Classifier Ensemble Methods for Diagnosing COPD from Volatile Organic Compounds in Exhaled Air

Ludmila I Kuncheva[1*], Juan J. Rodriguez[2], Yasir I Syed[3,4], Christopher O Phillips[5] and Keir E Lewis[3,4]

[1]School of Computer Science, Bangor University, UK
[2]Department of Civil Engineering, University of Burgos, Spain
[3]Institute of Life Sciences, College of Medicine, Swansea University, UK
[4]Respiratory Unit, Prince Philip Hospital, Llanelli, UK
[5]Welsh Centre for Printing and Coating, College of Engineering, Swansea University, UK

Abstract: The diagnosis of Chronic Obstructive Pulmonary Disease (COPD) is based on symptoms, clinical examination, exposure to risk factors (smoking and certain occupational dusts) and confirming lung airflow obstruction (on spirometry). However, most people with COPD remain undiagnosed and controversies regarding spirometry persist. Developing accurate and reliable automated tests for the early diagnosis of COPD would aid successful management. We evaluated the diagnostic potential of a non-invasive test of chemical analysis (volatile organic compounds - VOCs) from exhaled breath. We applied 26 individual classifier methods and 30 state-of-the-art classifier ensemble methods to a large VOC data set from 109 patients with COPD and 63 healthy controls of similar age; we evaluated the classification error, the F measure and the area under the ROC curve (AUC). The results show that classifying the VOCs leads to substantial gain over chance but of varying accuracy. We found that Rotation Forest ensemble (AUC 0.825) had the highest accuracy for COPD classification from exhaled VOCs.

Keywords: Chronic Obstructive Pulmonary Disease (COPD); Automatic diagnosis from exhaled breath; Pattern recognition; Classification; Classifier ensembles; Rotation Forest

Introduction

Chronic Obstructive Pulmonary Disease (COPD) is characterised by airflow limitation, which is not fully reversible. The causes are largely attributed to inhaling tobacco smoke, occupational exposure to dust and chemicals, and indoor and outdoor pollution (Rabe et al. 2007). COPD is a major public health problem and is the only one of the top five causes of death in the first world that is still rising. It is predicted to become the third leading cause of death by 2030, according to a study published by the World Bank/World Health Organization (WHO, 2008), and accounts for much chronic illness and morbidity. Yet, the Global initiative Obstructive Lung Disease (GOLD) report by Rabe et al. (2007) admits that COPD remains relatively unknown or ignored by the public as well as public health and government officials.

[*] Corresponding author. School of Computer Science, Bangor University, UKDean Street, Bangor Gwynedd. UK. LL57 1UT. Email: l.i.kuncheva@bangor.ac.uk

The current diagnosis of COPD is based on reported symptoms, patient's medical history (particularly exposure to risk factors), clinical examination, and then confirming lung air-flow obstruction (spirometry) where Forced Expiratory Volume in 1 second (FEV1) divided by Forced Vital Capacity is less than 0.80 and FEV1 predicted is less than 0.7 . (Rabe et al 2007)

Developing accurate and reliable automatic tests for early diagnosis of COPD is crucial for disease management as removing risk factors and early inhaled treatments has been shown to prevent progression, chronic ill health and premature death. (Rabe et al 2007). The current main test, spirometry, is effort dependent and often performed poorly. It can lead to over diagnosis in the young and underdiagnosis in the elderly. Moreover, it has not been validated in ethnic minorities. (Rabe 2007). The quest for a reliable biomarker in COPD is ongoing.

The smell of breath has long been linked with illness or physical conditions. Can volatile organic compounds (VOCs), measured from the exhaled breath, be used to identify COPD? Following on from Pauling's (1971) initial description of around 200 volatile organic compounds (VOCs) in exhaled breath, the trapping, detection and analysis of breath VOCs have been further developed. VOC analysis has been used to distinguish smokers from non-smokers (van Berkel et al, 2008), recognition of asthma (Ibrahim et al., 2011; Fens et al., 2009), lung cancer (Ulanowska et al. 2011; Machado et al., 2005; Philips et al. 2003; Bajtarevic et al., 2010, Barkar 2006) and tuberculosis (Phillips et al., 2007). Diagnosis of COPD from VOCs has also been attempted (Basanta et al., 2010, Fend et al., 2009, Van Berkel et al., 2010, Philips et al., 2012).

Here we study the diagnostic potential of the chemical signature of the exhaled breath for distinguishing between patients with COPD and healthy controls. We apply a large collection of state-of-the-art classification methods developed within the areas of pattern recognition, machine learning and data mining, with a special focus on classifier ensembles. We applied these methods to the largest data set so far derived from our previous work (Philips et al 2012). We demonstrate that the ensemble methods are superior to the individual classifier methods, resulting in better classification accuracy, F measure and the area under the ROC curve (AUC).

Material and Methods

Related work

Table 1 shows a summary of the classification methods and techniques used in the recent literature on diagnosis of lung disease based on breath samples. While the collection of sources is by no means comprehensive, it reveals that the possibilities offered by modern pattern recognition (Duda et al., 2001; Bishop, 2006), machine learning (Hastie et al., 2011; Schapire and Freund, 2012) and data mining (Witten and Frank, 2001) remain largely unexplored.

We applied the software package Weka (Hall, et al., 2009); a collection of machine learning algorithms for data mining tasks. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. Weka is open source software issued under the GNU General Public License, available at http://www.cs.waikato.ac.nz/ml/weka/. The classification methods we examined can be grouped into two large categories: individual classifiers and ensemble classifiers. The individual classifiers can be grouped into tree classifiers, rule classifier and other. Details about each of the classifiers can be found in the literature recommended in Weka.

Table 1. Summary of the classification methods and techniques for diagnosis of lung diseases based on breath samples.

Publication	Classification method	Comment
Basanta et al. 2010	Kruskal-Wallis ANOVA	20 subjects (20 COPD / 6 healthy smokers)
Bajtarevic et al., 2010	individual VOC analysis	96 subjects (65 lung cancer patients / 31 healthy volunteers)
Fens et al., 2009	4 PCA (98.7% explained variance) + Linear Discriminant Analysis (LDA)	90 subjects (30 COPD / 20 asthma / 20+20 controls)
Ibrahim et al., 2011	12 PCA (82% explained variance) + logistic regression	48 subjects (25 asthma / 23 controls)
Machado et al., 2005	22 PCA + canonical discriminant analysis, subsequently SVM	59 subjects (14 carcinoma / 45 controls)
Phillips et al., 2012	decision tree, 2 rule-based classifiers, SVM, 3 fuzzy classifiers and 3 ensemble methods: random forest, random subspace and bagging	182 subjects (119 COPD / 63 control)
Phillips et al., 2003	stepwise LDA	107 subjects (67 cancer / 41 controls)
Phillips et al., 2007	a fuzzy classifier and 'pattern recognition' classifier	101 subjects (23 tuberculosis / 19 suspected / 59 healthy controls)
Rogers et al., 2012	LDA	simulated breath
Sahin et al. ,2010	SVM	Not VOCs, only spirometry data; 499 samples (multiple samples from the same subject)
Ulanowska et al., 2011	factor analysis + LDA, CHAID trees	279 subjects (137 lung cancer patients / 142 controls)
Van Berkel et al., 2008	SVM	22 subjects (11 smokers and 11 non-smokers)
Van Berkel et al., 2010	SVM with feature selection	79 subjects (50 COPD / 29 controls)

Key: VOC -volatile organic compounds
LDA - linear discriminant analysis
SVM - support vector machine classifier
PCA - principal component analysis
CHAID a type of decision tree technique, based upon adjusted significance testing

Before we reason about the suitability of various classifier models to the data, we note the problems shared by many data sets in the medical domain.

- "Wide" data set are characterised by relatively small number of data points (called also instances or examples) compared to the number of features (attributes). In

VOC studies, the number of data points varies between 20 and several hundred (see Table 1) while the number of VOCs is typically in excess of 2000.

- Sparseness of the data comes from the fact that only a fraction of VOCs are likely to appear on the chemical signature of a breath sample. The remaining VOCs are in too small a quantity to trigger detection. Non-detections may also be a result of malfunctioning equipment.

These properties call for stringent experimental protocols to ensure that the element of serendipity is eliminated.

Participants

COPD patients were identified through hospital and primary care registers. All had previously confirmed obstructive lung disease on spirometry, were prescribed optimal inhaled medication, deemed stable by a respiratory clinician and none reported worsening symptoms within 6 weeks of testing. Healthy controls were drawn from spouses of patients, volunteers from local charity organisations and members of staff with no chest pain, breathlessness, cough nor wheeze on screening questions. Subjects gave written consent and the study was approved by our local ethics committee and registered (ISRCTN 82911859). 182 subjects participated . Following test for normality, baseline numeric variables were compared between the two groups using non-paired t-tests and the Mann Whitney U test. The categorical data were compared with chi square. Table 2 compares the two groups.

Table 2: Comparison of the two groups of subjects.

Variable (Mean±SD)	COPD (n=119)	Controls (n=63)	p-value
Age (years)	67.0±8.4	67.4±9.7	0.49
Male/Female	61% / 39%	47% / 53%	0.09
Smoking status (%) (never/ex-smokers/current)	0/66/34	62/29/9	0.000
Body Mass Index (kg/m^2)	25.7±4.6	27.0±4.4	0.11
Predicted FEV1	0.50±0.18	0.98±0.16	0.000
Oxygen saturations %	95.0±2.4	95.8±2.3	0.001

Key FEV1 Forced expiratory volume in 1 second

Procedure

Participants completed questionnaires for socio-demographic data, smoking status and any illnesses including current /recent symptoms then performed dry wedge spirometry (Vitalograph Alpha®,Buckinghamshire, UK) to achieve a Forced Expiratory Volume in 1 second (FEV1) as a marker of airflow obstruction. Smoking status was validated using exhaled carbon monoxide (CO) (Bedfont-Micro Smokerlyzer®). Resting peripheral oxygen saturations on air (Konica Minolta Pulsox-300, Konica Minolta Sensing Inc, Osaka, Japan) and Body Mass Index (BMI) were recorded.

Subjects were fasted for 4 hours, rested (sitting) for 30 minutes and the asked to exhale to slow vital capacity (maximum breath) into a trapping system. All samples were taken in the same hospital room with closed doors (to reduce background ambient air contamination of VOCs).

Three breath samples were taken from each subject, 2 minutes apart. A single background air sample was taken to monitor the ambient air at each sampling period. A commercially available sampler (Bio-VOC®, Markes International Limited, UK) was used to trap the last 129mL of breath from a full exhalation. This was then transferred to a thermal desorption tube containing carbon black sorbent which adsorbed the VOCs. Analysis was then performed using thermal desorption, gas-chromatography and mass spectrometry to extract, separate and identify the VOCs respectively. Details can be supplied.

Table 2 suggests that that any of the three variables alone: smoking status, predicted FEV1 or oxygen saturation, can distinguish reliably between COPD and controls with statistical significance $p<0.05$. Why continue? Our study is aimed at a further discovery of COPD diagnostic indicators. A classifier based solely on VOCs can contribute an 'independent opinion' to the collection of other, more traditional, diagnostic tests and cues. This combination may lead to a more reliable and accurate overall diagnostic tool, capable of detecting early stages of COPD.

Classifiers and Classifier Ensembles

Figure 1 sows a summary of the classification methods.

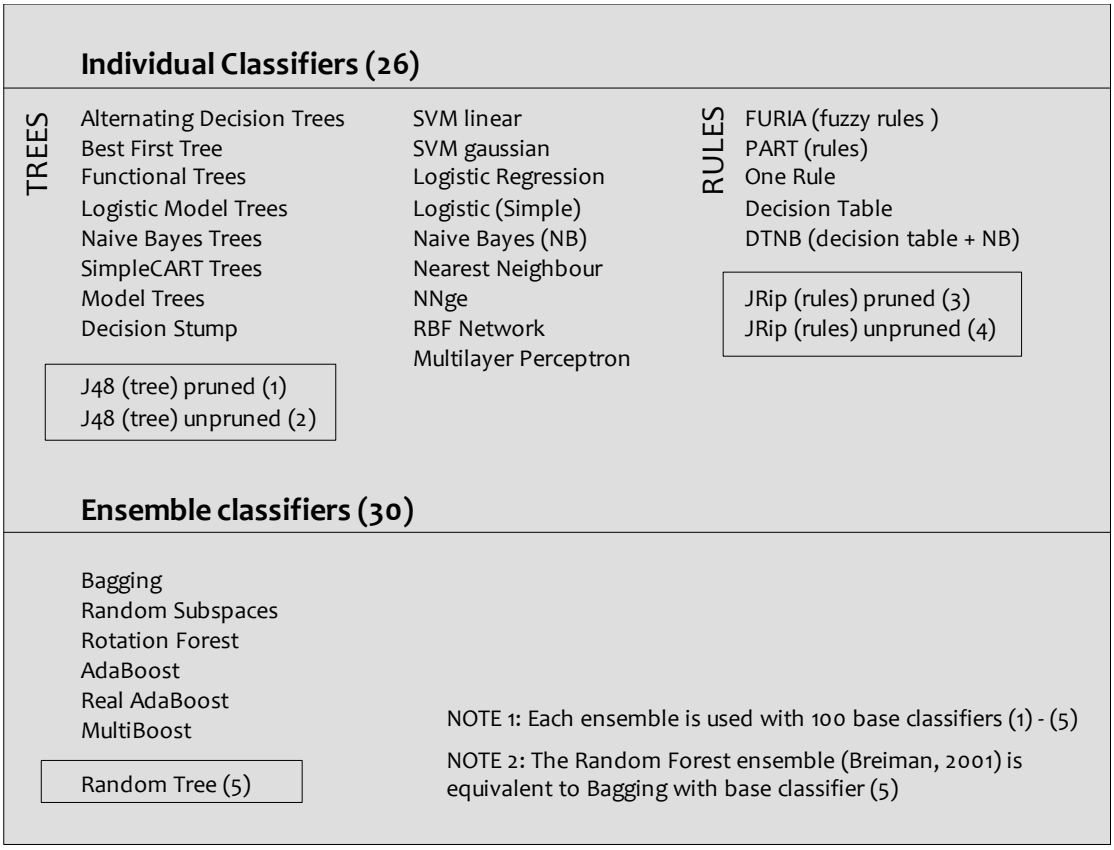

Figure 1. Individual and ensemble classifier methods used in this study

Below we give a brief comment on some of the individual classifiers included in the experiment

- SVM. Since its inception, the support vector machine classifier (SVM) has been gaining strength and is progressively eclipsing many earlier classifier models (Burges, 1998). This classifier is particularly suited to wide data type because it scales linearly along the feature dimension while tolerating the small sample size by ensuring large classification margins. SVM has a noticeable presence in the literature on VOCs classification (Table 1), not only as a classification method, but also as a powerful feature selection technique (Guyon et al., 2006). The SVM was applied with a linear and with a Gaussian kernel, with parameter values as pre-set in Weka.
- Decision tree classifier. This classifier is praised for its accuracy, robustness and interpretability (Breiman et al., 1984). These properties have prompted the development of many variants, some of which we used in our experiment. J48 is Weka's implementation of the decision tree classifier otherwise known as c4.5. This has been a popular choice of a base classifier in classifier ensemble studies.
- LDA. The linear discriminant analysis is a robust method for classification, and is commonly applied to biomedical data. We can speculate here that the wide use of LDA is a consequence of its availability in major statistical software packages and its acceptance within the statistics community. More flexible classifiers, backed by no less rigorous theory, may be overlooked in the process. One of the aims of our paper is to alert practitioners to the existence and availability of such methods.
- Rule-based classifiers. JRip is a version of a rule-based classifier which learns the geometry of the classes in the data (Cohen et al., 1995). Many rule-based classifiers exist, but unlike the decision tree family, the rule-base classifiers are not variants of one another, and may follow very different learning strategies.
- Neural Networks (NN). We included in the experiment a version of the Multi-Layer Perceptron (MLP) and the Radial Basis Function NN (RBF). One disadvantage of this type of classifier is the need of fine tuning. NNs enjoy a great success in the hands of experts. However, if they are used with the pre-defined parameters (which was our approach) they may not perform up to their full potential.

Classifier ensembles are now a well established and a widely acclaimed sub-field of pattern recognition, machine learning and data mining (Kuncheva, 2004; Rokach, 2010; Schapire & Freund, 2012; Zhou, 2012). Six most popular classifier ensemble methods were chosen for this experiment: Bagging (Breiman, 1996), Random Subspace (Ho, 1998), Rotation Forest (Rodríguez et al., 2006), AdaBoost (Freund & Schapire, 1997), Real AdaBoost (Schapire & Singre, 1999) and MultiBoost (Webb, 2002). A wealth of experimental work has been published trying to elect a winner among the ensemble methods. However, just like with the individual classifiers, the "No Free Lunch Theorem" holds, which states that no single method can be best on all possible data. This is why we chose all six methods for our study. The classifier ensemble methods are expected to be fairly robust with respect to the base classifier used. However, over the years, the decision tree classifier has been consistently elected as the most suitable base classifier. (Some of the ensemble methods are termed "forest" because of this choice.)

We note here that we have not neglected to include the Random Forest ensemble (Breiman, 2001), one of the most successful ensemble methods, especially suitable for

medical data. This method is equivalent to Bagging with Random Tree as the base classifier (Note 2 in Figure 1), which we have included in our set-up.

Experiment

Protocol

(i) The experiments were done using a 10 fold cross validation, repeated 10 times.
(ii) The cross validation was additionally indexed by the person identifier. The three records from the same person were placed in the same fold. This was done in order to ensure that the data for the same person did not appear simultaneously in the training and testing sets.
(iii) As a benchmark, we included in the comparisons the Majority classifier (also known as the Largest Prior classifier). This is the trivial classifier which always predicts the majority class label.
(iv) The performance of the classifiers and the ensembles was measured by the classification accuracy, the F measure, and also by the area under the ROC curve (AUC), taking COPD as the positive class. All calculations were done on the 100 testing folds of the 10-times 10-fold cross-validation. To illustrate the calculation, consider the following contingency table

		Labelled as	
		COPD	Healthy
True labels	COPD	*a*	*b*
	Healthy	*c*	*d*

One such table is produced for each of the 100 folds of the 10x10-fold cross-validation. The numbers in the table sum up to a tenth of the number of data points. In our data, the total number of samples was 3(breath samples)x182(subjects) = 546. In a typical CV run, 18 subjects would be left in the testing fold, hence $a + b + c + d = 18x3 = 54$.

The following quantity are calculated from the table

- classification accuracy = $(a + d) / (a + b + c + d)$
- sensitivity = $a / (a + b)$ = recall
- specificity = $d / (c + d)$
- precision = $a / (a + c)$
- F-measure = 2 ***** precision ***** recall / (precision + recall)

In principle, AUC measures the probability that the classifier will assign a higher score to a randomly chosen positive example than to a randomly chosen negative example. Recent studies have questioned the merit of AUC as a measure of classification performance (Hanczar et al., 2010), especially when the number of data points is not large. Hence we will refrain from offering statistical back up of our findings, and will exercise caution when interpreting the numerical results. In any case, the data was not sampled as i.i.d., hence the prior probability for COPD and healthy control cannot be estimated and used. This "misleads" the classifiers, which will try to accommodate the class prevalence, and therefore speaks in favour of AUC. Thus we decided to use all three measures.

Results

A total of 2075 different VOCs were recorded in samples. Of these, 146 VOCs had zero values for patient samples but were identified in background air samples only, suggesting they were ambient VOCs , leaving 1929 potentially useful VOCs.

Only 253 VOCs were observed in more than 5% of the subjects (in one or more of the three breath samples from the subject). In addition, we considered a measure of 'quality' (certainty) of the VOC detection against a library of standard mass spectra. VOCs whose quality was less than 50% were deemed unreliable. Reasons for a low quality score might be background 'noise' due to their low magnitude, or multiple compounds being insufficiently separated. After removing the low quality VOCs, 128 reliable and commonly detected VOCs remained as the input for the rest of the analyses

Tables 3 and 4 show the three measures for the 56 classifiers in the experiment. The values are averaged across the 100 testing folds of the CV. The cells with the largest value of the measure in the respective column are highlighted. To aid the interpretation of the results we included Figures 2 and 3. Each classification method is plotted as a dot (individual classifiers) or a cross (ensemble classifier). To visualise the relative position of the ensemble model that we recommend - the Rotation Forest - the 5 points corresponding to the different versions are plotted with a square marker.

Since all three performance criteria have their strengths and weaknesses, we derived the Pareto-optimal set of non-dominated classification methods. A classification method is called "non-dominated" if there is no other method in the set that has better or equal values on all criteria, such that at least one of the inequalities is strict. The Pareto optimal set is shown in Table 5, arranged in alphabetical order of non-dominated methods.

Conclusions:

(1) Classifier ensembles fare better than individual classifiers in diagnosing COPD from VOCs according to all three performance measures but there is no single classification method that is best on all criteria.

(2) The chance AUC value is 0.5. All classifiers clear this value by a large margin, suggesting high accuracy for the VOC classification in diagnosing COPD. The results with the other two performance measures are less impressive, more so for the F-measure.

(3.) The Rotation Forest ensemble (Rodríguez et al, 2006) achieves the highest value of the F-measure and the second highest for AUC, leading us to recommend this classification method for future analyses. The recommended base classifier is the rule-based classifier JRip (Cohen et al., 1995). However, the points for the Rotation Forest ensemble are closely clustered in Figures 2 and 3, which indicates that the method is reasonably robust to the choice of base classifier.

Table 3. Classification accuracy, F-measure and AUC for the individual classifiers.

	Accuracy		F-measure		AUC	
	mean	std	mean	std	mean	std
Majority	65.346	2.067	0.790	0.015	0.500	0.000
Alternating Decision Trees	69.836	7.524	0.767	0.067	0.777	0.068
Best First Tree	71.936	7.013	0.784	0.059	0.717	0.085
Functional Trees	69.547	4.849	0.760	0.049	0.664	0.109
Logistic Model Trees	69.774	6.670	0.769	0.069	0.733	0.070
Naive Bayes Trees	70.775	7.785	0.770	0.073	0.742	0.101
SimpleCART Trees	71.435	4.956	0.782	0.050	0.734	0.083
Model Trees	68.469	7.745	0.763	0.067	0.746	0.046
Decision Stump	65.346	2.067	0.790	0.015	0.685	0.072
J48 (tree) pruned	68.118	6.486	0.748	0.062	0.629	0.068
J48 (tree) unpruned	68.485	4.937	0.751	0.048	0.653	0.075
SVM linear	71.105	6.856	0.813	0.042	0.601	0.094
SVM gaussian	65.346	2.067	0.790	0.015	0.500	0.000
Logistic Regression	66.068	6.847	0.733	0.072	0.674	0.089
Logistic (Simple)	70.885	7.330	0.778	0.075	0.759	0.097
Naive Bayes	60.382	6.879	0.605	0.093	0.737	0.084
Nearest Neighbour	61.935	7.356	0.683	0.078	0.611	0.079
NNge	68.134	5.937	0.784	0.043	0.587	0.078
RBF Network	65.143	7.079	0.707	0.068	0.692	0.071
Multilayer Perceptron	61.871	12.996	0.635	0.227	0.690	0.104
FURIA (fuzzy rules)	68.827	6.093	0.779	0.047	0.686	0.086
PART (rules)	68.608	7.197	0.758	0.060	0.683	0.079
One Rule	63.472	4.143	0.738	0.033	0.565	0.057
Decision Table	69.470	6.270	0.806	0.039	0.587	0.083
DTNB (decision table + NB)	69.966	8.035	0.756	0.085	0.748	0.072
JRip (rules) pruned	68.801	9.483	0.766	0.078	0.649	0.097
JRip (rules) unpruned	71.082	6.867	0.794	0.048	0.649	0.088

Table 4. Classification accuracy, F-measure and AUC for the classifier ensembles.

		Accuracy		F-measure		AUC	
Ensemble method	Base classifier	mean	std	mean	std	mean	std
	Majority	65.346	2.067	0.790	0.015	0.500	0.000
Bagging	Random Tree	74.311	4.836	0.813	0.045	0.829	0.068
	J48 pruned	74.139	6.097	0.799	0.056	0.807	0.076
	J48 unpruned	73.963	6.278	0.799	0.056	0.811	0.067
	JRip pruned	71.211	6.588	0.795	0.051	0.784	0.079
	JRip unpruned	70.978	7.272	0.792	0.048	0.771	0.097
Random Subspace	Random Tree	73.194	5.335	0.807	0.038	0.813	0.063
	J48 pruned	73.907	3.487	0.806	0.038	0.820	0.048
	J48 unpruned	74.453	4.368	0.810	0.043	0.817	0.049
	JRip pruned	69.230	6.009	0.801	0.037	0.767	0.070
	JRip unpruned	68.492	4.306	0.798	0.028	0.771	0.064
Rotation Forest	Random Tree	74.318	5.014	0.814	0.039	0.824	0.062
	J48 pruned	73.738	6.837	0.801	0.061	0.817	0.073
	J48 unpruned	73.407	6.442	0.798	0.064	0.816	0.071
	JRip pruned	73.957	7.364	0.818	0.051	0.812	0.078
	JRip unpruned	74.325	5.732	0.821	0.037	0.825	0.076
AdaBoost	J48 pruned	74.635	6.226	0.811	0.057	0.761	0.059
	J48 unpruned	72.659	6.666	0.792	0.066	0.742	0.070
	JRip pruned	68.767	6.413	0.764	0.057	0.732	0.072
	JRip unpruned	72.794	6.395	0.797	0.050	0.730	0.106
	Random Tree	64.916	8.291	0.733	0.076	0.603	0.086
Real AdaBoost	J48 pruned	74.136	4.775	0.805	0.038	0.726	0.085
	J48 unpruned	74.049	7.085	0.801	0.075	0.737	0.084
	JRip pruned	68.472	4.505	0.764	0.047	0.715	0.086
	JRip unpruned	72.761	5.940	0.798	0.049	0.722	0.078
	Random Tree	73.355	4.398	0.803	0.038	0.716	0.073
MultiBoost	J48 pruned	75.012	5.223	0.811	0.048	0.743	0.063
	J48 unpruned	75.200	5.728	0.812	0.058	0.751	0.062
	JRip pruned	72.660	4.143	0.798	0.045	0.790	0.073
	JRip unpruned	71.717	6.219	0.786	0.062	0.734	0.068
	Random Tree	64.916	8.291	0.733	0.076	0.603	0.086

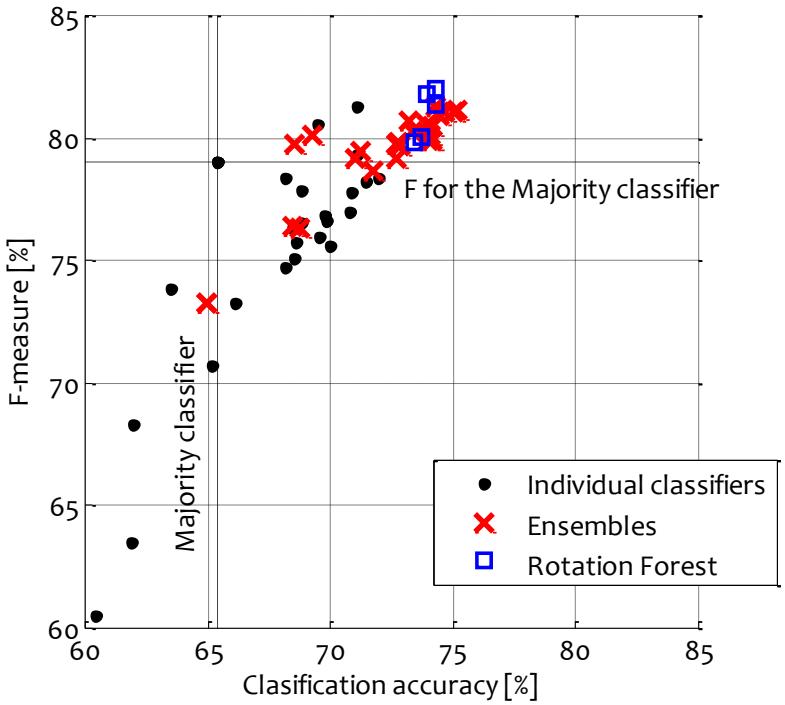

Figure 2. F-measure versus classification accuracy for the 26 individual classifiers and the 30 classifier ensemble methods.

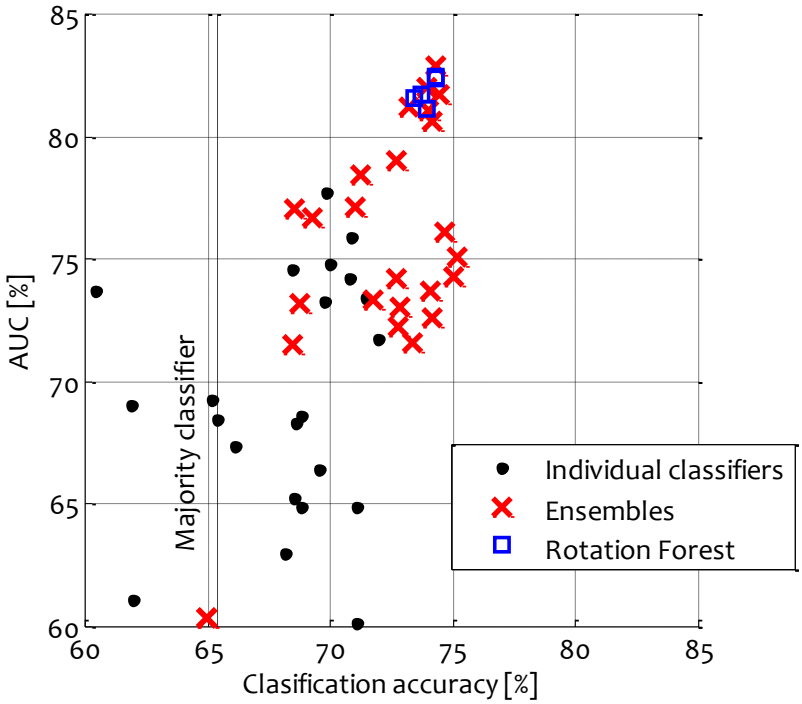

Figure 3. AUC versus classification accuracy for the 26 individual classifiers and the 30 classifier ensemble methods.

Table 5. Pareto frontier for the set of 56 classification methods

Method	Base classifier	Average of the three criteria [%]	Comment
AdaBoost	J48 pruned	77.267	
Bagging	Random tree	79.501	= Random Forest
Multiboost	J48 unpruned	79.049	
Random subspace	J48 unpruned	77.167	
Rotation Forest	JRip unpruned	79.615	

Discussion

Due to the technicality of trapping and VOC measurements in specialist laboratories, it is not yet clear whether VOC measurement and classification will contribute significantly to the routine diagnosis or monitoring of COPD. Future studies are needed to assess the "value for money" of VOC classification when combined with the traditional diagnostic tests.

However, this study helps advance new possibilities by pointing at the rich and unexplored yet armoury of methods offered by modern pattern recognition, machine learning and data mining. Many of these methods can be used as off-the-shelf classifiers, requiring minimal parameter tuning, if any. Although lacking interpretability (a common characteristic of the most powerful classification methods, for example, classifier ensembles), these methods can be used safely and reliably in the form of black boxes by non-specialist users. The user's trust will come from the clean and rigorous experimental protocol where the work of the methods will be assessed on unseen data.

With the development of modern technology, electronic noses and devices for VOC analysis may become an inexpensive tool available to the general practitioners and home monitoring. Exhaled breath analysis is non-invasive; it is a smaller and portable test, and involves no radiation exposure (unlike X-rays or computer tomography scanning). Unlike spirometry, VOCs analysis is not effort dependent.

Future research should focus on securing a clean and reliable data set by improving trapping and measuring techniques to reduce contamination and increase the number and the 'quality' ratings of the VOCs. This will lead to a larger number of useful VOCs, and hopefully a better classification performance. It is worth developing a stand-alone software tool allowing researchers and practitioners to experiment directly with the most successful ensemble methods discovered in our analyses.

Acknowledgements

This work was supported by the HMC2 initiative. Dr Syed's research salary was supported by Hywel Dda Health Board. The authors are indebted to the Probus Club of Llanelli who volunteered many members as controls as well as Dr Dai Hickmann, Ms Fran Griffiths, Dr Alan J Williams and Mr Jamie O'Grady for allowing access to their Primary Care COPD registers. Dr Rodríguez was supported by a mobility grant from the University of Burgos, Spain.

References

1. Basanta, M., Jarvis, R.M., Xu, Y., Blackburn, Tal-Singer, R., Woodcock, A., Singh D., Goodacre, R., Paul Thomas, C. L., & Fowler, S.J. (2010). Non-invasive metabolomic analysis of breath using differential mobility spectrometry in patients with chronic obstructive pulmonary disease and healthy smokers, *Analyst 135*, 315–320.
2. Bajtarevic, A., Ager, C., Pienz, M., et al. (2009). Noninvasive detection of lung cancer by analysis of exhaled breath. *BMC Cancer 9*, 348, doi:10.1186/1471-2407-9-348.
3. Barker .M, Hengst M., Schmid J., et al.(2006). Volatile organic compounds in the exhaled breath of young patients with cystic fibrosis *European Respiratory Journal 27*, 929-36.
4. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning,* Springer.
5. Breiman, L. (1996). Bagging predictors, *Machine Learning, 26*, 123-140.
6. Breiman, L. (2001). Random Forests, *Machine Learning, 45*, 5-32.
7. Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*, Wadsworth International.
8. Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery, 2*, 121-167.
9. Cohen, W.W. (1995). Fast effective rule induction, *Proceedings of the Twelfth International Conference on Machine Learning*, Morgan Kaufmann, 115-123.
10. Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification,* John Wiley & Sons.
11. Fens, N., Zwinderman, A.H., van der Schee, M.P., de Nijs S.B., Dijkers, E., Roldaan, A.C., Cheung, D., Bel, E.H., & Sterk P.J. (2009). Exhaled breath profiling enables discrimination of chronic obstructive pulmonary disease and asthma. *American Journal of Respiratory and Critical Care Medicine*, *180*, 1076-1082.
12. Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences, 55*, 119-139.
13. Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2006). Gene selection for cancer classification using support vector machines, *Bioinformatics (Oxford, England) 22* (19), 2348--2355.
14. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update, *SIGKDD Explorations, 11*.
15. Hanczar, B., Hua, J., Sima, Weinstein, C.J., Bittner, M., & Dougherty E.R. (2010). Small-sample precision of ROC-related estimates, *Bioinformatics (Oxford, England) 26* (6), 822--830.

16. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, Springer.
17. Ho, T. K. (1998). The random space method for constructing decision forests, *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*, 832-844.
18. Ibrahim, B., Basanta, M., Cadden, P., Singh, D., Douce, D., Woodcock, A., & Fowle, S.J. (2011). Non-invasive phenotyping using exhaled volatile organic compounds in asthma. *Thorax 66*, 804-809.
19. Kuncheva, L. I. (2004). *Combining Pattern Classifiers. Methods and Algorithms*, John Wiley and Sons.
20. Machado, R.F., Laskowski, D., Deffenderfer, O., Burch, T., Zheng, S., Mazzone, P.J., Mekhail, T., Jennings, C., Stoller, J.K., Pyle, J., Duncan, J., Dweik, R.A., Erzurum, S.C. (2005). Detection of lung cancer by sensor array analyses of exhaled breath. *American Journal of Respiratory and Critical Care Medicine, 171* (11), 1286-1291.
21. Pauling L., Robinson A.B., Teranishi R., & Cary P., (1971). Quantitative analysis of urine vapor and breath by gas-liquid partition chromatography, *Proceedings of the National Academy of Sciences USA, 68*, 2374--2376.
22. Phillips, C.O., Syed, Y., Mac Parthalain, N., Zwiggelaar, R., Claypole, T.C., Lewis, K.E. (2012). Machine learning methods on exhaled volatile organic compounds for distinguishing COPD patients from healthy controls. Journal of Breath Research, 6 (3), doi: 10.1088/1752-7155/6/3/036003
23. Phillips, M., Cataneo, R.N., Cummin, A.R., Gagliardi, A.J., Gleeson, K., Greenberg, J., Maxfield, R.A., Rom W.N. (2003), Detection of lung cancer with volatile markers in the breath. *Chest, 123*(6), 2115-2123.
24. Phillips, M., Cataneo, R.N., Condos, R., Ericksond, G.A.R., Greenberga, J., La Bombardie, V., Munawara, M.I., & Tietjef, O. (2007). Volatile biomarkers of pulmonary tuberculosis in the breath, *Tuberculosis 87*, 44-52.
25. Rabe, K.F., Hurd, S., Anzueto, A., Barnes, P.J., Buist, S.A., Calverley, P., Fukuchi, Y., Jenkins, C., Rodriguez-Roisin, R., van Weel, C., & Zielinski, J. (2007). Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. GOLD Executive Summary, *American Journal of Respiratory and Critical Care Medicine*, *176*, 532–555.
26. Rodríguez, J. J., Kuncheva, L. I., & Alonso, C. J. (2006). Rotation Forest: A new classifier ensemble method, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*, 1619-1630.
27. Rogers, P.H., Benkstein, K.D., & Semancik, S. (2012) Machine learning applied to chemical analysis: Sensing multiple biomarkers in simulated breath using a temperature-pulsed electronic-nose, *Analytical Chemistry, 84* (22), 9774-9781.
28. Rokach, L., (2010) *Pattern Classification using Ensemble Methods*, World Scientific.
29. Sahin D., Übeyli E. D., Ilbay G., Sahin M. & Yasar, A.B. (2010). Diagnosis of airway obstruction or restrictive spirometric patterns by multiclass support vector machines, *Journal of Medical Systems 34*, 967–973.
30. Schapire, R.E., & Freund, Y. (2012). *Boosting. Foundations and Algorithms.* MIT Press.
31. Schapire, R.E., & Singer.,Y. (1999). Improved boosting algorithms using confidence-rated predictions, *Machine Learning, 37*, 297–336.
32. Ulanowska, A., Kowalkowski, T., Trawińska, E., & Buszewski B. (2011) The application of statistical methods using VOCs to identify patients with lung cancer, *Journal of Breath Research, 5* (4), doi:10.1088/1752-7155/5/4/046008.

33. Van Berkel, J.J.B.N., Dallinga, J. W., Möller G.M., Godschalk, R.W.L., Moonen, E., Wouters, E.F.M. , Van Schooten, F.J. (2008). Development of accurate classification method based on the analysis of volatile organic compounds from human exhaled air, *Journal of Chromatography B, 861*, 101–107.
34. Van Berkel, J.J.B.N., Dallinga, J. W., Möller G.M., Godschalk, R.W.L., Moonen, E., Wouters, E.F.M. , Van Schooten, F.J. (2010). A profile of volatile organic compounds in breath discriminates COPD patients from controls, *Respiratory Medicine, 104* (4), 557-563.
35. Webb G. I. (2000). MultiBoosting: A technique for combining boosting and wagging, *Machine Learning40* (2), 159–196.
36. WHO. Chronic respiratory diseases. World health statistics. (2008). Retrieved January, 17, 2013 from http://www.who.int/respiratory/copd/World_Health_Statistics_2008/en/index.html.
37. Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann.
38. Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms,* Boca Raton, FL, Chapman & Hall.