International Journal of Scientific Research in Computer Science and Engineering
Vol.6, Issue.2, pp.41-50, April (2018)

Research Paper

E-ISSN: 2320-7639

Ensemble based J48 and random forest based C6H6 air pollution detection

Gagandeep Kaur[1*], Harmanpreet Kaur[2]

[1]Dept. of Computer Science & Engineering, Sri Sai college of Engineering and Technology, Manawala, India
[2]Dept. of Computer Science & Engineering, Sri Sai college of Engineering and Technology, Manawala, India

Available online at: www.isroset.org

Received: 07/Apr/2018, Revised: 12/Apr/2018, Accepted: 20/Apr/2018, Online: 30/Apr/ 2018

Abstract: Air pollution has become a critical challenge for today's world. An efficient monitoring of air pollution gases can help to reduce the pollution in the air. Air pollution cause us many diseases such as cancer etc. Benzene (C6H6) turn out to be more challenging issue in our society, because its sensors are costly to deploy and also not feasible to add too many sensors in urban areas. Therefore, in this paper an efficient monitoring of C6H6 gas has been done by using the ensemble approach. It is feasible to estimate C6H6 by using machine learning because there exists relationship between gases. Extensive experiments have been carried out to evaluate the effectiveness of the proposed technique. It has been found that the proposed technique significantly improves the performance of existing machine learning techniques.

Keywords: C6H6 • Air pollution • Random forest • Machine learning

1. INTRODUCTION

The adverse global environmental changes in respect of its atmosphere, such as, enormously rapid increment of greenhouse gas concentrations, air quality degradation, increase in the abundance of tropospheric oxidants including ozone, stratospheric ozone depletion, concomitant global warming followed by looming threat of climate change and bio diversity degeneration all are fuelled by human activities. [27]. the sources responsible for air pollution are of two categories which are natural sources and man-made sources. The natural sources include forest fires, volcanic eruption, and wind erosion of soil, natural radio activity and decomposition of organic matter by bacteria. The manmade sources are much diversified. These include automobile, industries, thermal power plants and agricultural activities. The fossils fuels (coal, oil, natural gas) are burnt in industries, thermal power plants and automobiles.Different hydrocarbons (methane, butane, ethylene, benzene) and suspended particulate matters (dust, lead cadmium, chromium, arsenic salt etc.) are also present in these emissions. These gases and suspended particulate matter (SPM) produced as result of burning fossils fuels are the greatest source of air pollution. The pollutants released from natural sources of air pollution are dispersed in a vast area and do not cause any serious damage.

Most of the health related air pollutants come from man-made sources of air pollution. In large cities, breathing the polluted air proves harmful to human health. Carbon monoxide, a serious air pollutant, reduces the oxygen carrying capacity of blood and causes nausea, headache, muscular weakness and slurring of speed. Oxides of nitrogen can damage the lungs, heart and kidneys of man and other creatures. The presence of hydrocarbon in air causes irritation to eyes, bronchial construction, sneezing and coughing. In densely populated cities, the air pollution may take the form of industrial smog and photo chemical smog.

Air pollution is one of the biggest public health issues confronting the world today. Air pollution is increasing at rapid rate in the world.The toxic levels of air pollution in and around world are creating quite a menace. The increase in population, emissions from industries and manufacturing activities, automobiles exhaust, etc., are reasons that are contributing to the air pollution. Many countries have declared it as major threat to human life. Currently air pollution is measured by utilizing spatially distributed sensors. However, due to sensor expenses and size limits the operational efficiency. Therefore, many researchers have proposed air pollution detection system using machine learning tools without deploying any particular kind of sensors. It reduces the cost of air pollution monitoring system. Benzene is considered to be a threat for various kinds of diseases. Therefore, an efficient monitoring of benzene becomes a challenging issue. Air pollutants such as benzene (C6H6) have accelerated the rate of cancer among human beings. Currently, atmospheric contamination is measured using spatially separated networks with limited sensors. However, the expenses involving multiple sensors with varying sizes limit the operational efficiency. Therefore, machine learning models to predict the concentration of

benzene in the air, without deployment of actual sensors for benzene detection. It is possible because there is a relation among various atmospheric gasses and thus regression can be performed to measure C6H6 if the concentration level of other gasses is known.

Air Pollution especially in and around urban areas have become, if not the most importantecologicalas well asevolvingstatesnearby the world. Air quality issues are most complex environmental problems and numerous research studies have already reported the impacts of atmospheric pollution on human health and the environment.

Air Pollution has been demarcatedas per any constituent present in the air as a result of anthropogenic activity or natural process that causes adverse effects to human or animal health & physiology, vegetation or materials. Thousands of different chemical compounds (almost all in gaseous forms) are present in our earth's atmosphere, many of which are trace in amount and beyond practical limits of detection of ordinary analytical set-ups. Many of these gases have potential to cause adverse action on the environment either directly or by interactions with other substances in the atmosphere. Pollutants are emitted directly into the atmosphere such as Oxides of Sulphur released by scorching of fossil fuels are known as primary pollutants.Secondary air pollutants are ones that are formed as products of reactions between chemical species existing in the atmosphere, heat (the thermal state of the species) and the radiations coming from the Sun. The most important of such secondary air pollutants is ozone (O3) that is produced trough complex photochemical reactions involving the primary air pollutant NO2 and aerobic oxygen under the influence of solar radiation of wavelength less than 424 nm. Concentration, location and time scale are important features that characterize the air pollution phenomenon of the atmosphere. Apart from these, meteorological conditions formulate the appropriate foundation of understanding of air pollution episode of a given region.

Sources of air pollution can be allocated into four kinds: mobile sources, stationary sources, area sources and natural sources as shown in Fig. 1.

1. Mobile sources include transportation
2. Stationary includes power plants, refineries, and industries
3. Area sources includes agricultural areas, fireplaces
4. Natural sources means wildfires and volcanoes

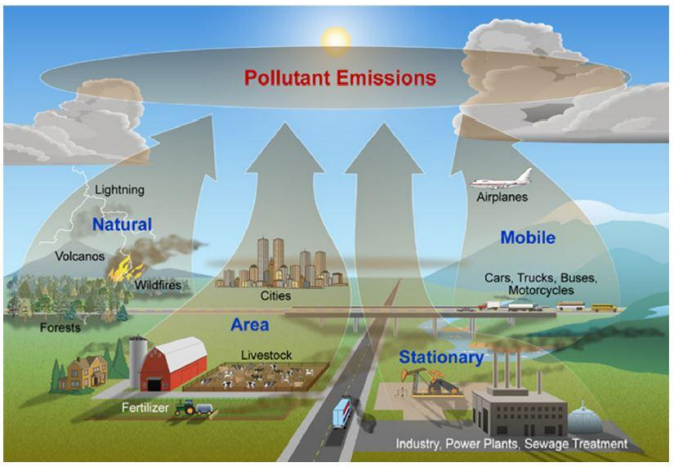

Fig. 1 Four categories of air pollution [5]

1.2 Consequences of air pollution

Polluted air is hazardous for health. Higher concentrations of pollutants cause breathing difficulties, chronic cough, and respiratory diseases. Higher level pollutants are injurious for lung function. According to estimation of world health organization during 2002, indoor air pollution was liable for 1.5 million people death whereas mortality of 2.4 million people directly attributable to air pollution in each year (WHO, 2002). In America, more than 500,000 people died in each year from cardiopulmonary disease related to breathing fine particulate air pollutants (American Chemical Society). 527700 numbers of people died due to air pollution in India (WHO, 2002).

1. Ozone (O3): Ozone that one is dreary and imperceptible, howeverregularlyensuesalongside with additional more observabletypein significant pollution proceedings. It is a gas that can form by a set of photochemical reactions in the presence of sun light and primary air pollutants.
2. Nitrogen Oxides (NOx=NO2+NO): NOx is the generic term for a group of highly reactive gases, which hold nitrogen as well as oxygen in variablequantities, such as nitric oxide (NO) as well as nitrogen dioxide (NO2).
3. Carbon Monoxide (CO): Carbon monoxide is an odor-less, color-less toxic gas. CO generates from incomplete combustion of fossil fuel like unvented kerosene,gas space heaters
4. Volatile Organic Compounds (VOCs): Volatile Organic Compounds (VOCs) is a collection of chemicals that comprise organic carbon, and readily disappear, changeable from liquids to gases whileshowing to air at normal temperature.
5. Sulfur Dioxide (SO2): Sulfur Dioxide is a colourless gas. It is transformed into sulphuric acid in the presence of water vapor. SO2 can be oxidized to form acid aerosols.

The impact of this paper is to improveinfluentialnumericallinkbetweenseveralimpurities. The impurityfeatures which will characterize the dataset

must be generic. Most importantrecipesopinions are as follows:

- The usage of combination of data mining methodscould be done to expand the accuracy rate supplementary for recognition of benzene.
- The combination of random forest and J48 has been disregardedthat can expand accuracy rate supplementary for benzene detection.
- The influence of performance metrics tuning is also unnoticed in presentworks.
- Hence, the studyeffort will becombined machine learning method which willcalculate the benzene from air fumes data in an effectivemode.

2. RELATED WORK

This section contains comprehensive review on existing well-known air quality prediction techniques by various researchers.

Siwek and Osowski (2016) [1] have discussed various data mining techniques of air pollution prediction. The techniques used for feature selection are genetic algorithm and step-wise fit approach. The study demonstrates that the pre-choice of the most primary attributes in a viable manner. The daily average air pollution for next day of various pollutants such as PM 10, SO2, NO2, and O3 is being predicted .Furthermore, various solutions of systems predicting such pollutants are being compared. The fundamental key point examined in the examination is the determination and generation of the prognostic feature which is important during the prediction of the air pollutants. Xiaoguang et al. (2015) [2] comprehensively evaluated and improved the daily air pollution prediction for 74 urban communities in china by various machine learning methods. Five different classification algorithms namely, Random forest model, gradient booting model, SVM model, decision tree model, and hybrid model of the four above models of machine learning techniques are adopted with exclusive feature groups which originated from WRF-Chem model. The best outcomes are acquired by various gathering of feature selection and model selection. In this study, experiment results gives the indication that when the more features are used, then at that point the likelihood to improve precision additionally increments. The environment agents forecasted in this study are $PM_{2.5}$, PM_{10}, SO_2, CO, NO_2, O_3.Yeganeh et al. (2017) [3] approximated the concentration of PM 2.5 by developing the satellite based model using ANFIS (Artificial Neuro Fuzzy Inference System). Authors have compared ANFIS with SVM (Support Vector Machine) and Back-Propagation artificial neural network adaptive model .The distinctive soft computing methods are used to build-up a satellite based model for evaluating the spatiotemporal variation of PM2.5. Sharma et al. (2015) [4]have employed adaptive neuro fuzzy inference system for forecasting air pollutants concentration. Pollutants such as Sulphur Dioxide (SO_2), and Ozone (O_3) in Delhi, India are being taken as environmental agents. A novel application of modified particles swarm optimization for training ANFIS for air pollutants forecasting is successfully investigated. The outcomes got are additionally compared with traditional gradient based method which is normally used for training ANFIS. Three performance metrics which are used for comparative study are MSE (Mean Squared Error), RMC (Root Mean Squared Error) and MAD (Mean Absolute Deviation) which are further being evaluated. Chen et al. (2016) [12] have discussed various machine learning algorithms for forecasting quality of air in urban areas. The discussed machine learning models are used to forecast the application of benzene in atmospheric. An effective machine learning based method is established for estimation of Benzene in the atmospheric. Fu et al. (2015) [5] addresses the danger of lung cancer and vision debilitation as one of the real worries for air quality. The prediction of PM2.5 and PM10 is done by developing an improved version of Feed Forward Neural Network model. The various health organizations in China have considered the high rate of concentration levels of PM2.5 and PM10 as one of the major issue of contamination of air. Yu et al. (2016) [6] discussed Random Forest Approach for forecasting air pollution in inner-city detecting system. The data incorporated is meteorology data, street data, ongoing traffic status and point of interest (POI) circulation. Execution of RAQ is assessed with genuine city information. The proposed approach accomplished better expectation exactness. Reviving outcomes are seen from the examinations that the air quality can be derived with incredibly in elevation precision from the information which is acquired from inner-city detecting. De Vito et al. (2008) [8] have assessed the utilization of neural networks together with on field information recordings for adjusting a multi-sensor device for benzene estimation. The situation is described by huge connections among a few contamination groups. The proposed sensor combination subsystem has been chosen for exploiting both single sensor specificity and situation related connections. Vlachokostas et al. (2011) [9] have discussed that there exist steady relationship between traffic-related air contamination and respiratory symptoms. Be that as it may, numerous urban regions are depicted by the nonappearance of the vital observing foundation, particularly for benzene (C6H6), which is a known human cancer-causing agent. The exhibited outcomes illustrated that the adopted approach is equipped for predicting C6H6 and ought to be considered as correlative to air quality predicting. Singh et al. (2013) [20] developed tree ensemble models for seasonal discrimination and air quality prediction. PCA (Principal Component Analysis) used to identify air pollution sources; air quality indices used for health risk. Bagging and boosting algorithms enhanced predictive ability of ensemble models. Ensemble classification and regression models performed better than SVMs. Proposed models can be used as tools for air quality prediction and management. Qi et al. (2017) [24] proposed a

general and effective approach to solve the three problems in one model called the Deep Air Learning (DAL). The main idea of DAL lies in embedding feature selection and semi-supervised learning in different layers of the deep learning network. The proposed approach utilizes the information pertaining to the unlabelled spatio-temporal data to improve the performance of the interpolation and the prediction, and performs feature selection and association analysis to reveal the main relevant features to the variation of the air quality. Researchers evaluated approach with extensive experiments based on real data sources obtained in Beijing, China. Experiments show that DAL is superior to the peer models from the recent literature when solving the topics of interpolation, prediction and feature analysis of fine-gained air quality.

The primary motivation behind this research work comes after conducting the survey of existing techniques through which followings gaps have been formulated: -

- Sufficient data: To improve proficient machine learning method, it is vital to have appropriate quantity of information for evolving accomplished method.
- Noise in data: Noise in information could mention as annoying information existent in database. Similarly, out of series information could also be termed as noise.
- Pre-mature convergence:It has been observed that majority of existing meta-heuristic-based machine learning models such as particle swarm optimization, genetic algorithm etc. suffer from pre-mature convergence issue. It limits the performance of air pollution prediction techniques.
- Data uncertainty:It has been found that the type-I fuzzy logic has not ability to support the degree of uncertainty. Therefore, it is required to design an efficient type-II fuzzy logic to improve the accuracy rate of the benzene (C6H6) prediction.
- Stuck in local optima:Majority of existing meta-heuristic-based benzene prediction models suffer from stuck in local optima issue.
- Computational speed: The majority of existing meta-heuristic based machine learning models suffer from poor computational speed.

3. METHODOLGY

Followings are the key assistances of this proposed method:

- This proposed work will attain improved accuracy as compared to machine learning method.
- air pollution detection method is appropriate for this proposed work since it usages minus space
- It recognizes benzene with the help of regression based methods. Proposed method will detect the association among additional gases with C6H6.

This work will utilize step by step methodology to attain the objectives of this research work.

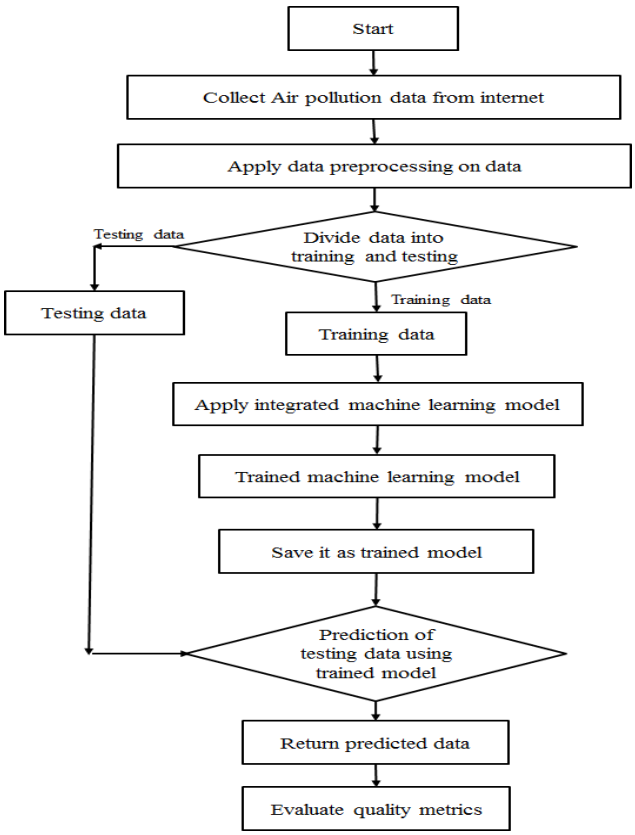

Figure 2: Flowchart of proposed air pollution detection method

4. RESULTS AND DISCUSSION

The proposed method is considered and applied in MATLAB software. Intel core i5 mainframe is applied with 8GB RAM and 2GB graphics card. In order to evaluate proposed model and perform a comparative analysis, following parameters were used:

1. Accuracy

Table 1 and 2 shows the accuracy analysis between the proposed methods as compared to other methods. In both Tables, information is trained and tested on similar dataset, so that's why named as training accuracy.

Therefore Table 1 and 2 reveal that the accuracy metric of the proposed method better outcome other methods such as linear model, neural network, Support vector machine (SVM) and J48.

Table 1: Training Accuracy

Dataset	40%	50%	60%	70%	80%
Liner model	93.7 ± 0.8	94.6 ± 1.1	91.4 ± 1.2	93.6 ± 0.9	93.4 ± 0.8
Neural Network	94.3 ± 0.8	95.4 ± 0.8	91.9 ± 1.0	94.6 ± 1.2	94.3 ± 0.9

SVM	95.0 ± 1.2	96.3 ± 0.9	93.6 ± 1.6	95.9 ± 0.8	96.1 ± 0.9
J48	95.9 ± 1.9	96.5 ± 1.2	94.4 ± 2.7	96.6 ± 1.2	96.2 ± 1.7
Random Forest	97.1 ± 0.9	97.8 ± 0.8	95.1 ± 0.7	98.6 ± 0.9	96.9 ± 0.8
Proposed	98.4 ± 0.7	99.4 ± 0.4	97.3 ± 0.7	99.4 ± 0.5	98.5 ± 0.7

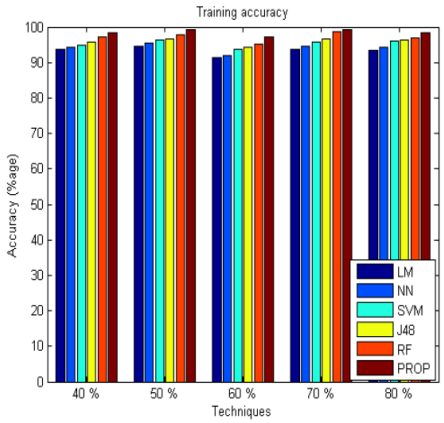

Figure 3: Accuracy analysis

Table 2: Testing Accuracy

Dataset	40%	50%	60%	70%	80%
Liner model	93.5 ± 1.1	93.9 ± 1.3	92.6 ± 1.6	95.9 ± 1.0	93.6 ± 1.8
Neural Network	94.2 ± 1.0	94.7 ± 1.3	93.4 ± 2.1	96.3 ± 1.1	94.3 ± 1.6
SVM	95.8 ± 0.7	96.7 ± 0.9	94.1 ± 1.0	98.3 ± 0.8	95.2 ± 1.8
J48	96.1 ± 0.9	97.1 ± 1.3	94.6 ± 1.8	98.5 ± 0.8	95.2 ± 2.0
Random Forest	96.9 ± 0.7	98.0 ± 0.8	96.1 ± 0.8	99.0 ± 0.7	97.4 ± 1.1
Proposed	98.1 ± 0.8	98.4 ± 0.8	98.2 ± 0.9	99.1 ± 0.7	98.2 ± 0.6

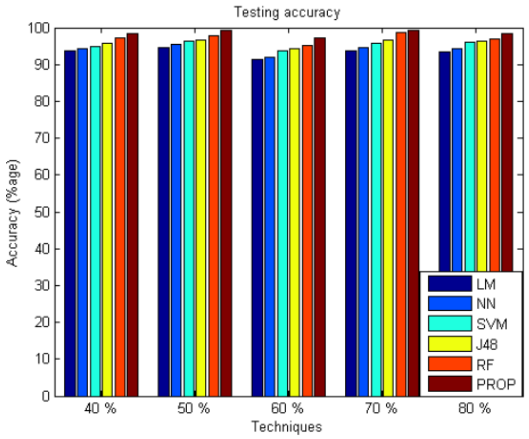

Figure 4: Accuracy analysis

2. Correlation

A table 3 and 4 shows the correlation analysis between proposed method and others. As identified in previous, correlation lies between [-1 1] and positive correlation approaches to 0 indicate that the proposed method provides significant results over other methods. Therefore, from Tables 4 and 5 it has been observed that the proposed method provides more significant results compared to earlier approaches.

Table 3: Training Correlation

Dataset	40%	50%	60%	70%	80%
Liner model	0.91 ± 0.04	0.90 ± 0.02	0.86 ± 0.09	0.86 ± 0.05	0.87 ± 0.07
Neural Network	0.92 ± 0.01	0.91 ± 0.03	0.87 ± 0.09	0.87 ± 0.06	0.88 ± 0.05
SVM	0.93 ± 0.02	0.92 ± 0.03	0.88 ± 0.06	0.88 ± 0.05	0.89 ± 0.04
J48	0.94 ± 0.03	0.93 ± 0.04	0.89 ± 0.06	0.89 ± 0.07	0.90 ± 0.05
Random Forest	0.94 ± 0.03	0.94 ± 0.04	0.90 ± 0.05	0.90 ± 0.05	0.92 ± 0.04
Proposed	0.97 ± 0.02	0.96 ± 0.03	0.92 ± 0.04	0.94 ± 0.05	0.95 ± 0.04

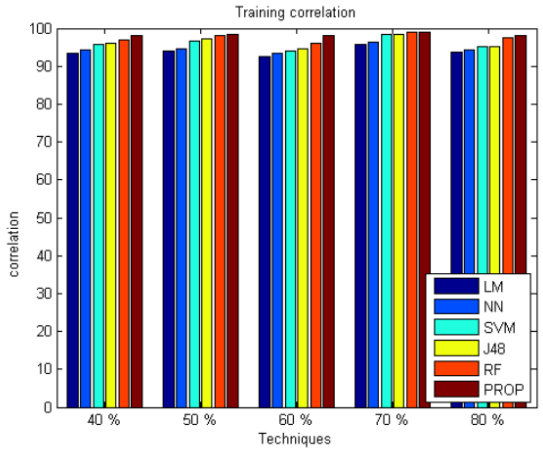

Figure 5: Training correlation analysis

Table 4: Testing Correlation

Dataset	40%	50%	60%	70%	80%

Liner model	0.84 ± 0.09	0.87 ± 0.04	0.85 ± 0.11	0.92 ± 0.06	0.93 ± 0.04
Neural Network	0.85 ± 0.10	0.88 ± 0.09	0.86 ± 0.11	0.93 ± 0.05	0.94 ± 0.04
SVM	0.86 ± 0.11	0.90 ± 0.08	0.87 ± 0.10	0.94 ± 0.03	0.89 ± 0.08
J48	0.87 ± 0.11	0.91 ± 0.07	0.88 ± 0.10	0.91 ± 0.07	0.92 ± 0.06
Random Forest	0.89 ± 0.09	0.92 ± 0.06	0.90 ± 0.07	0.96 ± 0.03	0.98 ± 0.01
Proposed	0.95 ± 0.04	0.96 ± 0.03	0.96 ± 0.02	0.98 ± 0.01	0.98 ± 0.01

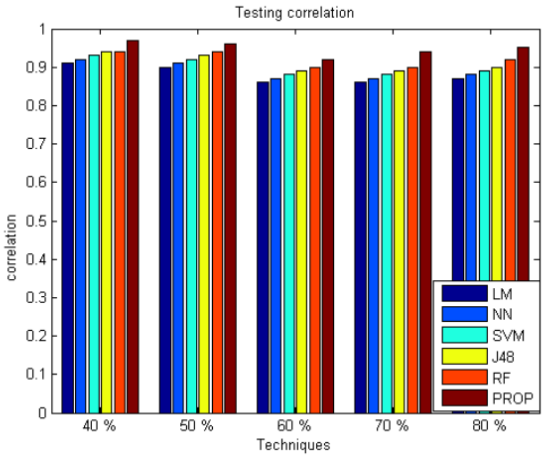

Figure 6: Testing correlation analysis

3. Root Means Squared Error (RMSE)

Tables 5 and 6 demonstrate the root means squared error (RMSE) analysis between proposed and other machine learning approaches. RMSE represent the difference among actual and predicted *C*6*H*6 values. Therefore, it should be minimum. From Tables 6 and 7 it has been observed that the proposed method provide lesser RMSE compared to others, therefore proposed method provides more significant *C*6*H*6 results.

Table 5: Training RMSE

Dataset	40%	50%	60%	70%	80%
Liner model	3.2 ± 0.63	3.9 ± 0.88	3.7 ± 0.44	2.9 ± 0.73	5.8 ± 0.49
Neural Network	4.4 ± 0.77	4.1 ± 0.81	4.1 ± 0.62	6.8 ± 0.58	5.9 ± 0.64
SVM	6.0 ± 0.82	4.5 ± 0.68	4.3 ± 0.57	4.7 ± 0.87	5.2 ± 0.91
J48	3.3 ± 0.95	4.6 ± 0.86	5.0 ± 0.78	5.6 ± 0.69	4.5 ± 0.72
Random Forest	2.0 ± 0.45	3.7 ± 0.43	2.4 ± 0.39	2.2 ± 0.47	3.3 ± 0.49
Proposed	1.7 ± 0.31	2.0 ± 0.39	1.1 ± 0.27	1.2 ± 0.29	2.2 ± 0.41

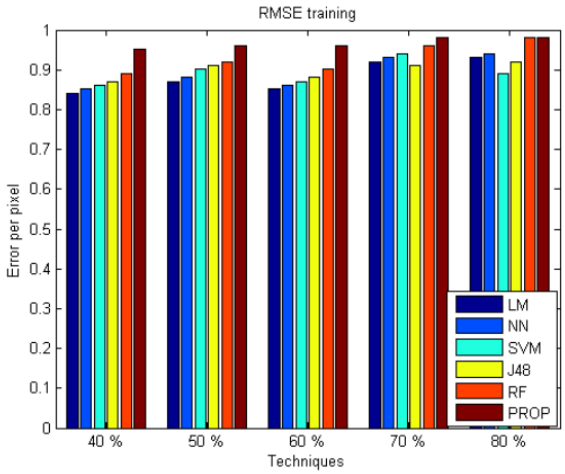

Figure 7: Testing Root mean square error analysis

Table 6: Testing RMSE

Dataset	40%	50%	60%	70%	80%
Liner model	4.4 ± 0.65	3.9 ± 0.61	3.8 ± 0.45	4.9 ± 0.47	5.0 ± 0.57
Neural Network	6.3 ± 0.39	6.5 ± 0.48	5.4 ± 0.51	3.5 ± 0.55	4.4 ± 0.41
SVM	4.0 ± 0.49	4.8 ± 0.45	4.1 ± 0.39	4.7 ± 0.38	4.6 ± 0.29
J48	5.3 ± 0.34	4.3 ± 0.29	4.1 ± 0.37	5.7 ± 0.26	4.8 ± 0.29
Random Forest	3.9 ± 0.23	2.9 ± 0.31	2.4 ± 0.29	2.5 ± 0.34	3.0 ± 0.27
Proposed	2.5 ± 0.24	1.9 ± 0.27	1.6 ± 0.19	1.9 ± 0.28	2.6 ± 0.32

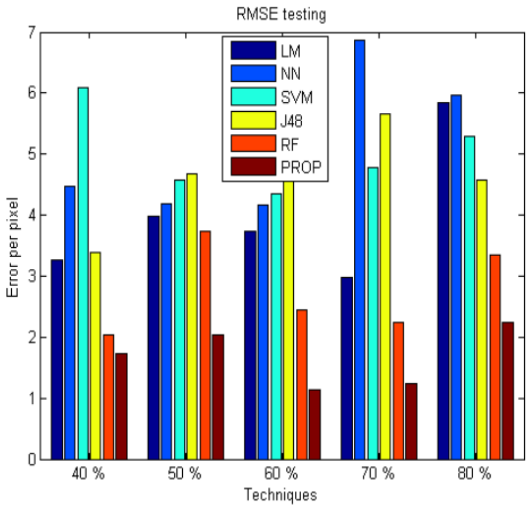

Figure 8: Testing Root mean square error analysis

4. Execution Time

Tables 7 and 8 shows computational time taken by proposed and other methods during training time and testing time in seconds, respectively. The computational time should minimum. However, proposed method take more time compared to other methods, therefore not so efficient in terms of running time.

Table 7: Training Execution Time

Dataset	40%	50%	60%	70%	80%
Liner model	94.4 ± 9.7	104.3 ± 13.7	96.5 ± 12.9	114.0 ± 11.6	107.6 ± 9.7
Neural Network	115.4 ± 9.6	94.4 ± 13.4	103.1 ± 12.4	93.9 ± 11.7	103.7 ± 9.3
SVM	120.6 ± 7.2	111.7 ± 8.4	92.6 ± 11.1	118.5 ± 8.3	80.5 ± 19.7
J48	89.6 ± 13.3	99.4 ± 12.4	90.9 ± 13.2	94.3 ± 11.7	93.7 ± 14.8
Random Forest	83.2 ± 15.3	78.8 ± 16.3	83.4 ± 13.4	82.6 ± 12.7	77.9 ± 11.6
Proposed	141.9 ± 4.7	157.6 ± 5.3	162.1 ± 9.3	171.3 ± 9.4	176.6 ± 9.7

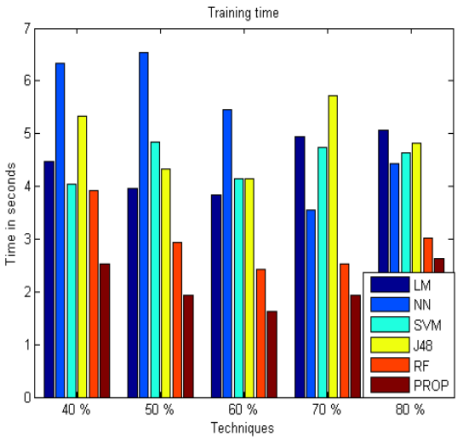

Figure 9: Training Execution time analysis

Table 8: Testing Execution Time

Dataset	40%	50%	60%	70%	80%
Liner model	88.3 ± 17.6	116.1 ± 14.9	115.7 ± 14.8	86.3 ± 15.7	90.7 ± 13.3
Neural Network	87.2 ± 10.7	107.7 ± 11.8	89.4 ± 9.7	81.0 ± 10.4	114.6 ± 9.9
SVM	103.4 ± 11.2	82.9 ± 12.9	101.3 ± 10.4	106.5 ± 11.8	116.6 ± 10.7
J48	81.9 ± 11.4	106.2 ± 14.6	85.0 ± 9.8	104.3 ± 13.1	88.4 ± 12.4
Random Forest	79.3 ± 9.1	80.2 ± 10.7	79.9 ± 7.3	78.4 ± 8.9	85.7 ± 10.4
Proposed	148.0 ± 7.4	159.0 ± 6.9	165.6 ± 5.4	172.1 ± 8.2	184.5 ± 7.9

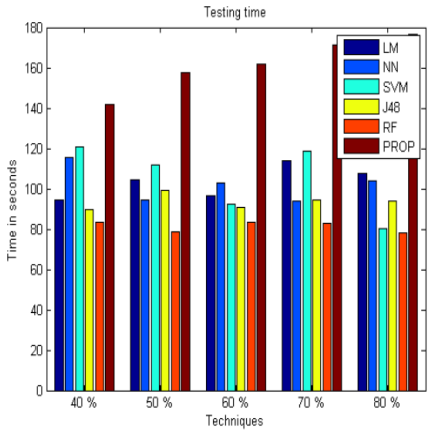

Figure 10: Testing Execution time analysis

5. Coefficient regarding determination (*R*)

Tables 9 and 10 depict the evaluation analysis regarding Coefficient regarding determination (*R*). It is the quotient on the variances of the installed beliefs plus witnessed beliefs on the dependent variable. R is a statistic which will provide some good info concerning the rewards regarding fit of the model. Within regression, the R is a statistical measure of how good the regression range approximates the genuine data points. *A R* regarding 1 signifies the fact that regression range correctly suits the data. From Tables 9 and 10 has been observed that the prosed method has better *R* as compared to other methods.

Table 9: Training Coefficient regarding determination (*R*)

Dataset	40%	50%	60%	70%	80%
Liner model	0.81 ± 0.08	0.82 ± 0.09	0.78 ± 0.12	0.76 ± 0.11	0.78 ± 0.08
Neural Network	0.82 ± 0.07	0.80 ± 0.06	0.79 ± 0.05	0.78 ± 0.07	0.79 ± 0.07
SVM	0.83 ± 0.05	0.83 ± 0.06	0.80 ± 0.05	0.79 ± 0.06	0.79 ± 0.07
J48	0.83 ± 0.05	0.84 ± 0.06	0.81 ± 0.04	0.80 ± 0.05	0.81 ± 0.06
Random Forest	0.84 ± 0.05	0.83 ± 0.05	0.82 ± 0.04	0.82 ± 0.06	0.83 ± 0.06
Proposed	0.86 ± 0.03	0.84 ± 0.03	0.83 ± 0.04	0.84 ± 0.05	0.85 ± 0.04

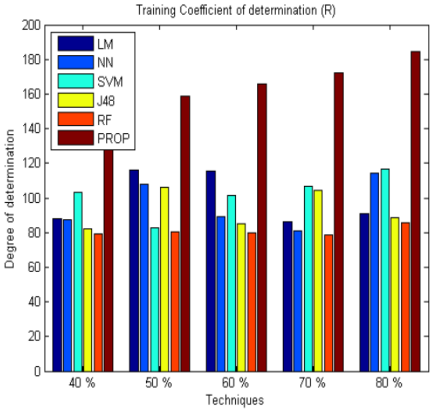

Figure 11: Analysis of Training Coefficient of determination (R)

Table 10: Testing Coefficient regarding determination (*R*)

Dataset	40%	50%	60%	70%	80%
Liner model	0.74 ± 0.08	0.72 ± 0.04	0.73 ± 0.07	0.71 ± 0.05	0.72 ± 0.04
Neural Network	0.75 ± 0.07	0.74 ± 0.04	0.76 ± 0.06	0.73 ± 0.06	0.74 ± 0.05
SVM	0.77 ± 0.07	0.75 ± 0.06	0.74 ± 0.08	0.73 ± 0.06	0.76 ± 0.04
J48	0.81 ± 0.07	0.79 ± 0.06	0.78 ± 0.07	0.76 ± 0.07	0.77 ± 0.06
Random Forest	0.84 ± 0.04	0.82 ± 0.06	0.84 ± 0.07	0.85 ± 0.03	0.86 ± 0.04
Proposed	0.85 ± 0.04	0.86 ± 0.06	0.87 ± 0.02	0.87 ± 0.04	0.88 ± 0.04

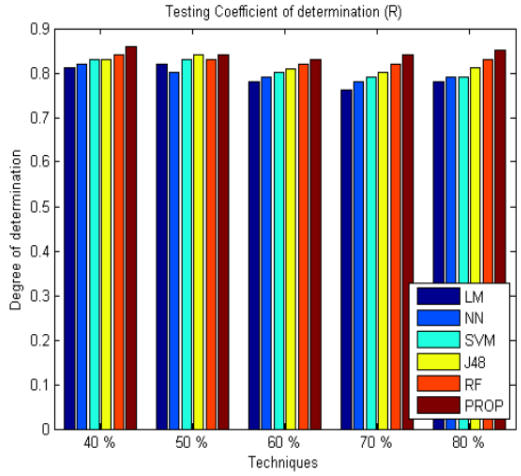

Figure 14: Analysis of Testing Coefficient of determination(R)

5.CONCLUSIONS AND FUTURE WORK

Human expertise of benzene is from a range of discerning and long-term unfavourable well being benefits and ailments, which include most cancers and aplastic anaemia. Direct exposure can happen occupationally and domestically on account of this huge using benzene-containing petroleum items, which include motor unit powers and solvents. Effective and passive expertise of cigarette can also be a important method of obtaining exposure. Benzene is especially erratic, and exposure comes about generally by means of inhalation. Public well being measures are needed to lessen the exposure connected with either employees and the typical human population to help benzene. Benzene (*C*6*H*6), simplest natural, great smelling hydrocarbon and also parent element of several significant great smelling compounds. Benzene is actually a colourless liquid which has a characteristic smell and is also principally utilised in producing polystyrene. The idea is extremely harmful and is also a regarded carcinogen; contact it may cause leukemia. Therefore, there are rigorous handles upon benzene emissions. The use of ensembling of data mining techniques have been done to improve the accuracy rate Benzene (*C*6*H*6) detection machine learning techniques. It has been achieved by using the integration of random forest and J48 based machine learning techniques. Root mean squared error (RMSE) tuning has also been achieved. Initially, to evaluate the performance of existing machine learning techniques for detection of *C*6*H*6. Then, proposed technique is implemented to evaluate the *C*6*H*6 in air. Then, comparisons have been done between the existing machine learning algorithms and proposed technique using:a) Root mean squared error, b) Correlation, c) Accuracy,
d) Error rate and e) Coefficient of determination.

Future work

Subsequent section describes future directions for the proposed work:

1. Ensembling of Random forest and J48 based machine learning technique does notguarantee the lowest error rate because random forest is limited to number of trees only. Therefore, in near future meta-heuristic techniques such as ant colony optimization, artificial bee colony etc. approaches will be considered to enhance the results further.

2. Also, in near future proposed technique will be applied on other fields such as biomedical processing, image machine learning etc. to evaluate the performance of the proposed technique for other applications.

3. Also, proposed technique will be applied on real-time data taken from sensors.

REFERENCE

[1]. Airbase - the european air quality database. http: //www.eea.europa.eu/data-and-maps/data/airbase-the-european-air-quality-database-8. Accessed: March 17, 2017.

[2]. C6H6pubchem open chemistry database. https://pubchem.ncbi.nlm. nih.gov/compound/C6H6. Accessed: March 17, 2017.

[3]. C6H6-B.csv national institute of standards and technology. https://www. nist.gov/file/36031. Accessed: March 17, 2017.

[4]. C6H6-nrm-part5.test.csv petravidnerovasensorsscikittest. https://github.com/PetraVidnerova/SensorsScikitTest/blob/master / data/C6H6-nrm-part5.test.csv. Accessed: March 17, 2017.

[5]. Shahid Ali and Sreenivas Sremath Tirumala. Performance analysis of svm ensemble methods for air pollution data. In *Proceedings of the 8th InternationalConference on Signal Processing Systems*, pages 212–216. ACM, 2016.

[6]. Yun Bai, Yong Li, Xiaoxue Wang, JingjingXie, and Chuan Li. Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions. *Atmospheric pollution research*, 7(3):557–566, 2016.

[7]. IliasBougoudis, Konstantinos Demertzis, and LazarosIliadis. Fast and low cost prediction of extreme air pollution values with hybrid unsupervised learning. *Integrated Computer-Aided Engineering*, 23(2):115–127, 2016.

[8]. IliasBougoudis, Konstantinos Demertzis, and LazarosIliadis. Hisycol a hybrid computational intelligence system for combined machine learning: the case of air pollution modeling in athens. *Neural Computing and Applications*, 27(5):1191– 1206, 2016.

[9]. Cole Brokamp, Roman Jandarov, M.B. Rao, Grace LeMasters, and Patrick Ryan. Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches. *Atmospheric Environment*, 151:1 – 11, 2017.

[10]. Ling Chen, Yaya Cai, Yifang Ding, MingqiLv, Cuili Yuan, and Gencai Chen. Spatially fine-grained urban air quality estimation using ensemble semi-supervised learning and pruning. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 1076–1087. ACM, 2016. 38

[11]. S De Vito, E Massera, M Piga, L Martinotto, and G Di Francia. On field calibration of an electronic nose for C6H6 estimation in an urban pollution monitoring scenario. *Sensors and Actuators B: Chemical*, 129(2):750–757, 2008.

[12]. Saverio De Vito, Grazia Fattoruso, Matteo Pardo, Francesco Tortorella, and Girolamo Di Francia. Semi supervised learning methods in artificial olfaction: A novel approach to classification problems and drift counteraction. *IEEE SensorsJournal*, 12(11):3215–3224, 2012.

[13]. Saverio De Vito, Marco Piga, Luca Martinotto, and Girolamo Di Francia. Co, no2 and nox urban pollution monitoring with on field calibrated electronic nose

[14]. Qingli Dong, Yong Wang, and Peizhi Li. Multifractal behavior of an air pollutant time series and the relevance to the predictability. *Environmental Pollution*, 222:444–457, 2017.

[15]. Husanbir Singh Pannu, Dilbag Singh, and Avleen Kaur Malhi. "Improved particle swarm optimization based adaptive neuro-fuzzy inference system for benzene detection." CLEAN–Soil, Air, Water (2018).

[16]. Eleni Fotopoulou, Anastasios Zafeiropoulos, Dimitris Papaspyros, Panagiotis Hasapis, George Tsiolis, ThanassisBouras, SpyrosMouzakitis, and Norma Zanetti. Linked data analytics in interdisciplinary studies: The health impact of air pollution in urban areas. *IEEE Access*, 4:149–164, 2016.

[17]. Yoav Freund and Llew Mason. The alternating decision tree learning algorithm. In *icml*, volume 99, pages 124–133, 1999.

[18]. N. Goudarzi, D. Shahsavani, F. Emadi-Gandaghi, and M. Arab Chamjangali. Application of random forests method to predict the retention indices of some polycyclic aromatic hydrocarbons. *Journal of Chromatography A*, 1333:25 – 31, 2014.

[19]. Hui Hu, Sandie Ha, Jeffrey Roth, Greg Kearney, Evelyn O. Talbott, and XiaohuiXu. Ambient air pollution and hypertensive disorders of pregnancy: A systematic review and meta-analysis. *Atmospheric Environment*, 97:336 – 345, 2014.

[20]. Ibrahim Anwar Ibrahim and Tamer Khatib. A novel hybrid model for hourly global solar radiation prediction using random forests method and firefly algorithm. *Energy Conversion and Management*, 138:413 – 425, 2017. 39

[21]. Younoh Kim, Scott Knowles, James Manley, and Vlad Radoias. Long-run health consequences of air pollution: Evidence from indonesia's forest fires of 1997. *Economics & Human Biology*, 26:186 – 198, 2017.

[22]. Ibai Lana, Javier Del Ser, Ales Padro, Manuel Velez, and Carlos Casanova-Mateo. The role of local urban traffic and meteorological conditions in air pollution: A data-based case study in madrid, spain. *Atmospheric Environment*, 145:424 – 438, 2016.

[23]. SirkkuManninen, Vitali Zverev, Igor Bergman, and Mikhail V. Kozlov. Consequences of long-term severe industrial pollution for aboveground carbon and nitrogen pools in northern taiga forests at local and regional scales. *Science of TheTotal Environment*, 536:616 – 624, 2015.

[24]. Nguyen Thi Trang Nhung, HereshAmini, Christian Schindler, MeltemKutlar Joss, Tran Minh Dien, Nicole Probst-Hensch, Laura Perez, and Nino KÃ 1 4nzli. Shortterm association between ambient air pollution and pneumonia in children: A systematic review and meta-analysis of time-series and case-crossover studies. *Environmental Pollution*, 230:1000 – 1008, 2017.

[25]. PJ García Nieto, Elías F Combarro, JJ del Coz Díaz, and Elena Montañés. A svm-based regression model to study the air quality at local scale in oviedo urban area (northern spain): A case study. *Applied Mathematics and Computation*, 219(17):8923–8937, 2013.

[26]. Pavel G Polishchuk, Eugene N Muratov, Anatoly G Artemenko, Oleg G Kolumbin, Nail N Muratov, and Victor E Kuzmin. Application of random forest approach to qsar prediction of aquatic toxicity. *Journal of chemical information and modeling*, 49(11):2481–2488, 2009.

[27]. Kanchan Prasad, Amit Kumar Gorai, and Pramila Goyal. Development of anfis models for air quality forecasting and input optimization for reducing the computational cost and time. *Atmospheric environment*, 128:246–262, 2016.

[28]. Dilbag Singh and Vijay Kumar. "Modified gain intervention filter based dehazing technique." Journal of Modern Optics 64, no. 20 (2017): 2165-2178.

[29]. Dilbag Singh and Vijay Kumar. "Comprehensive survey on haze removal techniques." Multimedia Tools and Applications (2017): 1-26.

[30]. Dilbag Singh and Vijay Kumar. "Dehazing of remote sensing images using improved restoration model based dark channel prior." The Imaging Science Journal 65, no. 5 (2017): 282-292.

[31]. Dilbag Singh and Vijay Kumar. "Dehazing of remote sensing images using fourth-order partial differential equations based trilateral filter." IET Computer Vision (2017).

[32]. Dilbag Singh and Vijay Kumar. "Defogging of road images using gain coefficient-based trilateral filter." Journal of Electronic Imaging 27, no. 1 (2018): 013004.

[33]. Dilbag Singh and Vijay Kumar. "Single image haze removal using integrated dark and bright channel prior." Modern Physics Letters B (2018): 1850051.

[34]. Husanbir Singh Pannu, Dilbag Singh, and Avleen Kaur Malhi. "Multi-objective particle swarm optimization-based adaptive neuro-fuzzy inference system for benzene monitoring." Neural Computing and Applications: 1-11.

[35]. Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining:Practical machine learning tools and methods*. Morgan Kaufmann, 2016.

[36]. John Yen and Reza Langari. *Fuzzy logic: intelligence, control, and information*. Prentice-Hall, Inc., 1998.

[37]. Yilmaz Yildirim and MahmutBayramoglu. Adaptive neuro-fuzzy based modelling for prediction of air pollution daily levels in city of zonguldak. *Chemosphere*,63(9):1575–1582, 2006.

[38]. Yongheng Zhao and Yanxia Zhang. Comparison of decision tree methods for finding active objects. *Advances in Space Research*, 41(12):1955–1959, 2008.

Authors Profile

Mrs. Gagandee Kaur pursed Bachelor of Engineering in Computer Science from CSVTU, Chhattisgarh . She is currently pursuing M.Tech in Computer Science from Sri Sai college of Engineering and Technology, Manawala, Amritsar .

Ms Harmanpreet Kaur pursed Bachelor of Engineering in Computer Science and Master of Science from Rayat Institute of Engineering and Information Technology,Ropar . She is currently working as Assistant Professor in Department of Computer, Sri Sai college of Engineering and Technology, Manawala, Amritsar .

