ISSN print: 1224-5984
ISSN online: 2248-3942

Rom Biotechnol Lett. 2022; 27(1): 3270-3281
doi: 10.25083/rbl/27.1/3270-3281

Received for publication, December, 14, 2021
Accepted, January, 16, 2022

Original paper

Hybrid Approach for Human Diseases Prediction Using Air Quality Index

VENU D[1], D. YUVARAJ[2], M. MURALI[3], NITIKA VATS DOOHAN[4]

[1] Department of ECE, Kakatiya Institute of Technology and science, Warangal
[2] Dept of Computer science, Cihan University – Duhok, Kuridsitan Region, Iraq
[3] Department of IT, Sona College of Technology
[4] Medi-Caps University, Indore

Abstract

Air pollution has become an extremely serious issue as the air pollutants emitted from motor vehicles has a greater impact on human health than other contaminants. Air quality forecasting plays a major role in giving warning to people and controlling air pollution. The single technique forecasting has various drawbacks such as low accuracy, low performance and low speed. Our present work overcome the above drawbacks by using a hybrid model approach. Our proposed method aims to forecast air quality to predict the hourly concentration of air pollutants using a hybrid model of data mining and machine learning. It predicts diseases due to emission of air pollutants from the motor vehicles based on Air Quality Index level. The CLusteringInQUEst algorithm is used to cluster geo-spatial data for specific input region. The Air Quality Index (AQI) for desirable set of important air pollutant features was calculated from the datasets produced by air pollutants from atmosphere. The calculated AQI was the input to the eXtreme Gradient Boosting (XGB) decision tree. It then classifies AQI level for the specific air pollutants. Then the diseases were classified using XGB algorithm.CLIQUE method has chosen than any other data mining techniques for which it can accurately predict diseases based on AQI values. XGBoost classifier is known for its good performance gradient bosting tree models which is very fast and an efficient one for both computation time and memory. Hence the above two techniques were combined as a hybrid approach to get the benefits of those features.The hybrid model produces a result with a higher performance, accuracy and speed compared to other models. In this paper, we have compared accuracy and precision rates for the hybrid approach with two single techniques such as Support Vector Machine and Random Forest.An accuracy and Precision rates of our proposed hybrid approach was 98.6% and 98.7% than Support Vector Machine has 93.85% and 94.8% & Random Forest has 94.28% and 94.52% which proves that hybrid approach is an efficient diseases prediction technique in real-time environment.

Keywords Data Mining, CLustering In QUEst, Machine Learning, eXtremeGradiant Boosting, Air Quality Index, high performance, high-speed, accuracy.

To cite this article: VENU D, YUVARAJ D, MURALI M, DOOHAN NV. Hybrid Approach for Human Diseases Prediction Using Air Quality Index. *Rom Biotechnol Lett*. 2022; 27(1): 3270-3281. DOI: 10.25083/rbl/27.1/3270-3281.

✉ *Corresponding author: VENU D, Department of ECE, Kakatiya Institute of Technology and science, Warangal
E-mail: stamizharasu@outlook.com

Introduction

Getting worse of air quality due to vehicle emissions has become a main global cause for decreasing of ambient air quality, premature mortality and morbidity living near major roadways. Stroke, pulmonary failure, lung cancer, and chronic respiratory conditions are also linked to ambient air emissions, which accounts for an unprecedented 4.2 million deaths each year.More than 95 percent of the world's population breathes toxic or dangerous food. Air Pollution is the major environmental issue which results around seven million deaths per year and it is attributed to about 9% of deaths around the world. It is also one of the leading risk factors for disease burden.

Majority of the population who suffers from the harmful effects of air pollution are children, elderly and people with respiratory and cardiovascular problems. According to recent findings, there are greater associations between air quality and the onset of respiratory and cardiovascular disorders. Ischemic cardiac failure, pre-existing respiratory condition, pneumonia, stroke, Chronic Obstructive Pulmonary Disease (COPD), lung cancer, and acute lower respiratory infections are common diseases induced by toxic air contaminants released by motor vehicles along roadways in children. "Critieria Air Contaminants" are toxic air pollutants that cause diseases and harm people's health and the environment. Greenhouse smoke, Particulate Matter, Nitrogen oxides (NOx), Carbon Monoxide (CO), Sulphur Dioxide (SO2), and other pollution contaminants emitted by motor vehicles are the most common substances causing morbidity and mortality across the planet.

To address this issue, we proposed a hybrid approach which helps in learning about the air pollution level due to the emission of harmful air pollutants from motor vehicles. The required region was clustered for air pollutant datasets such as Ozone (O3), Sulfur Dioxide (SO2), Carbon Monoxide (CO), and Nitrogen Dioxide (NO2). The Air Quality Index was measured for each air pollutant per hour to determine the degree of air contamination in the environment as a result of toxic gas emissions from motor vehicles. It is used to forecast diseases induced by air pollutants in relation to the Air Quality Index. The AQI of the air pollutants were used to classify the diseases using a Decision Tree algorithm called eXtreme Gradient Boosting (XGB), which simplifies the process of classification as it is versatile in nature. Of all decision tree algorithms, XGB is the most accurate, high speed decision tree algorithm and also it reduces overfitting in classifying the disease caused by air pollutants. The proposed model helps in predicting the diseases caused by harmful air pollutants which helps in protecting human beings from health hazards. The main contribution of this paper includes,

To propose an effective regional clusteringusing Grid based partitioning algorithm CLIQUE to cluster the geo-spatial data in real-time environment.

For efficient classification process, eXtreme Gradient Boosting (XGB) is used to classify diseases according to the Air Quality Index values.

LITERATURE SURVEY

Agarwal et al. [1] dealt with six major megacities in India as well as six major cities in Hubei province, China, where shutdown steps were strictly enforced. For about three months, real-time PM2.5 and NO2 concentrations were observed at different monitoring stations. Those data concentrations were translated into AQI. After one week of lockout, cities in China and India saw an average decrease in AQI PM2.5 and AQI NO2 concentrations of 11.32 percent and 48.61 percent, respectively, and 20.21 percent and 59.26 percent. As a consequence, the findings reveal that the decrease in AQI NO2 concentration was instantaneous as opposed to the incremental drop in AQI PM2.5 concentration. The lockout in China and India results in a final decline in AQI PM2.5 of 45.25 percent and 64.65 percent, respectively, as well as a reduction in AQI NO2 concentrations of 37.42 percent and 65.80 percent. Thus it shows the significance of vehicle smoke in the environmental pollution.

Sarkar et al. [2] proposed a system for determining improvements in air quality in the municipal corporations of Kolkata and Howrah in West Bengal, India, from the pre-lockdown era to the post-lockdown period. GIS-based methods such as interpolation were used to determine the spatial and temporal distribution of contaminants, whereas statistical methods such as analysis of variance (ANOVA) were used to calculate the mean differences between two distinct systems, and a correlation matrix was used to examine the evolving relationship of pollutants during the pre-lockdown and lockdown phases.

Sanchez et al. [3] proposed a protocol to identify the strategies that could be used for a comprehensive data map (SEM), which would recognise and characterise information on policy measures that could be adopted at the city level to minimise traffic congestion and/or TRAP from on-road mobile sources, mitigating human exposures and adverse health effects.

Markendeya et al. [4] investigated seasonal variations in air emission levels in Lucknow while also reviewing the city's indoor air quality and emphasising the health impacts of major toxins such as PM10, PM2.5, SO2, NO2, Pb, Ni, and aerosols.

Fazziki et al.[5] suggested a method for combining the advantages of agent technology with machine learning and big data tools. For air quality prediction and the least polluted path finding in the road network, an artificial neural networks model and the Dijkstra algorithm are used. HBase and MapReduce are used to execute all data processing operations in a Hadoop-based architecture.

Bigazzi et al.[6] reviewed the effectiveness of traffic management strategies (TMS) for mitigating emissions, ambient concentrations, human exposure, and health effects of traffic-related air pollution in urban areas.TMS can improve urban air quality and pollution-related health outcomes for exposed populations.

The feasibility of a deep learning algorithm for forecasting asthma risk was investigated and validated by Kim et al. [7]. The researchers were interested in the peak expiratory flow rates (PEFR) of 14 paediatric asthma

patients at Korea University Medical Center, as well as the amounts of indoor particulate matter PM10 and PM2.

An et al. [8] suggested a prototype of an optical transmission device with wavelength division multiplexing for electromagnetic interference-free indoor dust monitoring. The ability to view precise dust information in real-time is critical for patients with respiratory diseases. Indoor atmosphere information such as dust accumulation, temperature, and relative humidity were relayed by RGB light sources.

According to Xia et al. [9], a spatial correlation examination was used to determine the associations between PM 2.5 emissions and ARI admissions, and a time series study using a distributed lag non-linear model was also used to assess the associations between PM 2.5 emissions and ARI admissions. There were significant differences in affect between children and adults, most notably in acute lower respiratory infection.

Neto et al. [10] looked at forecasting schemes that used ensembles of Artificial Neural Networks (ANNs), with the aim of improving prediction precision and efficiency. Trainable and non-trainable hybrid approaches were used in this paper to forecast PM 10 and PM 2.5 time series forecasting (particles with aerodynamic diameters smaller than 10 and 2.5 micrometres, respectively) for eight separate locations in Finland and Brazil across various time scales.

METHODOLOGY

SUBSPACE CLUSTERING

Clustering is the process of making a group of abstract objects into classes of similar objects. Traditional methods have the problem called “curse of dimensionality”. To overcome this, a combination of Partitioning algorithm and Grid-based algorithm called grid partitioning (i.e., subspace algorithm) is used. The Partitioning clustering method classifies the information into multiple groups based on the characteristics and similarity of the data. The Grid-based clustering minimizes the computational complexity, in clustering of large datasets.

CLIQUE

Clustering InQUEst (CLIQUE) is a subspace clustering algorithm that constructs static grids from the ground up. It uses the Apriori approach to narrow the search area. CLIQUE manipulates multidimensional data by processing a specific level in the first stage and then moving upward to the higher one. The clustering process in CLIQUE starts by dividing the number of measurements into non-overlapping rectangular units called grids based on the prescribed grid size, and then calculating the dense region based on a defined threshold value. If the amount of data points in a unit crosses a certain threshold, it is said to be large. After that, the Apriori approach is used to generate clusters from all dense subspaces. CLIQUE has unique features, such as the ability to locate clusters in orbitrary form. It can also find any number of clusters based on any number of measurements. Clusters can exist in any subspace, which implies they can be in a single or overlapping subspace. Instances can belong to more than one cluster if the clusters overlap.

The geospatial data was clustered using the CLIQUE subspace clustering algorithm in this process.

The CLIQUE algorithm contains the following steps:

1. Locate spaces that comprise clusters.
   Divide the data space into equal parts and tally how many of each are contained within each partition cell.
   Determine which subspaces comprise clusters using the Apriori principle.
2. Locate clusters
   Find dense classes in any of the target subspaces.
   Count the number of associated dense units in each of the subspaces of interest.
3. Create a brief overview of the clusters.
   Determine the maximum regions that comprise each cluster's cluster of linked dense units.
   Determine the bare minimum of cover for each cluster.

As a result, a subspace clustering algorithm is used to cluster the spatial results.

The advantages of using the CLIQUE algorithm include the ability to,

As long as massive density clusters appear in those spaces, automatically define subspaces with the highest dimensionality.

It is untouched by the order of the records in the input and has no assumptions about data delivery.

Scales linearly with input size and scales well as the number of data dimensions increases.

APRIORI APPROACH

It is a classic algorithm that is useful in mining frequent item sets and associated association rules, according to the Apriori theorem. If a k-dimensional unit is dense, so all of its representations in k-1 dimensional space are also dense, meaning that a dense field in one subspace generates dense regions when projected onto lower dimensional subspaces. Since CLIQUE is based on the Apriori property, its quest for dense units in high dimensions is limited to the intersection of dense units in subspaces.

Datasets of atmospheric air toxins were gathered and used for feature collection. Until qualifying for the decision tree, the appropriate features were chosen during the feature selection phase. The aim of the feature selection techniques is to (i) improve the interpretability of the domain by reducing the feature area, and (ii) improve the efficiency of the machine learning algorithms. The Air Quality Index is determined for each acquired air pollutant dataset after the feature selection phase by taking the average value for the appropriate data. The AQI is a regular air quality indicator that is used to report on the quality of the air. It shows you how safe or dirty the air is, as well as what health affects you might be concerned with. The Air Quality Index (AQI) was developed to determine the volume of pollution in the air as well as their health effects. The Air Quality Index is a number between 0 and 500 that indicates how polluted the air is. The higher the AQI ranking, the higher the level of air emissions and, as a result, the greater the risk to one's health. This importance helps us in giving guidance on how

to defend ourselves from the dangers of air contamination, as well as protecting individuals from developing heart and lung diseases. For e.g, an AQI of 50 or less indicates good air quality, while an AQI of 300 or more indicates dangerous air quality (see Table.1). The main aim of the AQI is to help you consider how local air quality affects your wellbeing. The AQI is split into six parts to make it simpler to understand:

Table. 1. Air Pollution Level based on Air Quality Index

AIR QUALITY INDEX	AIR POLLUTION LEVEL
0-50	Good
51-100	Moderate
101-150	Satisfactory
151-200	Unhealthy
201-300	Very Unhealthy
Above 301	Hazardous

Each group represents a particular degree of public health concern. The six degrees of health risk and their meanings are as follows:

“Good” Your community's AQI ranges between 0 and 50. Air quality is deemed acceptable, and air contamination is deemed to pose minimal or no danger.

“Moderate” is a word that has a lot of different meanings depending on who Your community's AQI is between 51 and 100. The air quality is acceptable; nevertheless, certain contaminants may pose a modest health risk to a very limited number of citizens. People that are unusually vulnerable to ozone, for example, can develop respiratory symptoms.

“Unhealthy for Sensitive Populations” Members of vulnerable groups can feel health consequences when AQI levels are between 101 and 150. As a result, they are more likely to be harmed than the general population. People with lung cancer, for example, are more vulnerable to ozone damage, while people with either lung or heart disease are more vulnerable to particle emissions. When the AQI is in this size, the general population is unlikely to be harmed.

"Unhealthy" When AQI levels are between 151 and 200, anyone can begin to see health effects. Members of vulnerable groups can suffer more severe health consequences.

“Extremely Unhealthy” A health warning is issued when the AQI falls between 201 and 300, indicating that anyone may suffer from more severe health effects.

“Dangerous” Over 300 AQI values prompt health alerts of emergency situations. The community as a whole is more likely to be impacted.

Finally, the Air Quality Index values obtained from the given datasets were implemented for further classification using the decision tree algorithm, which is a machine learning technique. We use an efficient algorithm called eXtreme Gradient Boosting, abbreviated as XGBoosting, for this operation. It is then improved further in terms of efficiency and speed using a random scan technique.

XGBOOST CLASSIFIER

The XGB classifier is a machine learning classifier that is both fast and has good performance gradient boosting tree models. The XGBoost outperforms the competition on a variety of challenging machine learning tasks.XGBoost is a great machine learning model that consistently produces the best outcomes in terms of accuracy. XGBoost evolved from the simple gradient tree boosting concept to severe gradient boosting. In general, XGBoost is faster than other gradient boosting implementations. XGBoost is designed for speed and performance. Its engineering goal is to drive boosted tree algorithms' computing resource boundaries. The algorithm allows efficient use of both computation time and memory. When teaching the model, it makes for the most effective use of resources. It dynamically handles missing values in the dataset. By re-teaching the original model with new data, we will strengthen it. It is faster than other gradient boosting random forest implementations as opposed to other gradient boosting benchmarking random forest implementations. It is memory-friendly, fast, and precise.

The hyperparameters are tuned to enhance XGBoost efficiency. The following XGBoost model hyperparameters must be optimised: maximum depth (max_ depth), a tree depth level parameter; minimum child weight, the amount of hessian weights; subsample, a training data sample ratio; colsample by a tree, the sample column ratio when building each tree; A shrinkage parameter is the learning intensity, and the L1 regularisation parameter is alpha. The number of trees to suit is described by the n estimator hyper parameter in XGBoost. It's also the amount of epochs the algorithm uses to connect a tree before the total number of trees exceeds n estimator count, which improves the model's accuracy. The value of the n estimator is set to 100 by design.

For hyperparameter tuning, the random search technique is used. In the random search process, we generate a grid of potential hyperparameter values. Each iteration tries a random combination of hyperparameters from this grid, records the results, and finally returns the best combination of hyperparameters. Randomsearchcv performs a search over a set number of random parameter combinations. At the end of the search, you can view all of the results through the class's attributes. The best observed score and the hyperparameters that obtained the best score are perhaps the most critical attributes. If you've determined the best collection of hyperparameters, you can define a new model, set the values of each hyperparameter, and match the model to all available data. Scikit-optimize employs a Sequential model-based optimization algorithm to quickly identify optimal solutions to hyperparameter search issues.

After tuning the XGB hyperparameters, the model is qualified for classification and classifies malware and benign files with an accuracy of 99.5 %

Overview of the method

Air exposure is the main source of respiratory and cardiovascular diseases in our daily lives. When the world's population expands, so does the consumption of automobiles. As a consequence, mercury emissions from diesel cars cause severe heart and lung conditions. In this step, we suggest a hybrid model approach that uses data mining and machine learning technology to classify and forecast diseases. The CLIQUE and XGBoosting algorithms were used to accurately predict diseases.

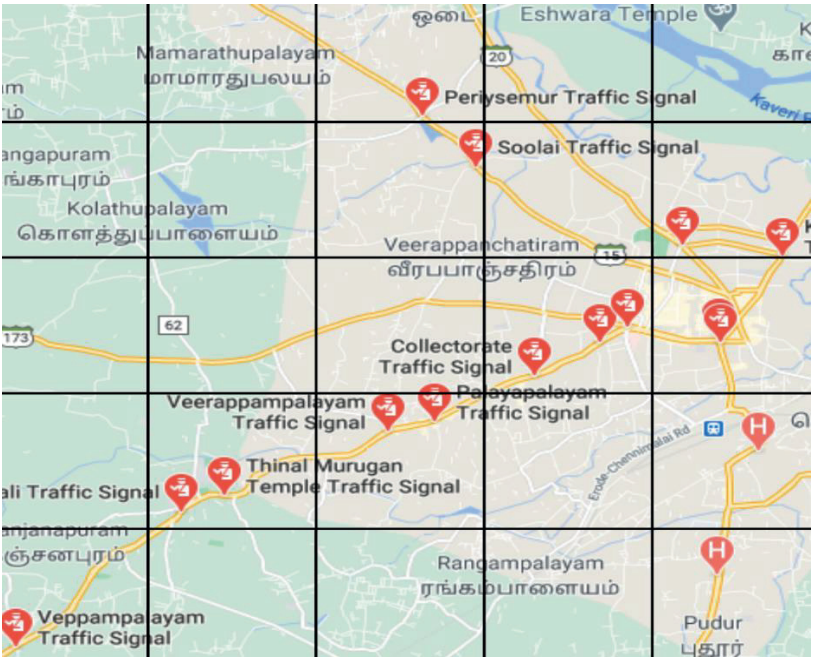

Figure 1. Geo-Spatial region

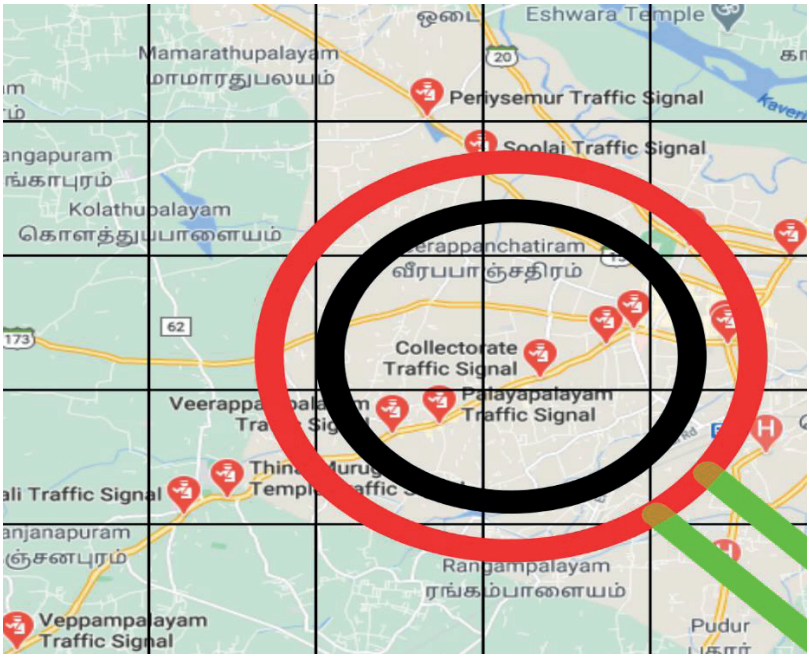

Figure 2. Clustering of region using CLIQUE algorithm

To start, geospatial data was clustered using a grid partitioning algorithm, a data mining technique. We use the CLustering In QUEst algorithm in our proposed work, which is a subspace clustering algorithm used to cluster geo-spatial regions in order to obtain all air pollutant datasets from the atmosphere. In this scenario, for example, we need air pollutant datasets near traffic signals in the erode region. The field is clustered using the grid clustering algorithm, as seen in Figure 1. If we choose a particular location, Signals in and around Palayapalayam, for example, are clustered by our algorithm using the CLIQUE subspace clustering method, as seen in Figure 2.

Our proposed grid partitioning algorithm clusters the input region in an efficient manner. It enhances the speed of clustering and accuracy. The acquired air pollutant features are then applied for feature selection process. The required air pollutant features like CO, Ozone, SO_2, NO_2 and Particulate Matters ($PM_{2.5}$, PM_{10}) are then selected in the feature selection process. Each and every selected air pollutant datasets are then calculated against Air Quality Index. Then, the AQI value is the input for eXtreme Gradient Boosting decision tree algorithm.

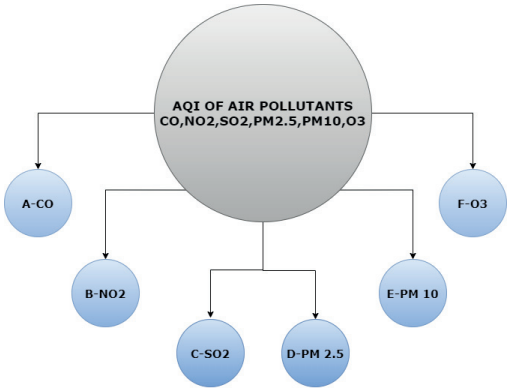

Figure 3. eXtreme Gradient Boosting Algorithm Decision Tree

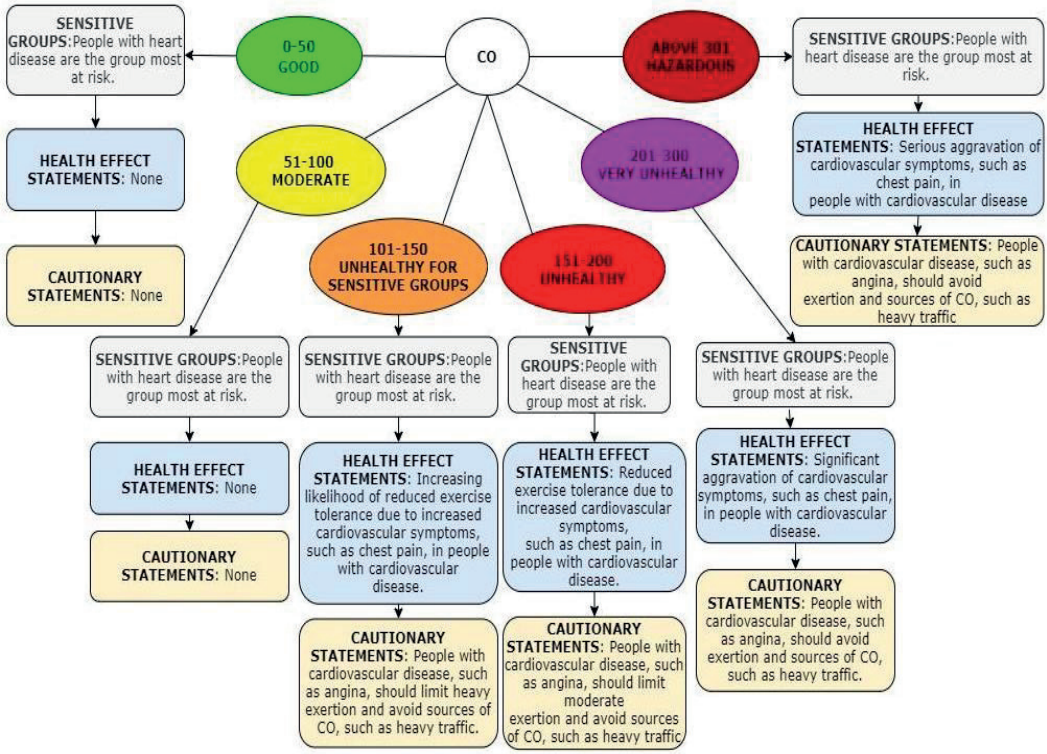

Figure 4. Classification for Carbon Monoxide A-(CO)

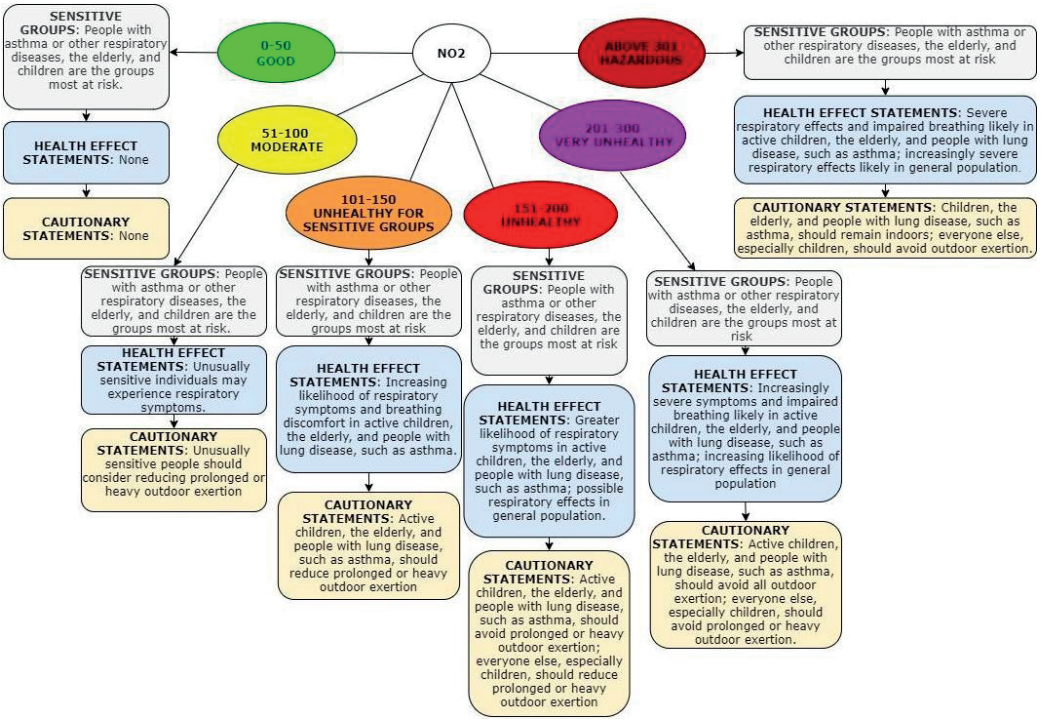

Figure 5. Classification for Nitrogen dioxide B-(NO_2)

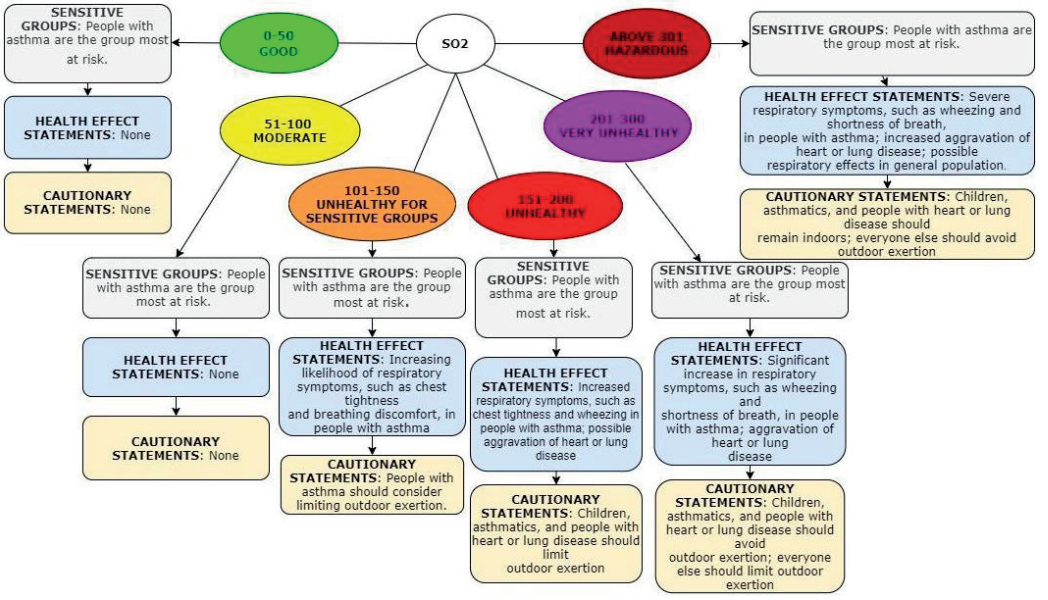

Figure 6. Classification for Sulphur dioxide C-(SO_2)

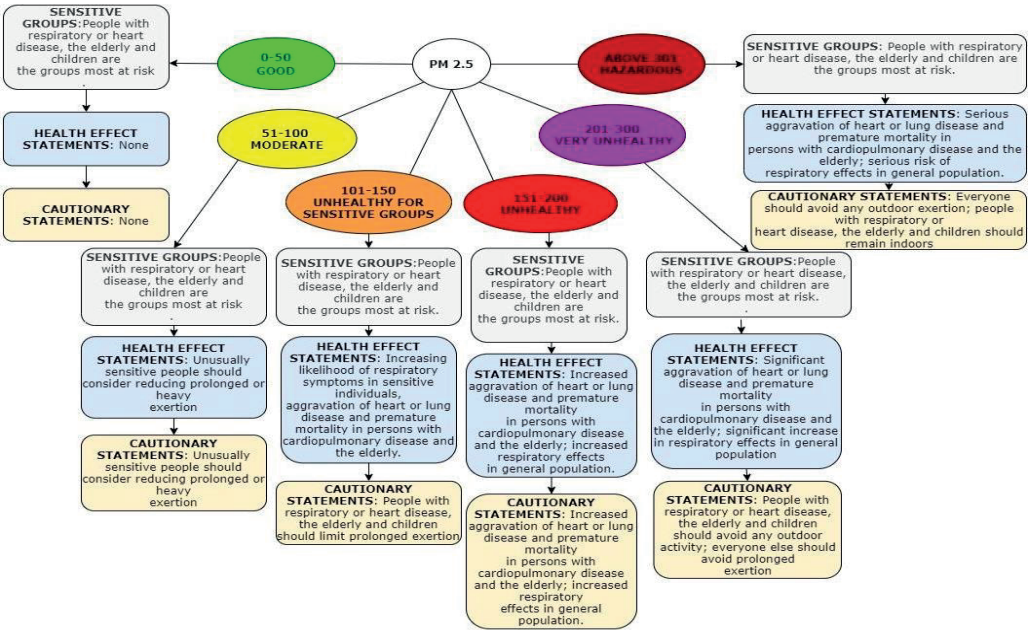

Figure 7. Classification for Particulate Matter 2.5 ($PM_{2.5}$)

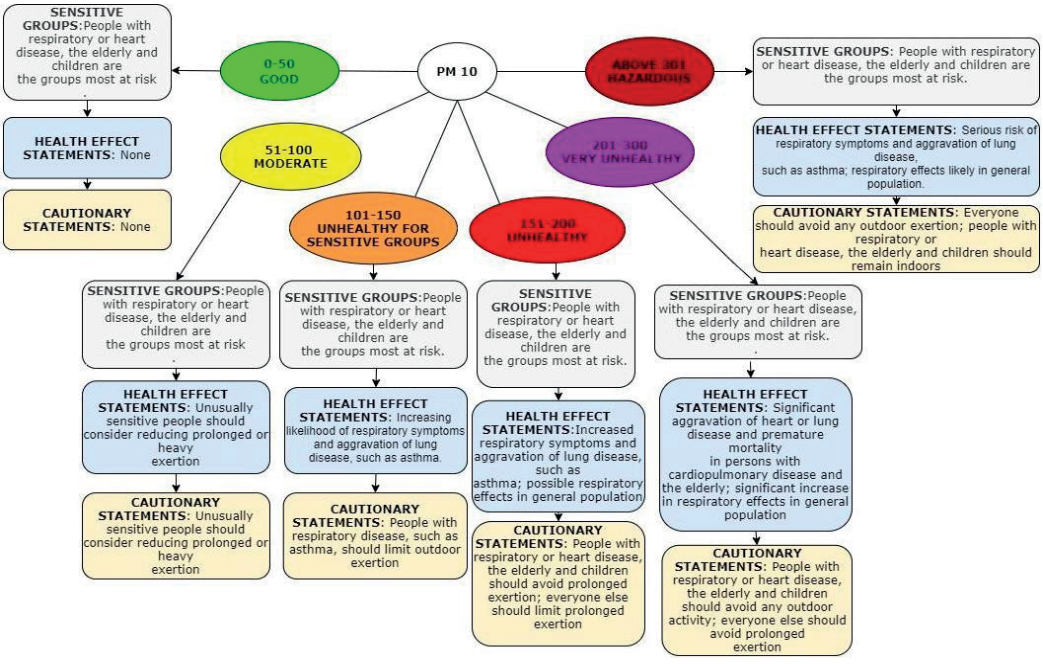

Figure 8. Classification for Particulate Matter 10 (PM_{10})

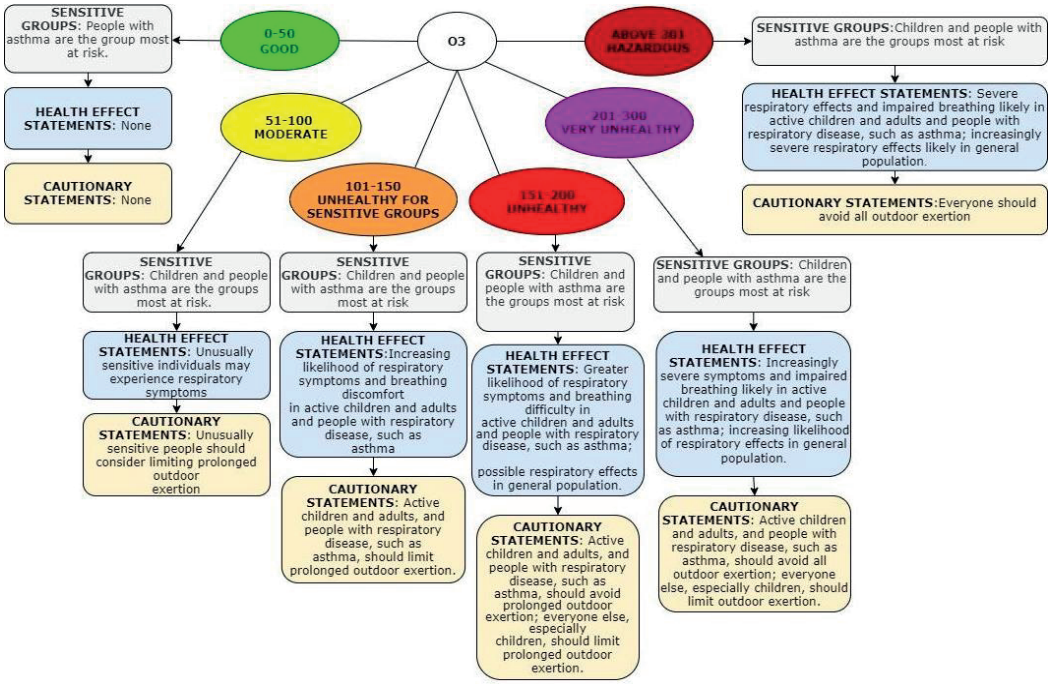

Figure 9. Classification for Ozone (O_3)

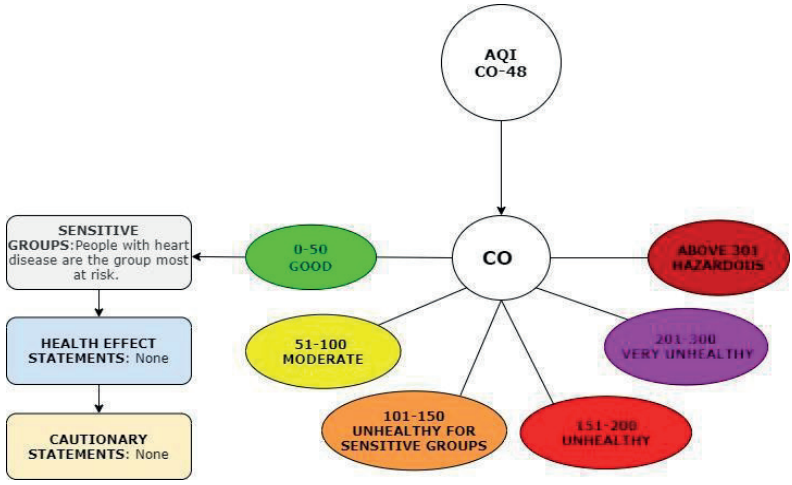

Figure 10. Example classification using eXtreme Gradient Boosting

The higher the AQI amount, the higher the level of air pollution and, as a result, the higher the health danger. The AQI value measured from the air pollutant databases is listed for AQI values. The condition is deemed good when the AQI is between 0 and 50. When the AQI value varies between 51 and 100, the condition is Moderate. When the AQI varies between 101 and 150, the condition is Satisfactory. When the AQI value varies from 151-200, the status is Unhealthy. When the AQI value varies between 201-300, the condition is Quite Good. When the AQI value

exceeds 300, the condition is declared hazardous. The classification is based on AQI ratios, and it is then further categorised for disease prediction. The resulting AQI amounts are listed for diseases in order to learn more about the health problem.

EXAMPLE

If we consider AQI of Carbon Monoxide gas as 48.So the input of the decision tree is taken as AQI-48.So the input of the decision tree is taken as AQI-48. It checks for the gas as well as AQI category for 48.It chooses the category from 0-50 and the AQI level is stated as GOOD. From this category, we can predict the groups which are sensitive as people with heart disease are the group most at risk. From this category, we can predict the groups which are sensitive as People with heart disease are the group most at risk, statements for health conditions as None because this level will not affect humans and also cautionary statements as None hence there is no health affects for humans.

The overall process flow diagram for this proposed work is shown in the Figure.9

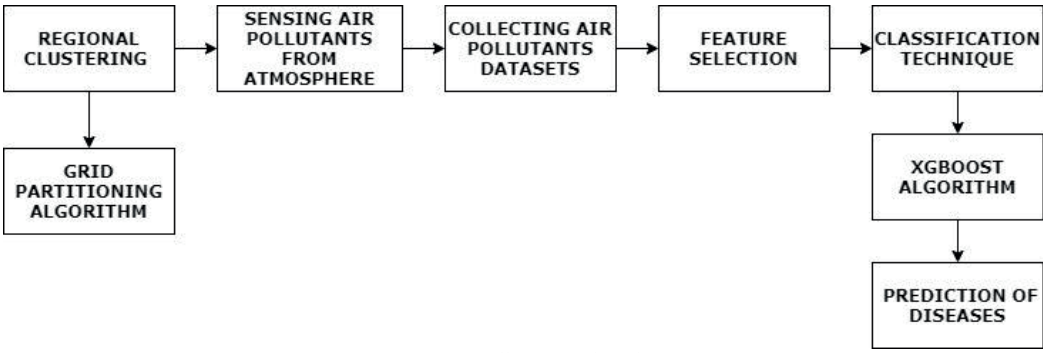

Figure 11. Overall process flow diagram

Results and discussion

Our proposed approach employs a mixture of data mining and machine learning algorithms for effective disease identification, assisting in the alerting of citizens to risk statements provided by air pollution released by motor vehicles. Grid partitioning is an elegant technique for clustering data points that is also recognised for its superior efficiency. The eXtreme Gradient Boosting algorithm which is a well-known for its parallel processing and is a non-greedy tree pruning for decision trees that reduces computational expense and time in decision trees. As a result, these two benefits of algorithms have been merged to conduct an effective procedure for disease prediction triggered by atmospheric air contaminants. Quality measures such as accuracy, precision, recall, and F-score have been used to test the proposed work. The average air pollutant datasets for PM2.5,PM10, NO2, CO, SO2, and O3 for a specific location are shown in Table.2. It is the dataset that reflects the possibility of diseases by taking into account the Air Quality Index (AQI), which is used to warn and alert the public regarding the likelihood of everyday emission levels. A considerable number of people are at risk of serious health problems when the AQI rises. An average human consumes about 11,000 litres of air every day. As a result, the air emissions released from motor vehicles cause health problems for anyone who breathe them in, which may even result in death. Figure.9 depicts a graph of AQI amounts for total air contaminants over the course of a day..

Table 2. Average air pollutant datasets

PLACE	DATE	$PM_{2.5}$	PM_{10}	NO_2	CO	SO_2	O_3	AQI	STATUS
Erode	03-04-2020	81.4	124.5	20.5	0.12	15.24	127.09	184	SATISFACTORY
Erode	04-04-2020	78.32	129.06	26	0.14	26.96	117.44	197	SATISFACTORY
Erode	05-04-2020	88.76	135.32	30.85	0.11	33.59	111.81	198	SATISFACTORY
Erode	06-04-2020	64.18	104.09	28.07	0.09	19	138.18	188	SATISFACTORY
Erode	07-04-2020	108.06	167.62	47.05	0.08	16.05	70.74	225	POOR
Erode	08-04-2020	100.75	172.04	53.94	0.11	14.05	59.2	251	POOR
Erode	09-04-2020	106.25	171.56	43.09	0.13	15.45	66.9	228	POOR
Erode	10-04-2020	83.79	141.83	47.47	0.3	13.35	77.54	223	POOR

Erode	11-04-2020	42.71	80.24	17.35	0.49	9.53	30.68	87	MODERATE
Erode	12-04-2020	54.73	94.12	12.79	0.58	8.21	30.21	89	MODERATE
Erode	13-04-2020	50.91	99.84	16.33	0.64	10.34	26.24	97	MODERATE
Erode	14-04-2020	38.5	106.7	16.82	0.58	11.02	26.62	100	MODERATE
Erode	15-04-2020	7.16	26.39	5.22	0.59	6.62	14.29	46	GOOD
Erode	16-04-2020	7.97	33.2	6.48	0.61	6.01	11.52	40	GOOD
Erode	17-04-2020	6.48	23.83	3.54	0.47	5.97	11.63	33	GOOD
Erode	18-04-2020	7.2	34.11	5.73	0.5	7.01	18.9	49	GOOD
Erode	19-04-2020	8.27	27.27	5.86	0.53	6.69	7.51	32	GOOD
Erode	20-04-2020	139.38	230.27	140.17	1.24	42.26	20.87	304	VERY POOR
Erode	21-04-2020	126.47	187.83	106.1	1.43	15.29	26.24	312	VERY POOR
Erode	22-04-2020	127.94	196.3	92.41	1.33	15.9	28.51	304	VERY POOR

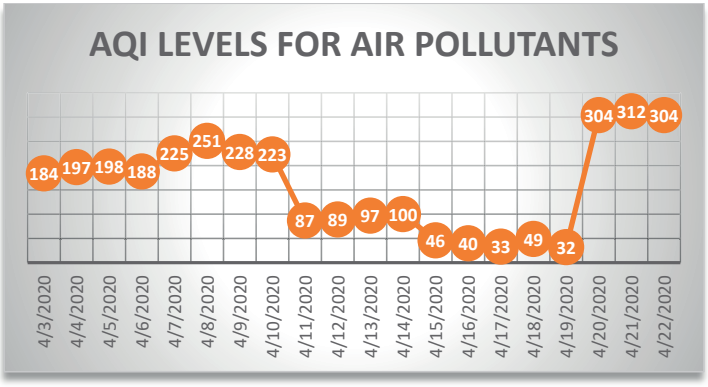

Figure 12. AQI levels for air pollutants

Figure.12 is the graphical representation of Air Quality Index levels for various air pollutants like carbon - monoxide, nitrogen-dioxide, sulphur-dioxide, ozone, particulate matter ($PM_{2.5}$, PM_{10}) for various time periods which is on daily basis.

The three different algorithms were evaluated for different performance metrics.Figure.13 shows the accuracy rate as 98.6% of a hybrid approach which was well greater than other two approaches such as support vector machine has 93.85% and random forest has 94.28%.

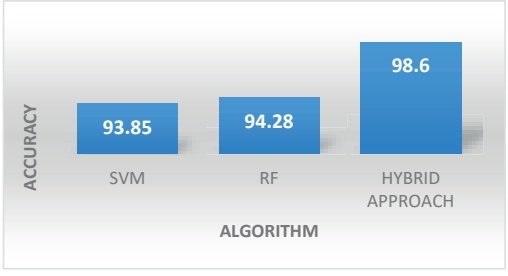

Figure 13. Comparison of algorithms against Accuracy

Figure.14 shows the precision rate of hybrid approach as 98.7% which was more precise than support vector machines has 94.8%and random forest has 94.52%.

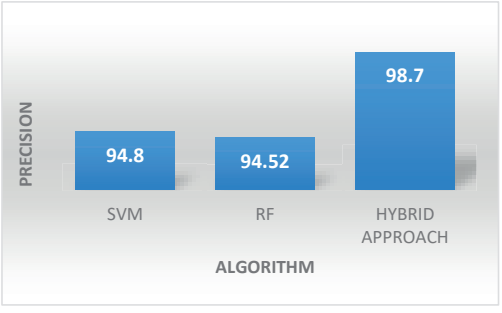

Figure 14. Comparison of algorithms against Precision

Figure.15 shows the recall rate of hybrid approach as 98.95% which was more precise than support vector machines has 96.4% and random forest has 96.75%.

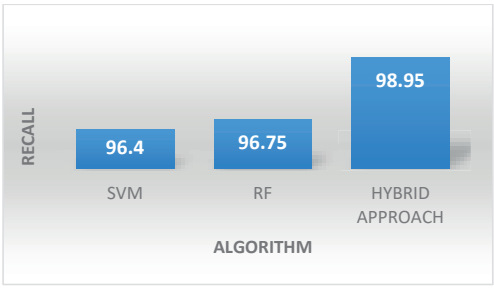

Figure 15. Comparison of algorithm against Recall

Figure.15 shows the recall rate of hybrid approach as 99% which was more precise than support vector machines has 97.23% and random forest has 97.88%.

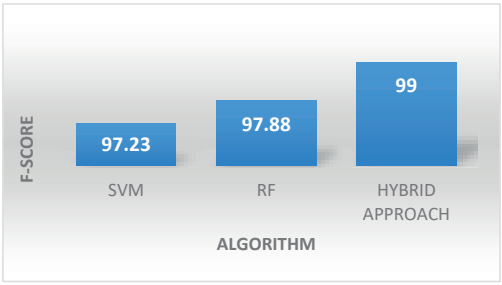

Figure 16. Comparison of algorithm against F-Score

Thus the overall efficiency of our proposed hybrid approach was 98% in comparison with other single techniques.

CONCLUSION

The fundamental aim of this paper is to predict the diseases The fundamental aim of this paper is to predict the diseases caused due to air pollutants emitted from the motor vehicles by an efficient approach of using hybrid algorithms. Here combination of two techniques were used such as data mining and machine learning. In this paper, two efficient algorithms were proposed such as grid partitioning clustering algorithm and eXtreme Gradient Boosting. This hybrid algorithm have been chosen on basis of its performance and accuracy. The poor performance and accuracy in the existing techniques have the difficulty of disease forecasting. Our proposed hybrid technique has performance and accuracy much greater than any other single technique for disease forecasting. Through this method we can attain an accurate disease prediction results using data mining and machine learning models. The grid partitioning clustering algorithm and XG Boosting algorithm are used for efficient prediction of diseases. Hence by this disease prediction method we can alert the people who are residing in nearby cities by preventing them from serious health effects and sometimes death. For the following algorithms such as the Vector support, the Random Forest and the Hybrid solution, the output measurements are compared. An overall accuracy of our proposed hybrid approach was 98 % which proves an efficient diseases prediction in real-time environment.

References

1. Agarwal, A., Kaushik, A., Kumar, S., & Mishra, R.K, (2020). "Comparative study on air quality status in Indian and Chinese cities before and during the COVID-19 lockdown period". Air Quality, Atmosphere & Health. doi:10.1007/s11869-020-00881-z
2. Sarkar, M., Das, A., &Mukhopadhyay, S, (2020). "Assessing the immediate impact of COVID-19 lockdown on the air quality of kolkata and Howrah, West Bengal, India". Environment, Development and Sustainability. doi:10.1007/s10668-020-00985-7
3. Sanchez, K.A., Foster, M., Nieuwenhuijsen, M.J., May, A.D., Ramani, T., Zietsman, J., &Khreis, H. (2020). "Urban policy interventions to reduce traffic emissions and traffic related air pollution: Protocol for a systematic evidence map". Environment International, 142, 105826, https://doi.org/10.1016/j.envint.2020.105826
4. Markandeya, Verma, P.K., Mishra, V., Singh, N.K., Shukla, S.P., & Mohan, D, (2020). "Spatio-temporal assessment of ambient air quality, their health effects and improvement during COVID – 19 lockdown in one of the most polluted cities of India". Environmental Science and Pollution Research. Doi:10.1007/s11356-020-11248-3
5. El Fazziki, A., Benslimane, D., Sadiq, A., Ouarzazi, J., &Sadgal, M. (2017). "An Agent Based Traffic Regulation System for the Roadside Air Quality Control", IEEE Access, 5, 13192–13201. doi:10.1109/access.2017.2725984
6. Bigazzi, A. Y., & Rouleau, M. (2017). "Can traffic management strategies improve urban air quality? A review of the evidence". Journal of Transport & Health, 7, 111–124. doi:10.1016/j.jth.2017.08.001
7. D. Kim, S. Cho, L. Tamil, D. J. Song and S. Seo, (2020). "Predicting Asthma Attacks: Effects of Indoor PM Concentrations on Peak Expiratory Flow Rates of Asthmatic Children", in IEEE Access, vol. 8,pp. 8791-8797,doi: 10.1109/ACCESS.2019.2960551
8. J. An and W. Chung,(2018). "Wavelength-Division Multiplexing Optical Transmission for EMI-Free Indoor Fine Particulate Matter Monitoring", in IEEE Access, vol. 6, pp. 74885-74894, doi: 10.1109/ACCESS.2018.2882576
9. X. Xia and L. Yao,(2019), "Spatio-Temporal Differences in Health Effect of Ambient PM2.5 Pollution on Acute Respiratory Infection Between Children and Adults", in IEEE Access, vol. 7, pp. 25718-25726, doi: 10.1109/ACCESS.2019.2900539
10. P. S. G. De MattosNeto et al., (2021). "Neural-Based Ensembles for Particulate Matter Forecasting", in IEEE Access, vol.9, pp.14470-14490, doi: 10.1109/ACCESS.2021.3050437.