Global J. Environ. Sci. Manage. 5(3): 309-318, Summer 2019

Global Journal of Environmental Science and Management
(GJESM)

Homepage: https://www.gjesm.net/

ORIGINAL RESEARCH PAPER

Application of ensemble learning techniques to model the atmospheric concentration of SO_2

A. Masih

Department of System Analysis and Decision Making, Ural Federal University, Ekaterinburg, Russian Federation

ARTICLE INFO

Article History:
Received 13 January 2019
Revised 15 April 2019
Accepted 07 May 2019

Keywords:
Air Pollution Modeling
Ensemble Learning Techniques
Multilayer Perceptron (MLP)
Random Forest
Bagging
Sulphur Dioxide (SO_2)
Support Vector Machine (SVM)
Voting

ABSTRACT

In view of pollution prediction modelling, the study adopts homogenous (random forest, bagging and additive regression) and heterogeneous (voting) ensemble classifiers to predict the atmospheric concentration of Sulphur dioxide. For model validation, results were compared against widely known single base classifiers such as support vector machine, multilayer perceptron, linear regression and regression tree using M5 algorithm. The prediction of Sulphur dioxide was based on atmospheric pollutants and meteorological parameters. While, the model performance was assessed by using four evaluation measures namely Correlation coefficient, mean absolute error, root mean squared error and relative absolute error. The results obtained suggest that 1) homogenous ensemble classifier random forest performs better than single base statistical and machine learning algorithms; 2) employing single base classifiers within bagging as base classifier improves their prediction accuracy; and 3) heterogeneous ensemble algorithm voting have the capability to match or perform better than homogenous classifiers (random forest and bagging). In general it demonstrates that the performance of ensemble classifiers random forest, bagging and voting can outperform single base traditional statistical and machine learning algorithms such as linear regression, support vector machine for regression and multilayer perceptron to model the atmospheric concentration of sulphur dioxide.

DOI: 10.22034/gjesm.2019.03.04

NUMBER OF REFERENCES	NUMBER OF FIGURES	NUMBER OF TABLES
44	6	4

*Corresponding Author:
Email: *adven.masikh@urfu.ru*
Phone: +79655100302
Fax: +79655100302

Note: Discussion period for this manuscript open until October 1, 2019 on GJESM website at the "Show Article.

INTRODUCTION

Considering the socio-economic ambiance of a territory, the health conditions of citizens are vastly dependent on its sanitary and ecological setting (Salnikov and Karatayev, 2011). Due to a vital connection between human health and air quality, atmospheric pollution level remains one of the several main aspects that influence human physical development. The consequences of inhaling contaminated air through bronchial tubes and windpipe can adversely affect alveoli – the point where dirt enters blood and lymph (Seinfild *et al.,* 1998). World Health Organization (WHO, 2014) and Samoli *et al.,* 2015 have drawn a strong correlation between health apprehensions (cardiovascular system, respiratory skin diseases etc.) and increased concentration of air pollutants i.e. oxides of nitrogen (NO_x), carbon monoxide (CO), sulphur dioxide (SO_2), ground level ozone (O_3). A constant monitoring of atmospheric pollutants is important because the introduction of chemicals and particulate matter (PM) in atmosphere beyond a certain limit results in an air pollution (Brunekreef, 2002). Due to negative impact of air pollutants on human health, other living organisms, crops and natural environment, it is one of many vital issue that metropolitan and industrial cities need to address (Masih, 2018a). Generally six air pollutants namely particulate matter ($PM_{2.5}$ and PM_{10}), SO_2, NO_2, CO, and O_3 are used to calculate Air Quality Index (AQI) that describes the pollution level of a region. Increased concentrations of SO_2 can be extremely dangerous due to its: nasty smell; ability to quickly react with suspended particles in air to form harmful acids which can cause acid rain; and most importantly short exposure to SO_2 can aggravate human respiratory system which make breathing difficult (Brunekreef, 2002; Xie *et al.,* 2016). Therefore, apart from a strict monitoring, practical significance demands the development of forecast models that can accurately predict the atmospheric concentration of SO_2. Chemical transport models and data driven statistical techniques are two major types of modelling approaches employed in atmospheric science to estimate and forecast the concentrations of air pollutants. However, applicability of chemical models is limited due to several reasons such as: a thorough understanding of transportation; its chemical mixing and transformation details; which is complex to investigate about; computation cost; and a complete list of air pollutants with concentration values which is difficult to get (Zhan, 2018). Due to limitations of chemical models, nowadays data driven approaches are catching the attention of a number of researchers to model atmospheric pollutants. With technological progression the modelling techniques are also progressing to be more efficient. For example recent environment science and engineering researches (Abdul-Wahab and Al-Alawi, 2002; Schlink, *et al.,* 2003; Chaloulakou *et al.,* 2003; Grivas and Chaloulakou, 2006; Baawain and Al-Serihi, 2014; Bedoui *et al.,* 2016) (Juhos et al., 2008; Wang et al., 2008; Rahimi, 2017) using machine learning techniques such as artificial neural network (ANN), Random Forest (RF), SVM etc. have been preferred over classical statistical methods due to their enhanced performances. Although, machine learning tools are able to deal with non-linearity of inventory data however, literature suggests that ANNs undergo from a problem of over fitting and local minima (Brunelli *et al.,* 2007). Whereas advanced machine learning techniques like ensemble learning (homogenous and heterogeneous) algorithms are capable of dealing with issues like local minima and overfitting. Therefore, RF, Bagging and Voting can be applied as alternative approaches (Yu *et al.,* 2016; Nawahda, 2016; Masih, 2018b). Literature review conducted in the context of this study suggests that, several studies in the field of air quality modelling have recently adopted ANNs for classification and regression purposes to predict the concentration of air pollutants such as NO_x, SO_2, O_3 etc. (Shaban *et al.,* 2016; Gardner and Dorling, 1999; Russo and Soares, 2014; Capilla, 2014; Lu *et al.,* 2003; Singh *et al.,* 2012). These articles confirm the ability of ANN based algorithms to handle non-linearity and complexity of the emission datasets, but also point out the problems of local minima and over-fitting, ANNs suffer from. Few attempts were made to overcome these problems, but the authors suggested that both cannot be solved simultaneously (Lu *et al.,* 2003; Wang and Lu, 2006). Later, the SVM performance was tested out against ANN based algorithm MLP by Lu and Wang, (2014) showed that on structural issues SVM can perform better than MLP. The work was considered a milestone in the field of atmospheric pollution prediction for further development. Similarly another study based on Athens Greece developed 84 different models and established that Tree and Rule classification algorithms perform significantly better than that of SVM and linear

regression (LR) (Riga *et al.,* 2009). Meanwhile, a comprehensive application of ensemble learning techniques in the field atmospheric modelling was considered by Singh *et al.,* 2013. The authors proposed an air quality model using principal component analysis (PCA) and ensemble learning algorithms – Bagging and Boosting for pollution forecasting. PCA was aimed at identifying the pollution sources. The study resulted in a significantly enhanced performance accuracy when compared with SVM. Similarly Cannon and Lord, (2000) built a model which used MLP and multilayer regression (MLR) as single base classifiers during the first phase to predict the maximum average of ground level ozone during the daytime. Whereas in second phase both classifiers were adopted within Bagging as base classifiers. The experiment results obtained suggest that, both classifiers have suffered from over-fitting and instability when used as independent learners, while adopting them as base classifiers within bagging both resulted in an enhanced stability and improved accuracy. Which confirms its ability to deal with over-fitting and performance instability. The utility of homogenous ensemble learning algorithms RF in atmospheric sciences has been explored in a Sydney based study conducted by Jiang and Riley, (2015). The authors developed two classification models for performance comparison between RF and classification and regression tree (M5P). The study concluded that RF achieves better accuracy as compared to single base tree. While another RF based approach (Yu *et al.,* 2016) have lately been constructed to predict AQI value by using urban public data based on road information, air quality and meteorological datasets in different regions of Shenyang. Upon a proficient assessment of the model, it was verified that RF can outperform state of the art classifiers such as Naïve Bayes, Logistic Regression, single decision tree and ANN. Ensemble learning algorithms work on a principle of developing multiple models, later when ensemble, result in an improved performance as compared to single based models. Discussing their applications, ensemble algorithms have successfully been employed in multiple fields such as finance, bioinformatics, computer security, marketing, and power for load prediction (Alfaro *et al.,* 2008; Gabralla and Abraham, 2014; Fathima *et al.,* 2014; Yang *et al.,* 2010; Tüfekci, 2014; Van Loon, *et al.,* 2007). Given these observations, the study has the following contributions: it draws a comparison between single base and ensemble learning classifiers; the results obtained were further tested as base classifiers, first within Bagging – a homogeneous ensemble learning technique and then within voting – a heterogeneous ensemble learning technique, with an aim to investigate about the best machine learning algorithm that can predict the atmospheric concentration of SO_2 with high precision by using emission and meteorological dataset. This study has been carried out in Russian Federation in 2019.

MATERIALS AND METHODS

The study uses meteorological and atmospheric gas concentration data, obtained from the official website (https://uk-air.defra.gov.uk) of Department of Environment Food and Rural Affairs (DEFRA). The dataset was recorded during 1 January to 8 May, 2013 at a sampling rate of one hour near Marylebone road located in London, United Kingdom. The descriptive statistics of 9 attributes including 4 atmospheric pollutants (SO_2, NO_2, CO, and HCl) and 5 meteorological parameters (temperature, wind speed, wind direction, relative humidity, and atmospheric pressure) is presented in Table 1.

All analyses were performed in an open tool kit for data analysis known as *Waikato Environment for*

Table 1: Descriptive statistics of air pollutants and meteorological parameters

Attribute	Unit	Valid N	Mean	Minimum	Maximum	Standard deviation
NO_2	$\mu g/m^3$	2768	79.57	7.25	206.2	38.42
SO_2	$\mu g/m^3$	2768	45.2	0.47	141.1	40.3
CO	$\mu g/m^3$	2768	106.04	1.27	436.78	79.54
HCl	$\mu g/m^3$	2768	4.57	2.00	101.03	1.95
Temperature	°C	2768	7.02	-9.2	28.1	6.22
Relative humidity	%	2768	66.8	45.0	101.57	6.98
Pressure	mmHg	2768	747.51	101.0	757.6	13.91
Wind speed	m/s	2768	3.83	0.1	10.1	1.79
Wind direction	Degree	2768	169.68	0.1	360	107.93

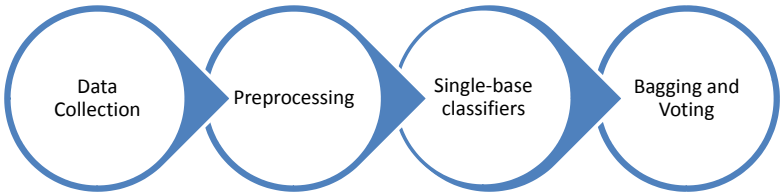

Fig 1: Data processing and modelling scheme

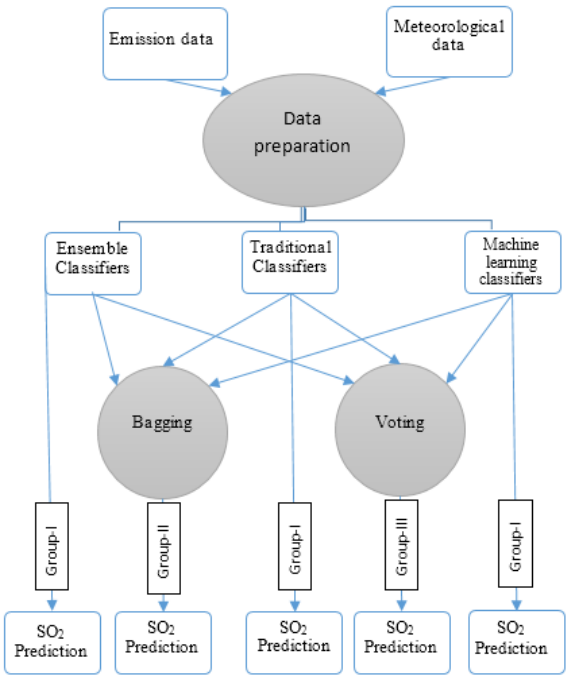

Fig 2: Experiment setting

Knowledge Analysis (WEKA). Data analysis conducted for this paper has several stages involved: 1) data collection; 2) data preprocessing; 3) Single base vs homogenous ensemble learning classifiers; 4) Bagging and Voting as explained in Fig. 1. Several observations were made during preliminary analysis of the dataset such as: 1) no clear pattern in dataset was found; 2) several missing values were observed in meteorological parameters; 3) few extreme values were recorded in both pollutant and meteorological features. Thus data needed a thorough cleaning. To do so, during data preprocessing, incomplete and inconsistence instances i.e. missing values with discrepancies were detected and replaced using imputation method. Similarly the noisy data containing outliers were removed by using WEKA built-in filter named interquartile range. Furthermore, a data transformation of wind direction was performed to make sure that 0 and 360 degrees are treated as one value. For that, Wind Speed (WS) and Wind Direction (WD) were combined in the form of two new orthogonal components: U=WS*cos (WD) and V= WS*sin (WD) to replace WS and WD. Lastly, the prepared dataset of 2768 instances were used for pollution modelling. The experiment design of the study is presented in Fig. 2. The work is divided into three different sets of experiments i.e. group-I, group-II and group-III. For group-I experiments altogether 7 models including 3 machine learning single base learners (MLP, SVM, M5P), 1 traditional statistical model (LR), and 3 homogenous ensemble learning models (RF, AR, RSS) were developed. To predict the concentration of SO_2, a comprehensive comparison was drawn between homogenous ensemble learning approaches and independent learners (i.e. single base learner algorithms). A single base classifiers are the algorithms which follow the basic rules machine learning. These algorithms take a training dataset and apply only one of the machine learning algorithm to build a prediction model e.g. LR, MLP, and SVM etc. Whereas, ensemble learning approaches usually use multiple algorithms to build a model. For group-II and III experiments, the study employs a meta-conformal ensemble approach (Balasubramanian *et al.,* 2014). The technique works on a principle of combining a base classifier h∈H with a meta-classifier m∈M to create a new combined classifier h:m which is more accurate than its base classifiers. An ensemble combines a series of k-learned models (m_1, m_2, m_3,...., m_k) with an aim to create an improved composite classification model (Balasubramanian *et al.,* 2014). Considering the concept of composite modelling, during group-II and III experiments, all group-I algorithms were combined with meta-learners – bagging and voting respectively as shown in Fig. 2. For model configuration, concentration of air pollutants as well as environmental parameters were tested one by one to predict the atmospheric concentration of SO_2

in form of R^2. The model performances were assessed under an experimental design of 80% training and 20% test data. To evaluate model performance four widely known scales: correlation coefficient (R^2), mean absolute error (MAE), root mean square error (RMSE), and relative absolute error (RAE) were used, their formulae are shown in Eqs. 1 to 4.

$$R^2 = \frac{\sum_i \left[(x_i - \bar{x})(y_i - \bar{y})\right]}{\left[\left(\sum_i x_i - \bar{x}\right)^2 \left(\sum_i y_i - \bar{y}\right)^2\right]^{1/2}} \quad (1)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i| \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2} \quad (3)$$

$$RAE = \frac{\left|\sum_i^n y_i - x_i\right|}{\left|\sum_i^n \bar{y} - x_i\right|} \quad (4)$$

Where, y_i and x_i are predicted and observed values respectively, $\bar{x}$ is the average predicted value and n is total number of instances.

RESULTS AND DISCUSSION

The successful application of ensemble learning techniques in several fields is an evidence that its applications in the field of environmental science and engineering is inevitable. Therefore, for SO_2 prediction, the performance accuracy of ensemble learning algorithms (RF, bagging and voting) have been assessed by using atmospheric pollution and meteorological data. Whereas for model validation the results were compared against other popularly known classifiers SVM and MLP. For experiment settings "Explorer"– a working environment implemented in WEKA was used. Besides applying WEKA implemented optimization algorithm named cross-validation parameter selection (CVPS) on all classifiers, for a fair comparison, the selected classifiers (SVM and MLP) were fine-tuned during preprocessing. In order to determine the optimal number of hidden layers for MLP, a range of experiments using 1 to 20 hidden layers were tested out. It can be seen in Fig. 3 that the MLP achieved the highest accuracy (R^2=0.93) at three different occasions when the hidden layers were 7, 14 and 20. However, the study adopts 14 hidden layers because the error value obtained using 14 hidden neurons was considerably low (RAE = 44.65%) as compared to 7 and 20 i.e. RAE = 52.41% and 48.23% respectively.

Similarly, to determine the best kernel for SVM algorithm, four different kernels namely Poly kernel, Normalized Poly Kernel, Pearson VII function-based Universal kernel (PUK) and radial basis function (RBF) kernel were tested. The accuracy achieved by using these kernels is presented in Fig. 4. It reflects that the performance of PUK is significantly better than that of other three kernels. Hence study adopts PUK in further experiments. The first part of the analysis involves the proficiency evaluation of models. For each model, the experimental design calculates correlation coefficient (R^2) under different train/test conditions. Altogether, 5 different training and testing scenarios based on different training and testing dataset ratios were tried and tested for each prediction algorithm as shown

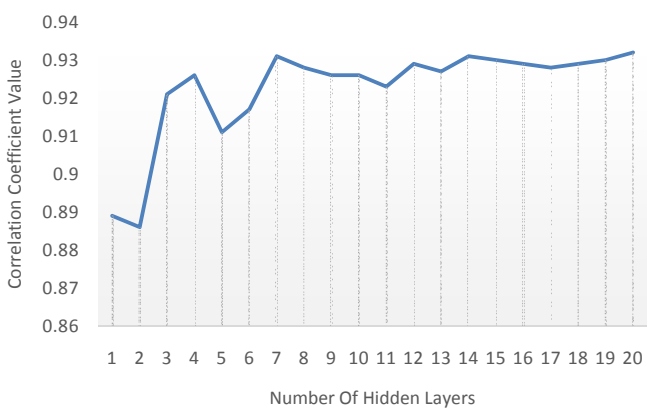

Fig. 3: Hidden layer selection using correlation coefficient value

in Fig. 5. Based on the best prediction performance, only RF, SVM, MLP and M5P out of seven classifiers were picked to determine the optimal training/testing scenario.

Prediction performance of RF, MLP, SVM and M5P, under different scenarios is presented in Fig. 5. It represents that, the performance of RF, in all scenarios has been the best, with a highest correlation coefficient equal to 0.95 under 80% –20% scenario, which no other classifier could achieve. While, the performance of M5P was the lowest among four with a correlation coefficient equal to 0.90. All four classifiers have performed the best under 80% –20% train/test scenario. The correlation coefficient obtained by each classifier under 5 different situations is fairly high. However, the consistency of RF to standout in all situations is noticeable. The performance of SVM to accurately predict the atmospheric concentration of SO_2 was nearly as good as that of RF under almost all situations. For an in depth analysis the prediction performance of all seven algorithms under 80% –20% train/test scenario was compiled in Table 2 with an aim to compare the performances of different single base algorithms against homogenous ensemble learning approaches. While results presented in Tables 3 and 4 were obtained by adopting group-I classifiers within bagging and voting respectively by using meta-conformity approach.

In Table 2, the performance of SVM and MLP have been fair with a correlation coefficient 0.94 and 0.93 respectively. But, RF (0.95) standout to be the best classification algorithms among all. Though the prediction performance of M5P is slightly shorter (R^2= 0.92) than that of MLP (0.93), however, interestingly, the error values obtained for M5P (RMSE= 0.0016, RAE=36.88%) are significantly better than MLP (RMSE=0.002 and RAE=44.65%). In comparison other algorithms such as AR, LG and RSS have

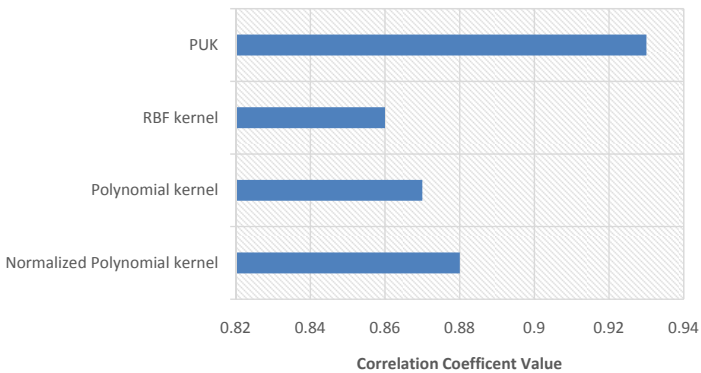

Fig. 4: SVM performance under different kernels

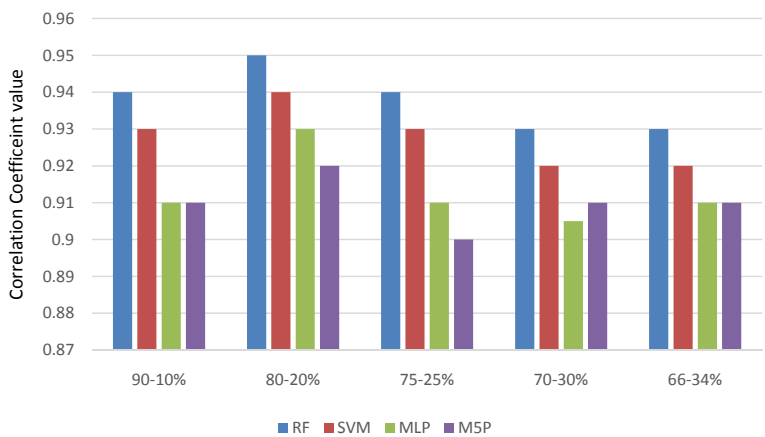

Fig. 5: Training and test scenarios

underperformed having low prediction accuracy and high error values.

During group-II experiments all seven group-I algorithms were employed within bagging as base classifiers to enquire if it affects the prediction accuracy of atmospheric SO_2. Results presented in Table 3 are evident of an obvious boost observed in correlation coefficient for AR (from R^2=0.85 to 0.88) and RSS (from 0.89 to 0.92) within bagging, whereas a slight improvement in prediction accuracy of SVM (R^2=0.94 to 0.95) was also reported. In fact the improvement in prediction accuracy of SVM within bagging has not just matched the group-I best classifier RF (R^2=0.95), but Bagged SVM performance in terms of error values (RMSE=0.0013, RAE=31.11%) is slightly better than Bagged RF (RMSE=0.0014, RAE=32.39%). On one hand Table 3 confirms the ability of homogenous ensemble learning technique – bagging to improve the prediction accuracy of different single base classifiers (AR, RSS, SVM, M5P), whereas on the other hand, it shows its limitation too. Because algorithms like RF and LR stayed unaffected within bagging. Interestingly, adopting MLP within bagging has not really affected its accuracy but in terms of error values a clear improvement of nearly 8% in RAE is recorded. A slight enhancement in Bagged MLP and Bagged SVM performances with a noticeable reduction in RMSE and RAE values in comparison with MLP and SVM as independent learners, is a confirmation of results reported by Windeatt, (2008) and Cannon and Lord, (2000) that apart from enhancing the prediction accuracy, bagging can solve over-fitting and local minima problems as well. Based on Tables 2 and 3, it is inferred that the performance of ensemble learning techniques (RF and bagging) is superior to that of traditional single base classifiers SVM, MLP, LR and M5P.

Furthermore, to test the prediction performance of heterogeneous technique – Voting, group-III experiments were conducted by the same meta-conformity rule. It is important to mention here that

Table 2: Prediction performance of different algorithms

Classifiers	R^2	MAE	RMSE	RAE (%)
LR	0.87	0.0015	0.002	46.61
MLP	0.93	0.0016	0.002	44.65
M5P	0.92	0.0012	0.0016	36.88
RF	0.95	0.001	0.0013	30.90
AR	0.85	0.0017	0.0022	51.64
RSS	0.89	0.0017	0.0022	51.09
SVM	0.94	0.001	0.0013	31.66

Table 3: Prediction performance of group-I classifiers in Bagging

Bagged classifiers	R^2	MAE	RMSE	RAE (%)
Bagged-linear regression	0.87	0.0015	0.002	46.51
Bagged-MLP	0.93	0.0012	0.0015	36.92
Bagged-M5P	0.93	0.0012	0.0015	35.88
Bagged-random forest	0.95	0.0011	0.0014	32.39
Bagged-additive regression	0.88	0.0015	0.002	45.33
Bagged-random subspace	0.92	0.0015	0.002	46.04
Bagged-SVM	0.95	0.001	0.0013	31.11

Table 4: Results of different group-I classifiers combined in voting

Experiment No.	Voting	R^2	MAE	RMSE	RAE (%)
1	RF, LR, MLP	0.93	0.0012	0.0016	37.45
2	RF,LR, M5P	0.93	0.0012	0.0015	36.05
3	RF, M5P, AR, RSS	0.93	0.0012	0.0016	37.31
4	LR, M5P, AR, RSS	0.91	0.0014	0.0018	41.21
5	MLP,LR, M5P,AR,RSS	0.92	0.0013	0.0017	39.89
6	SVM, MLP, LR, AR	0.93	0.0012	0.0016	37.26
7	SVM, RF, M5P	0.95	0.001	0.0013	30.88
8	SVM, RF, MLP	0.95	0.0011	0.0014	32.73

there is no specific rule that determines the minimum or maximum number of base classifier for an experiment in Voting. In Table 4, a total of 8 different combinations containing 3, 4 and even 5 group-I algorithms as base classifiers were tried. It is worth noting that voting uses the same group-I algorithms as base classifiers but ensemble them in different sets. First three experiments adopt RF due to its best performance (strong classifier), along with other weak classifiers such as MLP, M5P, LG, and AR in form of different composite models. In experiment 4 and 5 only weak classifiers were considered, whereas for experiment 6 and 7 SVM being best performer (strong classifier) was combined first with weak classifiers (MLP, LR, AR); and then with a weak (M5P) and a strong (RF) respectively. Last experiment only involves top group-I classifiers i.e. RF, SVM and MLP.

It was fascinating to see that in general all different sets of classifiers have resulted in an accuracy above 0.90. Specifically experiment 1, 2, 3 in which a strong classifier (RF) and 6 where SVM being strong contender was combined with weak classifiers, have performed exceptionally well by achieving a high correlation coefficient equal to 0.93 and low RAE value ranging from RAE = 36.05% to 37.45%. Although the prediction performance of experiment 4 and 5 is mediocre having R^2= 0.91 and R^2= 0.92 respectively, when only weak classifiers were considered in comparison with other results in Table 4, yet the accuracy values are good enough to compete popularly known single base classifiers such as MLP (R^2= 0.93), M5P (R^2= 0.92), and LR (R^2= 0.87). It establishes that ensemble mixing of single base learners within voting irrespective of strong or weak classifier sets, results in an enhanced prediction performance. Last two results listed in Table 4 are worth looking at because of two reasons: 1) their remarkably high prediction performance in terms of $R^2$and error values; and 2) the type of composite models tried. In both experiments RF and SVM were considered with third classifier M5P and MLP respectively. Table 4 is evident that the overall accuracy of both combinations have produced an all-time high correlation coefficient of R^2=0.95, but the error values obtained under experiment 7 when M5P (Tree classifier) joined RF and SVM is almost 2% lower than that when MLP was a third classifier. Which reflects the superiority of tree classifiers for their ability to predict with high predictive accuracy and low error values.

CONCLUSION

The work presented draws a comparison between three different types of classification schemes: single base learners (MLP, SVM, M5P, LR); homogenous ensemble (Random Forest, Bagging); and heterogeneous ensemble (Voting) techniques to predict the SO_2 concentrations in air by using 4 air pollutants (SO_2, NO_2, CO, and HCl) and 5 meteorological parameters (temperature, wind speed, wind direction, relative humidity, and atmospheric pressure). The results obtained suggest that 1) Random Forest performs significantly better as compared to single base classification algorithms and 2) Bagging has the ability to overall enhance the predictive accuracy of the single base learners. In fact SVM performed slightly better than RF when both were used as base classifiers within Bagging. The last set of experiments revealed that irrespective of type of classifiers (strong or weak), employing single base learners within voting, enhances their overall predictive accuracy. Whereas specifically, a set of classifiers containing SVM, RF along with M5P and MLP within heterogeneous ensemble classifier voting achieved the highest prediction accuracy (0.95) and lowest RMSE and RAE values (RMSE=0.001, RAE=30.88%). It indicates that, voting has the ability to efficiently compete with the best prediction classifiers RF (R^2=0.95, RMSE=0.0013, RAE=30.9%) and bagged SVM (R^2=0.95, RMSE=0.003, RAE=31.11%).

ACKNOWLEDGEMENT

The Author expresses its gratitude to the Management of Graduate School of Economics and Management, Ural Federal University, Ekaterinburg Russian Federation for providing necessary facilities to complete this study successfully.

CONFLICT OF INTEREST

The author declares that there is no conflict of interests regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancy have been completely observed by the authors.

ABBREVIATIONS

ANN	Artificial Neural Network
AQI	Air quality index
AR	Additive regression

CO	Carbon monoxide
CVPS	Cross validation parameter selection
DEFRA	Department for Environment Food and Rural affairs
HCl	Hydro chloric acid
LR	Linear regression
M5P	Regression Tree using M5 algorithm
MAE	Mean absolute error
MLP	Multilayer perceptron
MLR	Multi linear regression
NO_2	Nitrogen dioxide
O_3	Ozone
PCA	Principle Component Analysis
PM	Particulate matter
$PM_{2.5}$	Particles less than or equal to 2.5 micrometers in diameter
PM_{10}	Particles less than or equal to 10 micrometers in diameter
PUK	Pearson VII function based universal kernel
R^2	Correlation of determination
RAE	Relative absolute error
RBF	Radial basis function
RF	Random forest
RMSE	Root mean squared error
RSS	Random subspace
SO_2	Sulphur dioxide
SVM	Support vector machine
WD	Wind direction
WEKA	Waikato Environment for Knowledge Analysis
WHO	World Health Organization
WS	Wind speed

REFERENCES

Abdul-Wahab, S.A.; Al-Alawi, S.M.. (2002). Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks. Environ Modell Software, 17: 219-228 **(10 Pages).**

Alfaro, E.; García, N.; Gámez, M.; Elizondo, D., (2008). Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks. Decis. Support Syst., 45: 110-122 **(13 pages)**.

Baawain, M.S.; Al-Serihi, A.S., (2014). Systematic approach for the prediction of ground-level air pollution (around an industrial port) using an artificial neural network. Aerosol Air qual. Res., 14: 124-134 **(11 pages).**

Balasubramanian, V.; Ho, S.S.; Vovk, V.,(2014). Conformal prediction for relaible machine learning: theory, adoptions and applications. Newnes

Bedoui, S.; Gomri, S.; Samet, H.; Kachouri, A., (2016). A prediction distribution of atmospheric pollutants using support vector machines, discriminant analysis and mapping tools (Case study: Tunisia). Pollution, 2: 11-23 **(13 pages)**.

Brunekreef, B. and. (2002). Air pollution and health. The lancet, 360(9341): 1233-1242 **(10 pages)**.

Brunelli, U.; Piazza, V.; Pignato, L.; Sorbello, F.; Vitabile, S., (2007). Two-days ahead prediction of daily maximum concentrations of SO_2, O3, PM10, NO_2, CO in the urban area of Palermo, Italy. Atmos. Environ., 41, 2967-2995 **(29 pages)**.

Cannon, A.J.; Lord, E.R., (2000). Forecasting summertime surface-level ozone concentrations in the Lower Fraser Valley of British Columbia: An ensemble neural network approach. J. Air Waste Manage. Assoc., 50: 322-339 **(18 pages)**.

Capilla, C., (2014). Multilayer perceptron and regression modelling to forecast hourly nitrogen dioxide concentrations. WIT Trans. Ecol. Environ., 183: 39-48 **(10 pages)**.

Chaloulakou, A.; Saisana, M.; Spyrellis, N., (2003). Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens. Sci. Total Environ., 313: 1-13 **(13 pages)**.

Elangasinghe, M.A.; Singhal, N.; Dirks, K.N.; Salmond, J.A.; Samarasinghe, S., (2014). Complex time series analysis of PM10 and PM2. 5 for a coastal site using artificial neural network modelling and k-means clustering. Atmos. Environ., 94: 106-116 **(11 pages)**.

Fathima, A.; Mangai, J.A.; Gulyani, B.B., (2014). An ensemble method for predicting biochemical oxygen demand in river water using data mining techniques. Int. J. River Basin Manage., 12: 357-366 **(10 pages)**.

Gabralla, L.A.; Abraham, A., (2014). Prediction of oil prices using bagging and random subspace. Proceedings of the Fifth International Conference on Innovations in Bio-Inspired Computing and Applications, 343-354 **(12 pages).**

Gardner, M.W.; Dorling, S.R., (1999). Neural network modelling and prediction of hourly NOx and NO_2concentrations in urban air in London. Atmos. Environ., 33: 709-719 **(11 pages)**.

Grivas, G., and Chaloulakou, A. (2006). Artificial neural network models for prediction of PM10 hourly concentrations, in the Greater Area of Athens, Greece. Atmos. Environ., 40: 1216-1229 **(14 pages)**.

Jiang, N.; Riley, M.L., (2015). Exploring the utility of the random forest method for forecasting ozone pollution in Sydney. J. Environ. Protect. Sustainable develop., 1: 245-254 **(12 pages)**.

Juhos, I.; Makra, L.; Tóth, B., (2008). Forecasting of traffic origin NO and NO_2 concentrations by Support Vector Machines and neural networks using Principal Component Analysis. Simul. Mdel. Prat. Theory., 16: 1488-1502 **(15 pages)**.

Lu, W.Z.; Fan, H.Y.; Lo, S.M., (2003). Application of evolutionary neural network method in predicting pollutant levels in downtown area of Hong Kong. Neurocomputing, 51: 387-400 **(14 pages)**.

Lu, W.Z.; Wang, W.J.; Wang, X.K.; Xu, Z.B.; Leung, A.Y., (2003). Using improved neural network model to analyze RSP, NO x and NO 2 levels in urban air in Mong Kok, Hong Kong. Environ. Monit. Assess., 87: 235-254 **(20 pages)**.

Lu, W.Z.; Wang, D., (2014). Learning machines: Rationale and application in ground-level ozone prediction. Appl. Soft. Comput., 24: 135-141 **(7 pages)**.

Masih, A., (2018a). Thar Coalfield: Sustainable Development and an Open Sesame to the energy security of Pakistan. IOP Conference Series: Journal of Physics, 989 (1): 012004 **(8 pages)**.

Masih, A., (2018b). Modelling the atmospheric concentration of Carbon monoxide by using Ensemble Learning Techniques. Proceedings of the 5th International Young Scientists Conference on Information Technologies, Telecommunications and Control Systems, 2298: 12 **(8 pages)**.

Nawahda, A., (2016). An assessment of adding value of traffic information and other attributes as part of its classifiers in a data mining tool set for predicting surface ozone levels. Process Saf. Environ. Prot., 99: 149-158 **(10 pages)**.

Palani, S.; Liong, S.Y.; Tkalich, P., (2008). An ANN application for water quality forecasting. Mar. Pollut. Bull., 56: 1586-1597 **(12 pages)**.

Rahimi, A., (2017). Short-term prediction of NO_2 and NO_x concentrations using multilayer perceptron neural network: a case study of Tabriz, Iran. Ecol Processes., 6(4) **(9 pages)**.

Riga, M.; Tzima, F.A.; Karatzas, K.; Mitkas, P.A., (2009). Development and evaluation of data mining models for air quality prediction in Athens, Greece. Inf. Technol. Environ. Eng., 331-344 **(14 pages)**.

Russo, A.; Soares, A.O., (2014). Hybrid model for urban air pollution forecasting: A stochastic spatio-temporal approach. Math. Geosci., 46: 75-93 **(19 pages)**.

Salnikov, V.G.; Karatayev, M.A., (2011). Impact of air pollution on human health: Focusing on Rudnyi Altay industrial area. Am. J. Environ. Sci., 7(3), 286-294 (**9 pages**).

Samoli, E.;Atkinson, R.W.; Analitis, A.; Fuller, G.W.;Green, D.C.; Mudway, I.; Anderson, H.R.; Kelly, F.J., (2016). Association of short term exposure to traffic-related air pollution with cardiovascular and respiratory hospital admissions in London, UK. Occup. Environ. Med., 73: 300-307 **(8 pages).**

Schlink, U.; Dorling, S.; Pelikan, E.; Nunnari, G.; Cawley, G.; Junninen, H.; Greig, A.; Foxall, R.; Eben, K.; Chatterton, T.; Vondracek, J.; Richter, M.; Dostal, M.; Bertucco, L.; Kolehmainen, L.; Doyle, M., (2003). A rigorous inter-comparison of ground-level ozone predictions. Atmos. Environ., 37: 3237-3253 **(17 pages).**

Seinfeld, J.H., (1998). Atmospheric chemistry and physics: from air pollution to climate change. Phys. Today, 51: 88 **(13 pages)**.

Shaban, K.B.: Kadri, A.; Rezk, E., (2016). Urban air pollution monitoring system with forecasting models. IEEE Sens. J., 16: 2598-2606 **(9 pages)**.

Singh, K.P.; Gupta, S.; Rai, P., (2013). Identifying pollution sources and predicting urban air quality using ensemble learning methods. Atmos. Environ., 80: 426-437 **(12 pages)**.

Singh, K.P.; Gupta, S.; Kumar, A.; Shukla, S.P., (2012). Linear and nonlinear modeling approaches for urban air quality prediction. Sci. Total Environ., 426: 244-255 **(12 pages)**.

Tüfekci, P., (2014). Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. Int. J. Electr. Power Energy Syst., 60: 126-140 **(15 pages)**.

Loon, M.J.W.; Vautard, R.; Schaap, M.; Bergström, R.; Bessagnet, B.; Brandt, J.; Builtjes, P.J.H.; Christensen, J.H.; Cuvelier, C.; Graff, A.; Jonson, J.E.; Krol, M.; Langner, J.; Roberts, P.; Rouil, L.; Stern, R.; Tarrasón, L.; Thunis, P.; Vignati, E.; White, L.; Wind, P., (2007). Evaluation of long-term ozone simulations from seven regional air quality models and their ensemble. Atmos. Environ., 41, 2083-2097 **(15 pages).**

Wang, D.; Lu, W.Z., (2006). Interval estimation of urban ozone level and selection of influential factors by employing automatic relevance determination model. Chemosphere, 62: 1600-1611 **(12 pages)**.

Wang, W.; Men, C.; Lu, W., (2008). Online prediction model based on support vector machine. Neurocomputing, 71: 550-558 **(19 pages)**.

WHO, (2014). WHO's ambient air pollution database Update 2014.

Windeatt, T., (2008). Ensemble MLP classifier design. Comput. Intell. Paradigms., 133-147 **(15 pages).**

Xie, Y.; Zhao, L.; Xue, J.; Hu, Q.; Xu, X.; Wang, H., (2016). A cooperative reduction model for regional air pollution control in China that considers adverse health effects and pollutant reduction costs. Sci. Total Environ., 573: 458-469 **(12 pages)**.

Yang, P.; Hwa Yang, Y.; B Zhou, B.; Y Zomaya, A., (2010). A review of ensemble methods in bioinformatics. Curr. Bioinf., 5: 296-308 **(13 pages)**.

Yu, R.; Yang, Y.; Yang, L.; Han, G.; Move, O.A., (2016). Raq--a random forest approach for predicting air quality in urban sensing systems. Sensors, 16: 86 **(18 pages)**.

Zhan, Y.A., (2018). Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment. Environ. Pollut., 233: 464--473 **(10 pages)**.

AUTHOR (S) BIOSKETCHES

Masih, A., Ph.D. Candidate, Instructor, Department of System Analysis and Decision Making, Ural Federal University, Ekaterinburg, Russian Federation. Email: *adven.masikh@urfu.ru*

HOW TO CITE THIS ARTICLE

Masih, A., (2019). Application of ensemble learning techniques to model the atmospheric concentration of SO_2. Global J. Environ. Sci. Manage., 5(3): 309-318.

DOI: 10.22034/gjesm.2019.03.04

url: https://www.gjesm.net/article_35122.html