Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 132 (2018) 824–833

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Computational Intelligence and Data Science (ICCIDS 2018)

Evolving Differential evolution method with random forest for prediction of Air Pollution

Rubal[1], Dinesh Kumar[2]

M.Tech Scholar CSE Deptt. DAVIET Jalandhar, Punjab, India [1]
Associate Professor & Head IT Deptt. DAVIET, Punjab Technical University, India [2]
rubal.grewal29@gmail.com[1], erdineshk@gmail.com[2]

Abstract

The aim of this paper is to use a heterogeneous ensemble of differential evolution with random forest method for air pollution prediction. This is different from existing work (independent classifier of Bayesian network and multi-label classifier used for the estimation of air pollutants) as a method is proposing to combine state-of-the-art differential evolution strategies with random forest method instead of focusing on existing single technique. When the existing approach i.e. independent and multi-label classifiers are compared with proposed approach, it shows proposed approach leads to the performance gains. Continuous ambient air quality data of two cities Delhi and Patna from Central Pollution Control Board were publically made available, from where seven pollutants (C6H6, NO2, O3, SO2, CO, PM2.5 and PM10) dataset are collected with daily average concentration.

Keywords: Air pollution, Prediction; Naïve Bayes; Multi-label classifier; Differential evolution algorithm,;Random forest

Corresponding author email: rubal.grewal29@gmail.com

10.1016/j.procs.2018.05.094

1. Introduction

In addition to water and land, the prime resource for sustenance of life is air. Fresh unpolluted air is the basic need of each and every living being. There are different pollutants which are hampering the life on the earth. Air pollution is one the major cause which is affecting life the most. Exposure to air pollution has been associated with morbidity and mortality [4].Variety of air pollutants are given out into the atmosphere by anthropogenic sources, out of which sulfur dioxide, ozone, particulate matters, nitrogen dioxide, carbon monoxide and benzene are having the significant adverse impact on air quality.

An air pollutant can be produced by human activities or from some natural sources, that effect on human's health and the environment. Solid particles, gases or droplets of liquid can be substance of air pollutant. In this study, seven major air pollutants of two cities Delhi and Patna are predicted. The pollutants predicted in Delhi are C6H6, NO2 and CO. The pollutants predicted in Patna are SO2, O3, PM2.5 and PM10. One pollutant is considered at one time as a target variable and its prediction is made accordingly. And same treatment is made with the rest of the pollutants. These pollutants cause harmful effects which are as follows:-

1. Benzene: - Benzene (C6H6) evaporates into the air very quickly. Tobacco smoke is the major cause of Benzene. A disease like anaemia by reduction of blood cells has harmful effects on human health caused by benzene.
2. Nitrogen Dioxide: - Nitrogen dioxide (NO2) is a crucial air pollutant. Combustion of coal, oil, gases and fossil fuels are origin of NO2.
3. Ozone: - Ozone (O3) is present in ambience and builds under the activity of light. The reason of emphysema, throat irritation, asthma and coughing is breathing ozone.
4. Sulfur Dioxide: - Sulfur dioxide (SO2) is an unseeable gas. Breathing it leads to harmful effect on humans which caused to asthma and irritation of throat.
5. Carbon Monoxide: - Carbon monoxide (CO) is an inodorous and toxic air pollutant which has harmful effects to human body by reduction of oxygen. CO is formed by bursting of fossil fuels, oil, woods and vehicle emission.
6. Particulate Matter PM2.5: - Particulate matter (PM2.5) is the mass per cubic meter of air particles with a size (diameter) less than 2.5 micro meters. It is concoction of solid particles and liquid droplets in the air including dust, ash and sea spray.
7. Particulate Matter PM10: - Particulate matter (PM10) is 10 micro meters or less. Dust particles in range of PM10 go deep into the lungs and these particles entrap in the throat and nose.

A Bayesian network classifier can be used to figure out air pollutants. Also organized independent classifier which is based on Bayesian network used to betoken each class variable. Multi-label classifier predicts concurrent multiple air defilement variables. Authors [3] build a multi-label classifier based on Bayesian networks and they try out three different case studies for prediction of PM2.5 and Ozone.

In the computer science literature, Independent and simple bayes are also recognized as other names of naïve bayes. This is a simple technique for constructing classifiers [1]. It defines as all the features in its structure are independent to each other and all are dependent on class variable only. For classification, it takes small number of training data to calculate the parameters and is considered as an advantage of naïve bayes classifier. Multi-label classification is a classification problem where multiple target labels can be assigned to each observation instead of only one. A multi-label classifier symbolizes the joint distribution of each class variables and the features which are used for their prediction. Following are constraints under this classifier: each class can also be parent to the other class; number of classes can be parent to the one feature and also the one feature can be parent to the other. They call METAN (Multi-label ETAN) the resulting classifier [3].

DE is an Evolutionary algorithm. A basic variant of DE algorithm works by having a problem of candidate solution [5]. The DE algorithm incorporates simple arithmetic operators with traditional operators of mutation, recombination and selection to evolve global optima (minima/maxima) from randomly generated candidate solution. The process is defines in four steps: -

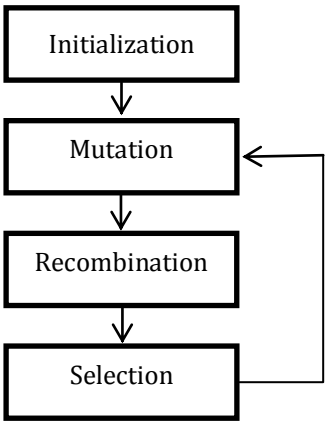

Fig 1: Differential Evolutionary Algorithm Procedure

Random forest classifier is the best learning technique. In machine learning and pattern recognition it is very popular and potent technique. The principle of random forests is to build binary sub-trees using the training bootstrap samples coming from the learning sample L and selecting randomly at each node a subset of X [8]. The classification with highest votes among all the trees in the forest opts by decision forest.

2. Differential Evolution strategies with random forest

This study proposes a new hybrid technique by combining differential evolution with random forest method. The process of differential evolution is based on four steps. The first step of DE is initialization, which is work by having population of a candidate solution. The next step mutation generates new candidate solution using different differential evolution strategies or mathematical formulae. At the third step of DE i.e. recombination, in this new candidate solution is combined with the existing candidate solution using its score or fitness value. Selection is the fourth and last step of differential evolution in which if the new solution is an improvement then it is selected and make part of the population. For random forest method, DE solutions of different strategies represent the training set, while the candidate solution act as validation set. Based on these two sets the random forest method predicts the new candidate solution.

3. Proposed Methodology

The flow chart shown in fig. 4 elucidates the process of predicting air pollutants via Differential evolution method with random forest. Firstly, the data will be read and then the same will be divided into training and testing set. Secondly, differential evolution algorithm will be applied to predict the polluted gases. The algorithm incorporates simple operations of initialization, mutation, recombination and selection to get optimum results from randomly generated candidate solution. In selection criteria, if randomly generated new value is improved from existing value then it is selected otherwise, it is discarded andprocess will start from mutation step. At last, random forest method will be applied, that select number of variables at each split, use final value of differential evolution method, number of rows and number of trees to generate decision trees using training data. Finally it will predict the most common output from decision trees as the final output.

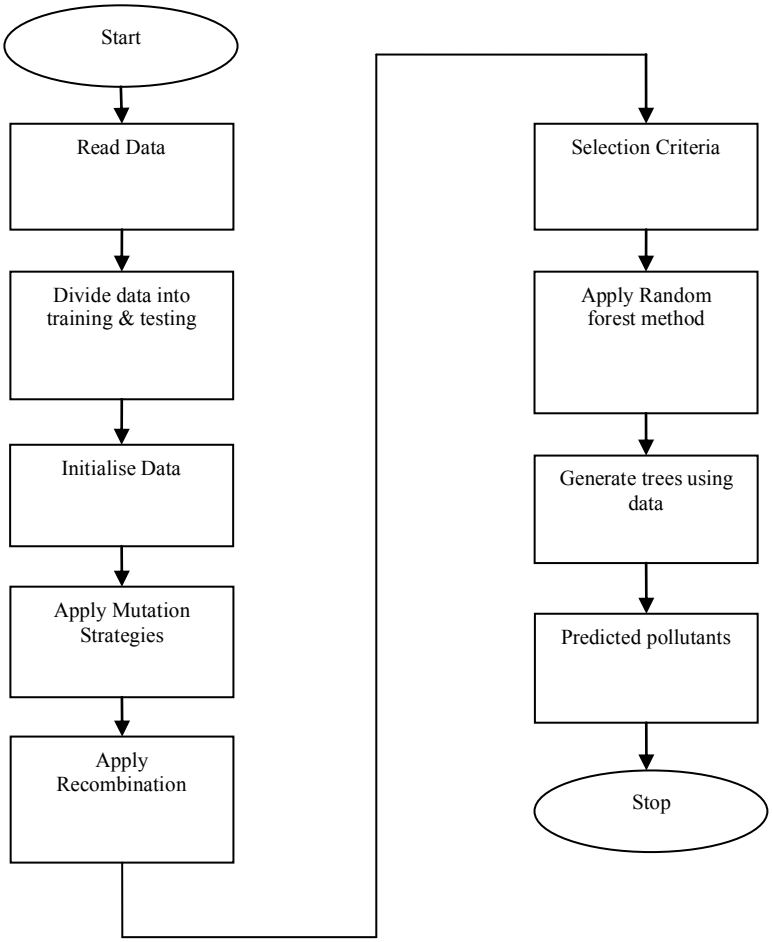

Fig 2: Flow Chart of differential evolution with random forest method

This process also elucidates in below algorithms:-

3.1 Algorithms:-

Algorithm: - Main (D, trainData, testData)
Input: (D is Dataset, trainData is training data, testData is testing data)
// File name of Dataset is Air_Pollution
Step 1: Import Dataset
Dataset = Air_Pollution
// Divide data into training and testing
Step 2: Set trainData = (Dataset * training /100)
Step 3: Set testData = (Dataset – trainData)
Step 4: c = Call Ran_For (mtry, r, n, trainData)
Step 5: Display c
Step 6: Exit

Algorithm: - Diff_Evo(D,F,r1,r2,r3,x,v)
Input: (D is Dataset, F is mutation factor, r1,r2,r3 are three random variables, x is target variable and v is variable used for calculating random variables)
//Initialization
Step 1: Read Data File
Step 2: Shuffle records of data file
Step 3: Set F=Mutation Factor
// Select three random values and one target variable randomly from dataset for seven pollutants.
Step 4: Set r1 = First Random value from Dataset
Set r2 = Second Random value from Dataset
Set r3 = Third Random value from Dataset
Step 5: Set v = r1+F*(r2-r3)
// Recombination
Step 6: If v < x Then
Set New_v = v
Else
Go to step 4
[End If]
// Selection
Step 7: Return New_v

Algorithm: - Ran_For(mtry, r, n, trainData)
Input: (mtry is number of variables selected at each split, r is number of rows selected to build a tree, n is the total number of trees)
// Divide data into mtry, rows(r) and number of trees(n)
Step 1:Read n
Step 2:Seti = 1
Step 3: Repeat while i ≤ n
Input mtryi
Input ri
ki = Call Diff_Evo(D,F,r1,r2,r3,x,v)
ti = Create tree using mtryi, ri and ki
i = i+1
// Calculate average prediction value
Step 4:Pbest = (t1 + t2 + t3 +……+ tn)/n
Step 5: Return Pbest

4. Experimentation

In the evaluation the following dataset is used, detailed in Table 1. The main source of the data of this study is Central Pollution Control Board, India. Data extracted from the period of January'2015 to 3 August'2017 and includes total 946 readings/recordings and are elucidated in the table 1 mentioned below. The seven pollutants are derived from two cities. Another factors like, humidity, temperature, wind speed and wind direction are also taken in to consideration as a meteorological data.

Table 1: Details of the datasets

Pollutants	Concentration	Location
CO	1384.05	Delhi
NO2	74317.15	
Benzene	1118.88	
PM2.5	146831.53	Patna
PM10	219061.47	
SO2	28236.11	
Ozone	38663.85	

In Table 2 the functioning of proposed approach is compared with independent and multi-label classifier. The proposed differential evolution with random forest method outperforms the independent and multi-label approach conceding higher accuracy, higher area under curve, higher success index, higher correlation and lower cost. In the table below the performance of these three techniques are based on equal partitioning of training and testing i.e. 50% data taken for training and 50% data taken for testing.

Table 2: Performance of DE with random forest, multi-label and independent classifier

Parameters	Independent	Multi-label	Proposed
Accuracy	0.65	0.73	0.80
AUC	0.34	0.36	0.52
C5	1.62	1.38	1.02
C10	2.74	2.27	2.1
SI	0.16	0.20	0.32
Correlation	0.36	0.45	0.57

5. Results & Discussions

The performance of proposed hybrid technique has been compared with the independent approach and multi-label classifier. The study includes various performance indicators like accuracy, area under curve, success index, correlation, and cost of X etc. to compare the proposed classifier with existing classifiers. In this comparison, all the values have been received higher than the earlier comparison, excluding the parameter of cost. The base of training has been taken at different percentages and defining the testing data accordingly. If a training percentage is 60% spontaneously 40% data will be considered as data for testing and so on.
Above data set has been taken into consideration for the desired results. These results are summarized as follows:

a) Accuracy: Accuracy is the degree to which something is true or exact. High accuracy implies small error. If both predicted and actual class matches then prediction is considered as accurate.

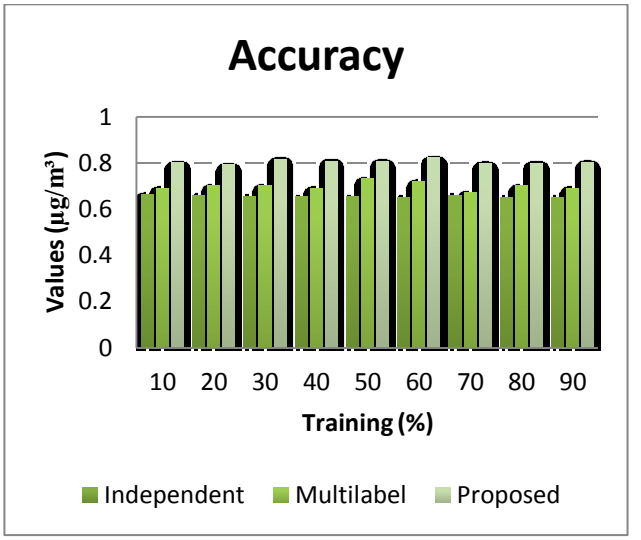

Fig 3: Graph showing accuracy comparison between three techniques on Delhi and Patna case study

b) Area under Curve: The mean of area under the receiver operating characteristic curve is used most of the time by area under curve (AUC). The proportion of random positive is ranked before random negative.

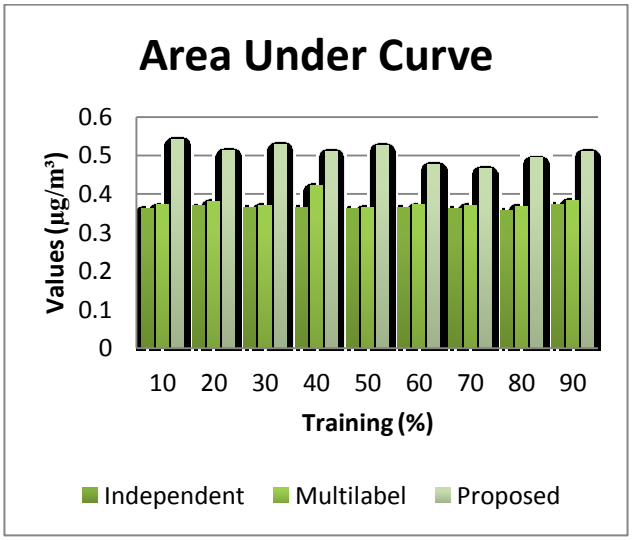

Fig 4: Graph showing area under curve comparison between three techniques

c) Success Index: The difference between true positive rate and false positive rate is known as success index i.e. $SI = tpr - fpr$ .If the actual class is 1 and predicted class is also 1, then it is known as true positive. If the actual class is 0 and predicted class is 1, then it is known as false positive.

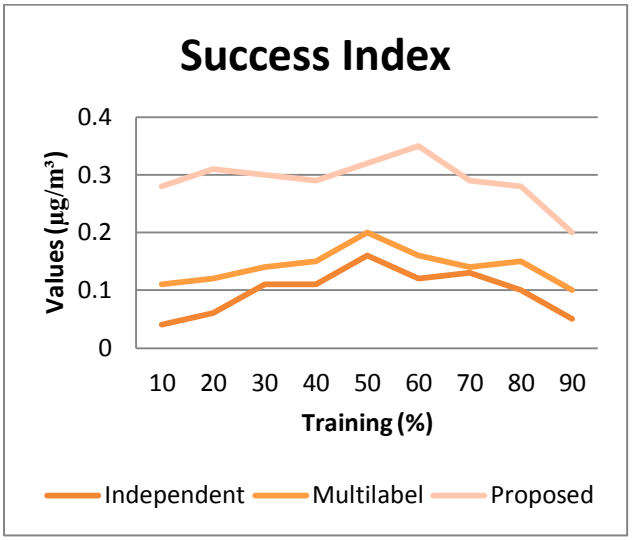

Fig 5: The success index showing performance comparison between three techniques

d) Cost of C5: In this cost of C5 is the mean cost-per-decision where the cost of false positive is 1 and cost of false negative is 5.

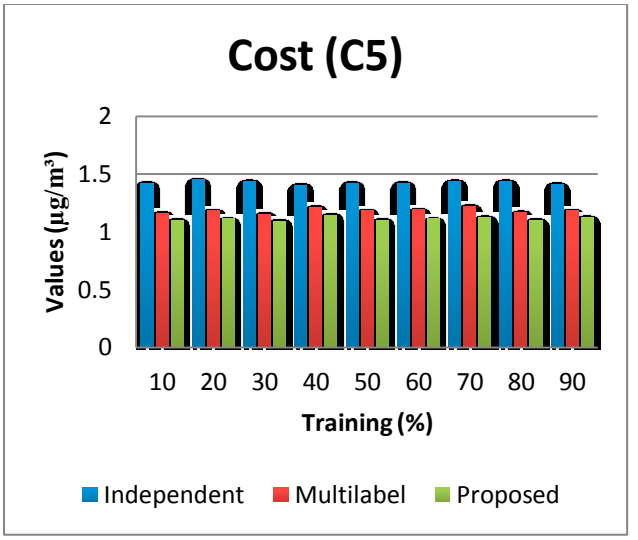

Fig 6: Graph showing cost (C5) comparison (Lower is better)

e) Cost of C10: In this scenario the mean cost-per-decision is C10. Where the cost of false positive is 1 and cost of false negative is 10.

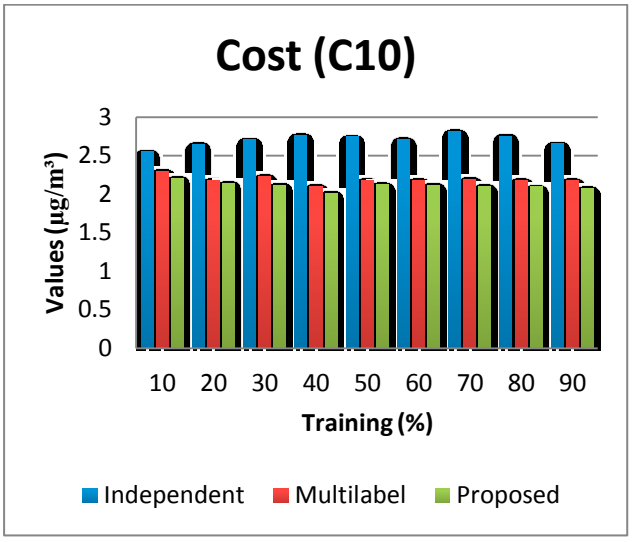

Fig 7: Graph showing cost (C10) comparison (Lower is better)

f) Correlation: The measurement of correlation coefficient is define by degree to which movement of two variables are associated. The range -1.0 to 1.0 is considered for correlation coefficient. A value -1.0 shows a perfect negative correlation and a value 1.0 shows perfect positive correlation.

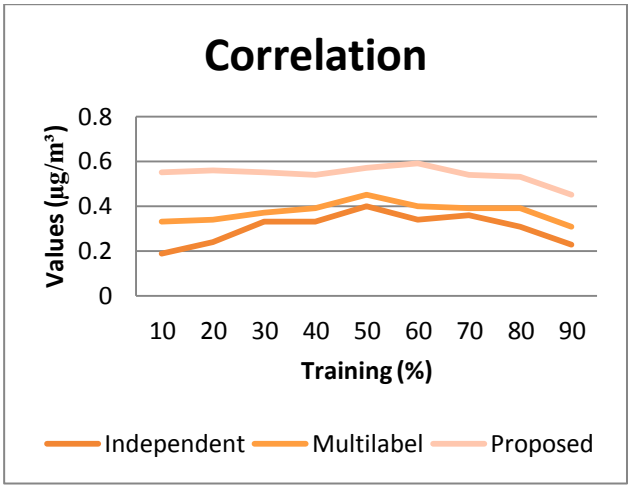

Fig 8: Graph showing comparison of three techniques on the basis of correlation

6. Conclusion & Future Scope

Air pollution is serious, social and environmental problem in India.In this work, the prediction of concentration values of different pollutants is done. The main aim is to predict accurate values and provide precise forecasting information. The air quality index (AQI) is the parameter to monitor the daily air quality. Continuous ambient air quality data of two cities i.e. Delhi and Patna from Central Pollution Control Board (CPCB), India were publically made available. The prediction task is to portend different air pollutants, if the air quality exceeding standard of all seven pollutants C6H6, NO2, Co, O3, SO2, PM2.5, PM10 is high then it result as yes otherwise no at each two

stations. This yields seven class variables to be predicted.From the results derived in previous sections, the study can overall conclude that a combined differential evolution strategy with random forest method can outperform the stand alone with independent classifier of Bayesian network and multi-label classifier technique. Also improved results are received when the parameters including accuracy, area under curve, success index, C5, C10 and correlation compared with existing techniques. This comparison has taken to the point where the new approach has shown the better results than the existing techniques. In future, this approach can also be applied in many other areas of environmental modelling to determine the presence of clouds and prediction of wind power. Moreover,various prediction techniques can merge to predict the data or any information over an environment like artificial neural network.

References

[1] Jie Cheng, et al. "Comparing Bayesian network classifiers" Department of computer science university of Alberta T6G 2H Canada.

[2] Jesse Read, "Multi-label classification" Department of signal theory and communications Madrid, Spain July, 2013.

[3] Giorgio Corani, et al. "Air pollution prediction via multi-label classification." Environmental Modelling & Software 80 (2016): 259-264.

[4] Mario Catalano, et al. "Improving the prediction of air pollution peak episodes generated by urban transport network" Environmental Science and policy 60 (2016) 69-83.

[5] Price, K.V, "An introduction to Differential evolution" New ideas in optimization, McGraw Hill, London 1999.

[6] Tirimula Rao Benala, et al. "Differential evolution in analogy based software department effort estimation." Swarn and evolutionary computation (2016).

[7] Yu Zheng, et al. "U-Air: when urban air quality inference meets big data". In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2013).

[8] Ahmad Taher Azar, et al. "A random forest classifier for lymph diseases". Elsevier Computer methods and Programs in biomedicine 465-473 (2013).

[9] Ramapati Kumar, et al. "Patna ambient air quality report" Centre for environment and energy development (CEED) 2016.

[10] Patricio Perez, "Combined model for PM10 forecasting in a large city," Atmospheric Environment, vol. 60, pp. 271-276, 2012.

[11] Muhammad Waseem Ahmad, et al. "Comparison between random forest and ANN for high resolution prediction of building energy consumption" Energy and buildings 147 (2017) 77-89.

[12] Central Pollution Control Board http://cpcb.nic.in/RealTimeAirQualityData.php

[13] Muhammad Atif Tahir, et al. "Multi-label classification using heterogeneous ensemble of multi-babel classifier". Pattern Recognition Letters 33 (2012) 513-523.

[14] Central pollution control board, Ministry of environment and forests average report criteria. [Online]. http://www.cpcb.gov.in/CAAQM/frmUserAvgReportCriteria.aspx

[15] Air pollution in Punjabi Bagh, Delhi: Real time air quality index visual map. [Online]. http://aqicn.org/map/delhi/punjabi-bagh/

[16] Air pollution in IGSC planetarium complex, Patna: Real time air quality index visual map. [Online]. http://aqicn.org/map/india/patna/igsc-planetarium-complex/

[17] DejanPetelin, et al. "Evolving Gaussian process models for prediction of ozone concentration in the air," Simulation modelling practice and theory, vol. 33, pp. 68-80, 2013.

[18] Jianjun He, Sunling Gong, Ye Yu, Congbo Song and Hongjun Mao, "Air pollution characteristics and their relation to meteorological conditions during 2014-2015 in major chinese cities," Elsevier Environmental Pollution, pp. 1-13, 2017.

[19] J. Alonso Montesinos and M. Martinez Durban "The application of Bayesian network classifiers to cloud classification in satellite images," Elsevier Renewable Energy, vol. 97, pp. 155-161, 2016.

[20] Nir Friedman, Dan Geiger and Moises Goldszmidt, "Bayesian network classifiers," Kluwer Academic Publishers, Machine learning, vol. 29, pp. 131-163, 1997.

[21] D. Domanska and M. Wojtylak, "Explorative forecasting of air pollution," Atmospheric Environment, vol. 92, pp. 19-30, 2014.

[22] A. Lahouarand J. Ben HadjSlama, "Hour ahead wind power forecast based on random forests," Renewable Energy, vol. 17, pp. 0960-1481, 2017.

[23] EleftheriosGiovanis, "The relationship between teleworking, traffic and air pollution," Elsevier Atmospheric pollution research, pp. 1-14, 2017.