., & Dandona, L. (2021). Health and economic impact of air pollution in the states of India: the Global Burden of Disease Study 2019. The Lancet Planetary Health, 5(1), e25-e38.
17. Vapnik, V.; Cortes, C. Support Vector Networks. Mach. Learn. 1995, 20, 273–297.
18. Drucker, H.; Burges, C.J.; Kaufman, L.; Smola, A.J.; Vapnik, V. Support vector regression machines. In Advances in Neural Information Processing Systems; MIT Press: Cambridge, MA, USA, 1997; pp. 155–161.
19. Boningari, T.; Smirniotis, P.G. Impact of nitrogen oxides on the environment and human health: Mn-based materials for the NOx Abatement. Curr. Opin. Chem. Eng. 2016, 13, 133–141. [CrossRef].
20. https://docs.aws.amazon.com/machine-learning/latest/dg/training-parameters.html.
21. https://www.kaggle.com/datasets/fedesoriano/air-quality-data-in-india