Modelling Ground-Level Ozone Concentration using Ensemble Learning Algorithms

Eman S. Al Abri[1], Eran A. Edirisinghe[1], and Amin Nawadha[2]
[1]Department of Computer Science, Loughborough University, United Kingdom
[2] Environment Research Centre, Sohar University, Sultanate of Oman.

Abstract— *Environmental risks caused by exposure to ground level ozone have significantly increased during recent years. One main producer of ozone is the photochemical reaction between volatile organic components and the anthropogenic nitrogen oxides created by vehicular traffic. Therefore the measurement and monitoring of atmospheric ozone concentration levels is important. In this paper we propose a study of the use of state-of-the-art machine learning approaches in modelling the concentration of ground level ozone. The prediction is based on concentrations of seven gases (NO_2, SO_2, and BTX (Benzene, Toluene, o-,m-,p-Xylene) and six meteorological parameters (ambient temperature, air pressure, wind speed, wind direction, global radiation, and relative humidity). The analysis of the results indicates that accurate models for the concentration of ground level ozone can be derived with the best performance accuracies indicated by the Ensemble Learning Algorithms. The investigation carried out compares the use of different machine learning classifiers and show that the Ensemble-classifier Bagging performs superior to standard single classifiers, such as Artificial Neural Networks and Support Vector Machines, popularly used in literature. In addition, we study the performance of the meta-classifier Bagging when different base classifiers are used in optimised configurations and compare the results thus obtained. The research conducted bridges an existing research gap in big-data analytics related to environment pollution prediction, where present research is largely limited to using standard learning algorithms such as Neural Networks and Support Vector Machines often available within popular commercial software packages.*

Keywords: Ozone, Atmospheric pollution, machine learning, Environment Science, Ensemble classifiers

1 Introduction

Ozone is a trans-boundary air pollutant that can be formed by photochemical reactions between anthropogenic nitrogen oxides(NO_x) and Volatile Organic Compounds (COVs) in the presence of sunlight[1]. When O_3 is formed, it remains suspended in the lower atmosphere (ground level ozone) for hours to days depending on the meteorological conditions and can endanger local and regional receptors.

In recent years, the environmental risks caused by exposure to ground level ozone (O_3) from both stationary and mobile sources have increased annually[2]. Several studies that analyse the effects of meteorological conditions on the formation and transport of O_3 have been listed in the work of [3]. Further, statistically significant relationships have been identified between elevated concentrations of O_3 and environmental risks in [4],[5].

A number of studies in the field of environmental science and engineering have focused their interest on constructing models to predict the concentrations of gases that result in air pollution. The majority of environmental researchers tend to use Artificial Neural Networks (ANN) and Support Vector machines (SVM) to predict ozone concentration [6]-[11]. Although there are more developed data mining / machine learning techniques, such as Ensemble learning approaches [12], only two attempts have investigated their use in predicting the ozone concentration; they are the work of [13] and [14]. This research showed that improvements in predictions can be obtained using bagging[15] as against using the popular single classifiers such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM). However these investigations were limited in the fact that Bagging was used only with the default single classifier RepTree[12] in WEKA (Waikato Environment for Knowledge Analysis) toolkit [16] as the base classifier. In the field of air pollution monitoring, no attempt has been made to test other ensemble classifier, select the best base classifier or to optimise the performance of the base classifier based on various possible parameter selections, all of which can lead to significant improvements in prediction accuracies. On the other hand, several attempts have been made in areas beyond air quality prediction in the use of ensemble classifiers, such as in bioinformatics, medicine and marketing, to build predictive models [17]-[21]. This work has shown that ensemble classifiers outperform the corresponding single classifiers and that the ultimate answer to the question, which classifier works best, depends on the dataset. It is clear that different datasets, in particular from different application domains, are statistically different and this has a high impact on the variability of results obtainable from different classifiers.

From the review of literature conducted and summarised above, a lack of research into effectively utilising Ensemble learning to predict ozone concentration was identified. Therefore the research proposed in this paper aims to find accurate models that can be used to predict ground level ozone concentrations, given a multitude of environmental parameters. An investigation was carried out comparing the performance of several machine learning techniques. Multiple predictive models were built using popular single classifiers namely Multilayer Perceptron (MLP) and Support Vector Machines and two ensemble learning algorithms, namely Bagging and Random Forests[22], using the WEKA toolkit. In addition, comparative analyses were performed to determine the algorithm that produced the best performance and to optimize the performance of each selected approach. The dataset considered in this work was obtained from Sohar University, Oman, which used a DOAS instrument [23] to gather the environmental data. The dataset includes concentrations of eight gases (O_3, NO_2, SO_2, and BTX (Benzene, Toluene, o-,m-,p-Xylene)) and six meteorological parameters (ambient temperature, air pressure, wind speed and direction, solar radiation, and relative humidity).

As implantations of the machine learning algorithms used for pre-processing/data-cleaning, feature selection, optimizing classifier parameters, modelling and performance analysis, WEKA has been used throughout this paper. Initially, training phases based on different classification algorithms for predicting O_3 concentration were performed. Subsequently, the prediction performance of different algorithms, were examined using ten-fold cross validation as implemented in WEKA. Various evaluation metrics have been utilised to analyse the results. It should be noted the key focus of the research conducted is not time-series analysis of O_3 concentration (i.e. predicting how O_3 concentration changes with time) but how to predict O_3 concentration based on the concentrations of the primary pollutant gases and the environmental parameters that can have an impact. In particular when O_3 creation is assumed to be due to the production of primary pollutant Nitrogen Dioxide, generated by vehicular traffic in this area, the time dependent analysis is not essentially useful.

For clarity of presentation this paper is divided into several sections. Apart from this section that provided the reader with an insight to the research context and identified research gaps, section-2 provides the background to data collection and presentation. Section-3 details the experimental procedure followed and section-4 provides the experimental results and a detailed analysis of the results. Finally section-5 concludes with an insight into future research.

2 Data collection and representation

This section provides details of the data collection approach used and how this data was represented for subsequent analysis.

2.1 The sampling site

Measurements were recorded across the Sohar Highway (SHW), Oman, in front of the main entrance to the Sohar University (SU) with a Differential Optical Absorption System (DOAS) instrument that was professionally installed (see Fig.1. for an aerial view of the system). The light beam travels a round-trip of 477 meters from A, which is located on the roof of the main administrative office building of SU, to B, where a reflector (or receiver) is installed on the top of another building situated across the road, as illustrated in Fig.1. The SHW has two lanes in each direction and an additional two single carriageway roads, in parallel, on both sides, bringing the total number of lanes to eight. Additionally, there is the SU car park, marked as C, where vehicular traffic may be present and thus would result in higher levels of O_3 concentrations. In order to capture the rapid variations of the concentrations of gases present in the space of the monitoring path, evaluations of light captured by the DOAS instrument is performed every 30 seconds for the measurement of the concentrations of O_3, NO_2, and SO_2 gases and every one minute for measurement of the concentrations of BTX. Additionally, the meteorological parameters, including wind speed and direction, relative humidity, pressure, temperature, precipitation, global solar radiation etc., are separately measured by sensors located on the roof of the SU building at A. The height of the instruments from ground level was approximately 12 metres.

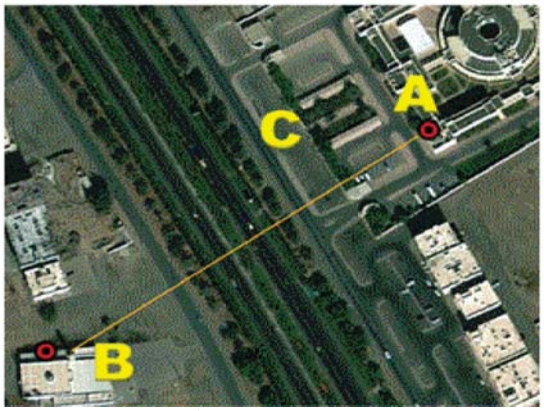


Fig. 1. Sampling path of the DOAS instrument; A = light emitter location, B = reflector location and C = car park.

2.2 Ozone dataset

The dataset used for the experiment was captured by the Sohar University DOAS system during 2013/2014. However, due to a technical fault in the system, the dataset collected during the specified period is not continuous. Nonetheless, a sufficiently large dataset was gathered to make the experiments statistically relevant. This dataset was analysed to investigate the modelling algorithm that gives the best prediction accuracy.

In the dataset used so far, there are a total of 6,744 instances spread across the years 2013-2014, as detailed in Table I.

TABLE I
DATASET DESCRIPTION

	2013			2014			Total number of records
	Start Date	End Date	No. of Rec.	Start Date	End Date	No. of Rec.	
Dataset	1st April 2013	23rd Aug. 2013	3480	1st March 2014	14th July 2014	3264	6744

2.3 Dataset representation

The target dataset is a sequence of measurements presented in a time series. The measurements are concentration values of eight gases measured in μgm^{-3} and readings of six environmental parameters. Table II lists the 14 attributes of each measured data value with their descriptive statistics.

TABLE II
ATTRIBUTES OF THE DATASET

2013-2014	Unit	Min	Max	Standard deviation	Mean
Sulphur Dioxide (SO_2)	μgm^{-3}	1.61	15.11	2.33	4.96
Nitrogen Dioxide (NO_2)	μgm^{-3}	0.02	83.99	16.65	18.24
Ozone (O_3)	μgm^{-3}	0.85	139.50	24.25	43.25
Benzene (C_6H_6)	μgm^{-3}	0.05	19.56	4.17	6.13
Toluene (C_7H_8)	μgm^{-3}	0.73	47.14	7.77	15.16
p-Xylene (C_8H_{10}(p))	μgm^{-3}	0.10	8.75	1.18	3.30
m-Xylene (C_8H_{10}(m))	μgm^{-3}	0.69	5.44	0.52	2.44
o-Xylene (C_8H_{10}(o))	μgm^{-3}	0.80	58.15	6.91	29.56
Temperature	ºC	16.19	45.06	3.53	31.10
Relative Humidity	%	8.47	93.57	19.33	64.38
Pressure	kPa	98.94	102.89	0.56	100.19
Global Radiation	W/m^2	-2.75	1120.24	247.95	201.13
Wind speed	m/s	0.31	6.266	1.02	1.77
Wind Direction	degree	0.11	359.99	91.50	137.52

Having collected the above dataset section-3 presents the method adopted for its analysis and detailed investigation.

3 Proposed method

The proposed approach adopts standard data mining procedure that involves data pre-processing prior to data modelling using machine learning. WEKA (version 3.7.11) is a toolkit that supports open source software implementation and operation of a large number of options for both data pre-processing and modelling. In this section we introduce the reader to the specific data pre-processing and modelling algorithms that have be adopted within the research context of the proposed work, as implemented by WEKA. Note that for our data analysis and method evaluation comparison purposes both Explorer and Experimenter software environments have been used, as appropriate.

3.1 Data pre-processing

Outlier Removal: In the data captured by the DOAS, missing values are recorded as -999.00. A careful analysis of the captured data also revealed that there are data measurement outliers, which could have resulted from instances of temporary sensor malfunctioning due to dust, high temperatures and overheating. Therefore a data cleaning operation within WEKA (listed under pre-processing) was utilised for the removal of outliers. The two filters interquartileRange (filters -> unsupervised -> attribute -> interquartileRange) and removeWithValues (filters -> unsupervised -> instances -> removeWithValues) were used respectively to clean the data in hand. Note that the first filter adds two extra columns to the data to indicate instances which contains the outliers and extreme values and the second filter removes such data by refereeing to the extra columns added by the first filter. After this cleaning process, only approximately 62% (4,173 out of 6,744 instances) of the original dataset were utilised for the next stage (modelling phase).

Data transformations: Since the wind direction is originally measured as an angle from the north in a clockwise direction, with values ranging from 0-360 degrees, the originally recorded witnd related data will have to be re-represented to avoid 0 and 360 degree directions being considered as different. The Wind Speed (WS) and Wind Direction (WD) have been combined and divided into two orthogonal compenents u = WS×cos(WD) and v = WS×sin(WD). (u,v) parameters will replace (WS, WD) in order to compensate for the above issue with regards to the original value of WD.

Attribute selection: Reduction of the attributes by eliminating the msot insignificant attributes can lead to both improved accuracy and speed of data modelling. The use of three popular feature selection filters have been investigated in the proposed research, namely, CFS Subset Evaluator with Best First and Greedy Stepwise Search methods, ReliefFAttributeEval with attribute ranking (removed last three attributes), and Principal Components. In the experiments conducted it was revealed that none of these filters uenhanced the accuracy of modelling although in the case of using the RelieFAttributeEval filter three of the most insignificant features were removed from used in modelling thus impacting positvely on speed.

3.2 Modelling the ozone concentration

As previously stated WEKA consists of implementations of a large number of classifiers that includes all state-of-the art and the popular traditional classifiers, such as, the Artificial Neural Networks and the Support Vector Machines. Our detailed experiments were designed to test all possible classifiers as single classifiers and as combined approaches (as appropriate). The purpose of this exhaustive investigation was to find the best classifiers / classifier combinations that outperformed traditional approaches for the prediction of air (Ozone) pollution thus generating new and useful knowledge for the community involved in environmental science and engineering research.

The initial exhaustive list was reduced to investigating 16 learning algorithms in detail from WEKA classifier categories, namely, Functions (4 different functions), Lazy (3), Meta (2), Rule (2) and Tree (5). The two meta-classifiers included the two popular Ensemble learning approaches Bagging and Random Forests. Furthermore, more detailed investigations were conducted with the Bagging ensemble classifier due to the initial indication of its superiority of performance. Within the detailed experiments thus conducted all the single learner classifiers initially experimented, were utilised as the base classifier of the ensemble classifier, Bagging.

Within the experimental context of this paper only six classification algorithms are analysed and discussed in detail. These include the two most popular single learning algorithms used in research that focus on air pollution analysis, Artificial Neural Networks [ANN] (implemented in WEKA as Multi-Layer Perceptron [MLP]) and Support Vector Machines [SVM] (implemented in WEKA as SMOreg) and the basic Ensemble Classifier, Random Forest [RF]. In conducting more detailed performance analysis of Bagging, the above three experiments are complemented with using them within Bagging as a base-classifier, namely Bagging with MLP, Bagging with SMOreg and Bagging with Random Forest. Although a large number of other classifiers and classifier combinations were evaluated, the detailed analysis of only these algorithms is presented in section-4. The accuracy of the algorithms are evaluated using two widely used evaluation metrics: Correlation Coefficient, Mean Absolute Error.

To present a fair performance comparison between the prediction models presented, optimal parameters for each classifier has been examined prior to conducting detailed modelling. The CVParameterSelection optimisation algorithm of WEKA has been used for this purpose.

The Explorer GUI environment of WEKA has been used to construct individual classifier models using their optimal parameters settings. Hence, the performance of different classifiers have been analysed and compared, using the same dataset (see section 2) using the Explorer. Since the Explorer does not provide the statistical significance of the improvements achievable by different classifiers, WEKA's Experimenter GUI environment was utilised to obtain additional information. A statistical test (Paired T-Tester corrected) was used to calculate the statistical significance between the different predictive models. The performance of the classifiers were examined using 10 fold cross validation and was compared using the Correlation Coefficient.

4 Experimental results and analyses

Experiments were conducted to analyse and compare the performance of six different classifiers: MLP (WEKA's ANN implementation), SMOreg (WEKA's SVM implementation), Random Forest (RF), Bagged-MLP, Bagged-SMOreg and Bagged-RF. Further detailed experiments were also conducted to determine the potential impact of feature reduction / selection and in the selection of classifier parameters in optimising classifiers, in the overall accuracy obtainable from each of the six evaluated classifiers. Further the original readings recorded for wind direction was a measure in the range 0-360 degrees. In order to compensate for the fact that 0 and 360 degree readings mean the same, we have combined wind direction (WD) with wind speed (WS) to replace them with two orthogonal components WS×cos(WD) and WS×sin(WD).

It is noted that all of the classifiers investigated (i.e. regardless of whether the classifier is of the single classifier type or the ensemble classifier type) consist of a number of input parameters that may have a vital impact on the accuracy of predictions obtainable. Although WEKA provides default parameter values for each classifier, our preliminary experiments suggested that these values do not result in optimised prediction. Therefore it was vital to select a set of parameters which provide optimal prediction accuracy. For this purpose we made use of WEKA's CVParameterSelection filter. Table III tabulates the prediction accuracy obtainable via each approach in terms of correlation coefficient. The results indicate that the optimal parameter selection has a positive impact only when use the single classifiers MLP (i.e. ANN) and SMOreg (i.e. SVM for regression). When using ensemble classifiers Random Forest and Bagging, the optimal parameter selection algorithm has no impact, indicated by the accuracy figures that remain unchanged. It is noted that even though the CVParameterSelection filter changes some parameters in its attempt to optimise the accuracy, no change is indicated in comparison to the accuracy obtainable using default settings.

For clarity of comparison Table IV tabulates overall prediction accuracies obtainable by each classifier presented in terms of the Co-relation Coefficient and Mean

Absolute Error with both using the default parameter settings of WEKA and with optimised parameter settings.

Fig.2 illustrates graphs representing the actual Ozone concentration vs the predicted Ozone concentrations. The graphs illustrate the better prediction capability of Bagged Random Forest classification approach as compared to the others. Data points lie closer to the line of approximation (less spread) than in the other graphs.

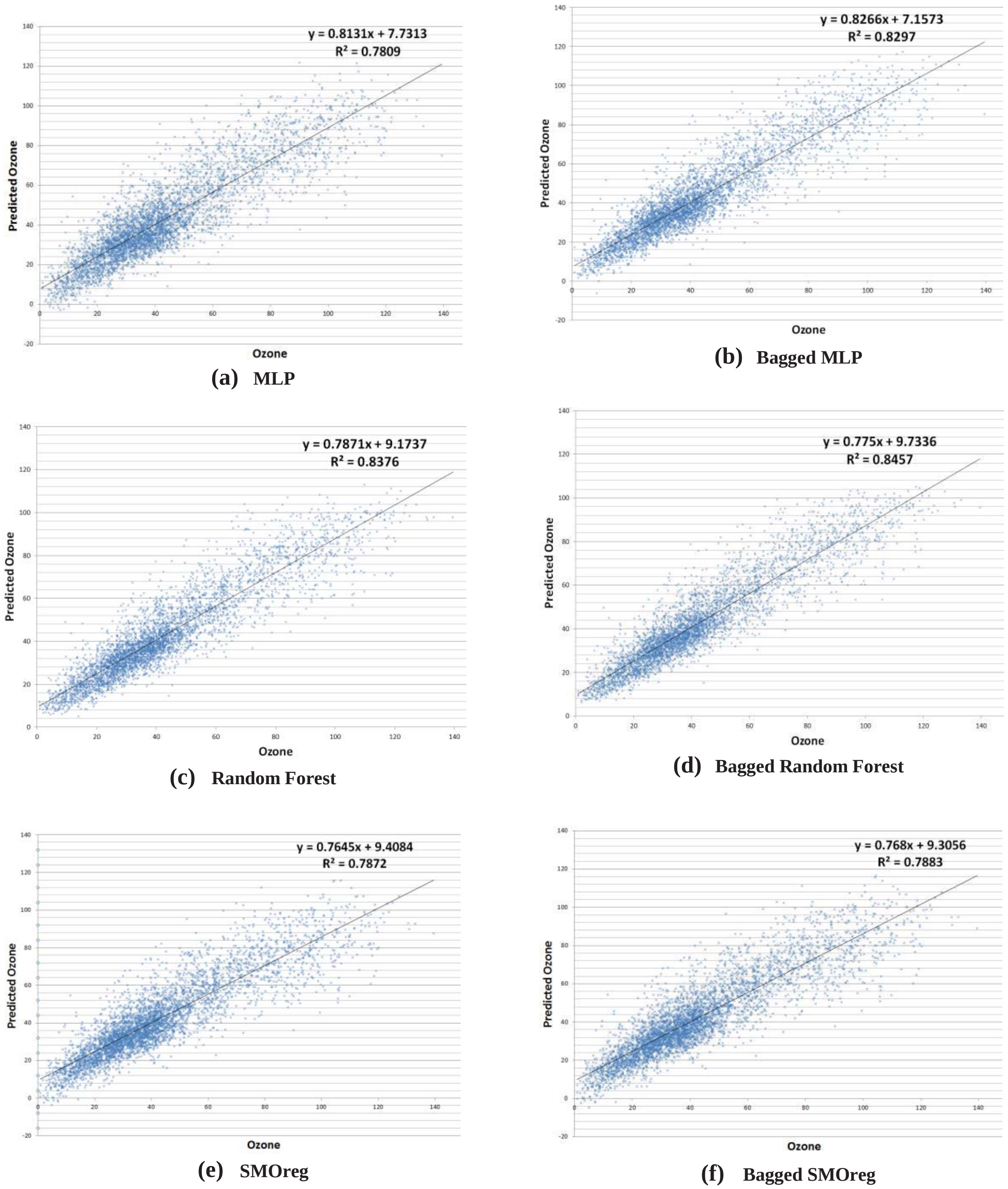
TABLE III
EXPERIMENTS TO OPTIMISE THE CLASSIFIERS

Classifier Name	Default settings	Correlation Coefficient	Optimal Parameters	Correlation Coefficient
MLP	Learning Rate (L)=0.3 Momentum /(M)=0.2 Hidden layer= a (attribute/class)/2	0.85	Learning Rate(L)=0.1 Momentum (M)=0.1 Hidden layer= 5	0.88
Bagged MLP	Bagging: bag size percent (P)=100 Number of iteration(I)=10 Seed (S)=1 num-slots =1 MLP: Learning Rate (L)=0.3 Momentum /(M)=0.2 Hidden layer= a (attribute/class)/2	0.90	Bagging: bag size percent (P)=100 Number of iteration(I)=10 Seed (S)=1 num-slots =1 MLP: Learning Rate(L)=0.1 Momentum (M)=0.1 Hidden layer= 5	0.90
Random Forest	NumTree (I)=10 NumFeature (K)=0	0.92	NumTree (I)=20 NumFeature (K)=0	0.92
Bagged RandomForest	Bagging: bag size percent (P)=100 Number of iteration(I)=10 Seed (S)=1 num-slots =1 Random Forest: NumTree (I)=10 NumFeature (K)=0	0.92	Bagging: bag size percent (P)=100 Number of iteration(I)=10 Seed (S)=1 num-slots =1 Random Forest: NumTree (I)=20 NumFeature (K)=0	0.92
SMOreg	C:1.0 Kernal: polyKernel	0.84	C:1.0 Kernel: NormalizedPolyKernel	0.89
Bagged SMOreg	Bagging: bag size percent (P)=100 Number of iteration(I)=10 Seed (S)=1 num-slots =1 SMOreg: C:1.0 Kernal: polyKernel	0.84	Bagging: bag size percent (P)=100 Number of iteration(I)=10 Seed (S)=1 num-slots =1 SMOreg: C:1.0 Kernal: NormalizedPolyKernel	0.89

TABLE IV
RESULTS OF THE PREDICTION MODELS

	Default Parameters		**Optimal Parameters**	
Classifier	**Correlation Coefficient**	**Mean Absolute Error**	**Correlation Coefficient**	**Mean Absolute Error**
MLP	0.85	9.81	0.88	8.51
SMOreg	0.84	9.54	0.89	8.05
RandomForest	0.91	7.52	0.92	7.16
Bagged MLP	0.90	7.64	0.91	7.27
Bagged RandomForest	0.92	7.08	0.92	7.05
Bagged SMOreg	0.84	9.54	0.89	8.04

Table V tabulates the accuracy values obtained when using four different attribute filtering approaches implemented within WEKA, namely, CFS Subset Evaluator, with Best First and Greedy Stepwise search, Relief Attribute Evaluator and Principle Component Analysis. The results indicate that no improvement of accuracy is achieved in comparison with using all attributes. We also investigated the impact of removing wind direction from being considered, taking only the wind speed into account (from the original data recorded). It was seen that the wind direction has negligible impact on the Ozone concentration prediction accuracy. This is justifiable as the measurements for Ozone was done across the road, i.e. at its source, as it was vehicular traffic that was suspected to create the Ozone from the Nitrogen Dioxide emissions.



(a) **MLP**

(b) **Bagged MLP**

(c) **Random Forest**

(d) **Bagged Random Forest**

(e) **SMOreg**

(f) **Bagged SMOreg**

Fig. 2. Scatter Plots of the actual and predicted Ozone for 6 Models

TABLE V
RESULTS OF APPLYING FEATURE SELECTION

	MLP	**SMOreg**	**Random Forest**	**Bagged MLP**	**Bagged SMOreg**	**Bagged RandomForest**
CFS-Best First	0.82 (-3)	0.82 (-2)	0.89 (-3)	087 (-3)	0.82 (-2)	0.90 (-2)
CFS-Greedy Stepwise	0.81 (-4)	0.82 (-2)	0.88 (-4)	0.86 (-4)	0.82 (-2)	0.90 (-2)
Relief Att. Eval.	0.83 (-2)	0.83 (-1)	0.91 (-1)	0.89 (-1)	0.83 (-1)	0.92 (0)
PCA	0.84 (-1)	0.83 (-1)	0.87 (-5)	0.89 (-1)	0.83 (-1)	0.89 (-3)
Using All Attributes	**0.85**	**0.84**	**0.92**	**0.90**	**0.84**	**0.92**

Ensemble learning is an approach that uses different classification techniques to build up a single model. Proposed by Breiman, 1996, Bootstrap Aggregation (Bagging) is a common type of an ensemble learning approach. Bagging resamples the original data, by using the bootstrap method, randomly, but with replacement (some can be selected repeatedly while other may not). The data produced are different from each other, however, the size of these samples are equal. Subsequently, a tree is built up from each sample. Later a classification model is developed from each sample using a single learning algorithm. Subsequently the outputs of different models are integrated into a single predication model. It uses either the weighted vote or average vote, depending on the type of task (i.e., a classification task or regression task, respectively). Due to the above process adopted by Bagging it resolves the data over-fitting problem associated with most classifiers, in this case with MLP and SVM in particular.

This is the reason for the significantly better prediction accuracies obtainable from using the Ensemble Classifier Bagging as against the accuracies obtainable from the traditional single classifiers commonly used in predicting Ozone, ANN and SVM.

5 Conclusion and future works

In this paper we have compared the performance of six machine learning algorithms in predicting the ground level atmospheric ozone concentrations. The prediction was based on concentrations of seven gases (NO2, SO2, and BTX (Benzene, Toluene, o-,m-,p-Xylene) and six meteorological parameters (ambient temperature, air pressure, wind speed, wind direction, global radiation, and relative humidity). Results prove the ability of ensemble learning algorithms, Random Forests and Bagging to perform significantly better than the traditional single classifier based learning algorithms, Artificial Neural Networks and Support Vector Machines.

We are currently extending the research presented within this paper to predict Ozone concentration variations over long periods of time, extending beyond a five year period, attempting to identify patterns and trends.

6 Acknowledgment

Authors would like to acknowledge the support of the research Council and the Ministry of Manpower of the Sultanate of Oman for providing the funding for this research and Loughborough University UK for providing research support. The research leading to these results has received research project grant funding from the Research Council of the Sultanate of Oman, research grant agreement no [ORG SU EBR 12 013].

7 References

[1] U.S. Environmental Protection Agency, "Guidelines for Developing an Air Quality (ozone and PM2.5) Forecasting Program," 2003.

[2] Region 7 Air Program, "Health Effects of Air Pollution," *EPA*. [Online]. Available: http://www.epa.gov/region07/air/quality/health.htm. [Accessed: 28-Feb-2015].

[3] D. M. Agudelo–Castaneda, E. C. Teixeira, and F. N. Pereira, "Time–series analysis of surface ozone and nitrogen oxides concentrations in an urban area at Brazil," *Atmos. Pollut. Res.*, vol. 5, pp. 411–420, 2014.

[4] M. Jerrett, R. T. Burnett, A. P. I. C, K. Ito, G. Thurston, D. Krewski, Y. Shi, E. Calle, and M. Thun, "Ozone exposure and mortality.," *N. Engl. J. Med.*, vol. 360, p. 2788; author reply 2788–2789, 2009.

[5] WHO Regional Office for Europe, "Health risks of ozone from long-range transboundary air pollution," p. 111, 2008

[6] M. S. Baawain and A. S. Al-Serihi, "Systematic approach for the prediction of ground-level air pollution (around an industrial port) using an artificial neural network," *Aerosol Air Qual. Res.*, vol. 14, pp. 124–134, 2014.

[7] N. Loya, I. Olmos Pineda, D. Pinto, H. Gómez-Adorno, and Y. Alemán, "Forecast of air quality based on ozone by decision trees and neural networks," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7629 LNAI, pp. 97–106, 2013.

[8] B. Mileva-Boshkoska and M. Stankovski, "Prediction of missing data for ozone concentrations using support vector machines and radial basis neural networks," 2007.

[9] A. S. Luna, M. L. L. Paredes, G. C. G. de Oliveira, and S. M. Corrêa, "Prediction of ozone concentration in tropospheric levels using artificial neural networks and support vector machine at Rio de Janeiro, Brazil," *Atmos. Environ.*, vol. 98, pp. 98–104, Dec. 2014.

[10] S. . Abdul-Wahab and S. . Al-Alawi, "Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks," *Environ. Model. Softw.*, vol. 17, no. 3, pp. 219–228, Jan. 2002.

[11] A. Coman, A. Ionescu, and Y. Candau, "Hourly ozone prediction for a 24-h horizon using neural networks," *Environ. Model. Softw.*, vol. 23, no. 12, pp. 1407–1421, Dec. 2008.

[12] I. H. Witten, E. Frank, and M. Hall, *Data Mining Practical Machine Learning Tools and Techniques*, 3rd ed. Elsevier, 2011.

[13] K. P. Singh, S. Gupta, and P. Rai, "Identifying pollution sources and predicting urban air quality using ensemble learning methods," *Atmos. Environ.*, vol. 80, pp. 426–437, Dec. 2013.

[14] A. J. Cannon and E. R. Lord, "Forecasting Summertime Surface-Level Ozone Concentrations in the Lower Fraser Valley of British Columbia: An Ensemble Neural Network Approach," *J. Air Waste Manage. Assoc.*, vol. 50, no. 3, pp. 322–339, Mar. 2000.

[15] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[16] WEKA; the University of Waikato, "Weka 3 - Data Mining with Open Source Machine Learning Software in Java." [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/index.html. [Accessed: 27-Feb-2015].

[17] P. Yang, Y. Hwa Yang, B. B. Zhou, and A. Y. Zomaya, "A Review of Ensemble Methods in Bioinformatics," vol. 5, pp. pp.296–308, 2010.

[18] E. Alfaro, N. García, M. Gámez, and D. Elizondo, "Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks," *Decis. Support Syst.*, vol. 45, no. 1, pp. 110–122, Apr. 2008.

[19] L. A. Gabralla and A. Abraham, "Prediction of Oil Prices Using Bagging and Random Subspace," Advances in Intelligent Systems and Computing, Proceedings of the Fifth International Conference on Innovations in Bio-Inspired Computing and Applications IBICA 2014 P. Kömer, et al. (eds)., Volume 303, pp. 343–354, 2014.

[20] A. Fathima, J. A. Mangai, and B. B. Gulyani, "An ensemble method for predicting biochemical oxygen demand in river water using data mining techniques," *Int. J. River Basin Manag.*, vol. 12, no. 4, pp. 357–366, Oct. 2014.

[21] P. Tüfekci, "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.*, vol. 60, pp. 126–140, Sep. 2014.

[22] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[23] OPSIS, "UV DOAS Technique," 2014. [Online]. Available: http://opsis.se/Techniques/UVDOASTechnique/tabid/632/Default.aspx. [Accessed: 09-Feb-2015].