A Novel Method for Automatic Identification of Respiratory Disease from Acoustic Recordings

Xuen Hoong Kok, Syed Anas Imtiaz and Esther Rodriguez-Villegas

***Abstract*— This paper evaluates the use of breath sound recordings to automatically determine the respiratory health status of a subject. A number of features were investigated and Wilcoxon Rank Sum statistical test was used to determine the significance of the extracted features. The significant features were then passed to a feature selection algorithm based on mutual information, to determine the combination of features that provided minimal redundancy and maximum relevance. The algorithm was tested on a publicly accessible respiratory sounds database. With the testing dataset, the trained classifier achieved accuracy of 87.1%, sensitivity of 86.8% and specificity of 93.6%. These are promising results showing the possibility of determining the presence or absence of respiratory disease using breath sounds recordings.**

I. INTRODUCTION

Respiratory diseases are estimated to affect in excess of 1 billion people in the world, out of whom 4 million per year suffer from premature mortality [1]. This high prevalence together with rising healthcare costs and the loss of productivity, not just from patients themselves, but from those caring for them, have massive negative socio-economic impact globally [1].

The most common method for early diagnosis of respiratory disease is auscultation. Auscultation is the process of listening to sounds generated by the heart and lungs with a stethoscope. It is a non-invasive and straight-forward procedure, and hence, because of this, it is possibly the most widely used diagnostic method in medicine. However, to draw meaningful conclusions and perform diagnosis, auscultation requires the practitioner to undergo training and have prior experience in the field. Based on the findings from auscultation, further tests can be prescribed to confirm or refine the diagnosis.

Breathing sounds can be separated into two categories: normal and abnormal (adventitious). Examples of abnormal sounds include stridors, wheezes, crackles and rhonchi. The presence of adventitious respiratory sounds have been shown to indicate underlying respiratory conditions. For instance, the occurrence of wheezes and crackles have been linked to diseases such as asthma and Chronic Obstructive Pulmonary Disease (COPD) [2].

Unfortunately, considering the highly subjective nature of the auscultation process, unintentional misdiagnosis is not unusual, if performed by an unskilled or inexperienced person [3]. This can lead to patients not receiving the appropriate treatment and care, and consequently the deterioration of their condition. Many studies have investigated automatic and computerized methods to detect the presence of abnormal lung sounds and reduce the subjectivity of the process [4]. The drawback of reported methods, which is also a problem with short time auscultation, is that they are based on individual breath segments or phases. Breath segments typically last between 1.5 to 3 seconds long, and breath phases are even shorter. Moreover, wheezes and crackles are also occasionally present in breathing sounds of subjects who do not have respiratory diseases [5]. Hence, the presence of adventitious sounds in breath segments is insufficient to determine if the subject suffers from respiratory disease.

An alternative that has not been thoroughly explored is the use of longer term acoustic breath sound recordings acquired from wearable continuous monitoring devices. Longer recordings will provide a clearer picture if the occurrences of abnormal lung sounds are isolated incidents or if they are indicative of a respiratory condition.

This paper aims to show the feasibility of using acoustic breath recordings to determine if a subject suffers from respiratory disease. Section II discusses the database and methods used in the experiment. Section III examines the features explored, and the selection of features using hypothesis testing and a mutual information (MI) based method. Section IV provides an overview of the classification algorithm used, as well as the classifier evaluation metrics. Section V presents the results obtained from testing the algorithm and the discussion.

II. MATERIALS AND METHODS

The database used in this paper was obtained from the 2017 ICBHI Challenge [6]. This database consists of 920 recordings obtained from two different research teams and contains breath sounds recordings from 26 healthy subjects as well as 100 patients suffering from a range of respiratory diseases, such as Chronic Obstructive Pulmonary Disease (COPD), bronchitis, lower and upper respiratory tract infections. The recordings were also segmented and annotated by healthcare professionals to indicate the presence of wheezes, crackles or both, wheezes and crackles, in each segment. The average duration of the recordings in this database is 21.5 seconds.

The number of recordings from the respiratory disease patient group is 885. The average duration of these recordings is 21.6 seconds. Annotation files for 259 of these recordings do not indicate the presence adventitious sounds. The median

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC), UK / grant agreement no. EP/P009794/1.

X. H. Kok, S. A. Imtiaz and E. Rodriguez-Villegas are with the Wearable Technologies Lab, Department of Electrical and Electronic Engineering, Imperial College London, SW7 2BT, United Kingdom. E-mail: {x.kok17, anas.imtiaz, e.rodriguez}@imperial.ac.uk

number of segments of all annotated abnormal sounds per recording is 3, with a standard deviation of 3.3. Individually, the number of labelled wheezes, crackles and both wheezes and crackles are 0 ± 2.1, 0 ± 2.9 and 0 ± 1.5 respectively. The remaining 35 recordings were obtained from healthy subjects, with an average duration of 20 seconds. Seven of these contain adventitious respiratory sounds. Six recordings contained crackles, two had wheezes and one had both wheezes and crackles. These are summarized in Table I.

TABLE I
DETAILS OF HEALTHY SUBJECT RECORDINGS CONTAINING ADVENTITIOUS SOUNDS

Subject No.	Recording No.	#Crackles	#Wheezes	#Both	#Normal
113	66	2	0	0	4
114	67	3	0	0	7
147	302	0	1	0	11
162	455	2	0	1	5
	456	2	1	0	4
163	463	5	0	0	2
	464	2	0	0	2

A high level overview of the proposed method is shown in Fig. 1. The entire process can be broken down into three parts: pre-processing, feature extraction, and classification. The signal pre-processing steps include resampling the signals to 8000 Hz, to ensure a consistent sampling rate is used across all recordings. The resulting signal is then passed through a band-pass filter to retain components in the range of 100 to 2000 Hz, since these correspond to frequencies of breath sound signals [7]. As the data were collected from different sources and devices, the recordings were all normalized to the range [-1,1] before features were extracted.

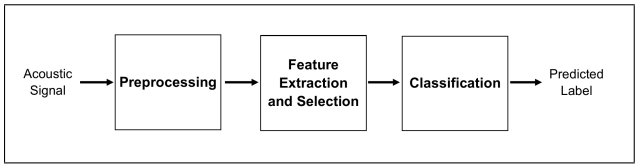

Fig. 1. High level overview of the proposed method

III. FEATURE EXTRACTION AND SELECTION

The abnormal sounds annotated in the database, i.e. adventitious sounds, are sounds which appear superimposed onto normal respiratory ones. But, in addition, lung sounds in patients with respiratory disease also have a lower intensity [2]. The additional frequency components as well as reduction in amplitude support the hypothesis that the right chosen features would have differing values between healthy subjects and respiratory disease patients. The following features were explored as potential candidates, to lead to an algorithm which would distinguish recordings from healthy and respiratory disease patients.

A. Features Extracted

1) Mel-Frequency Spectral Coefficients (MFCC): MFCC is widely used in the domain of automatic speech recognition. The success of this feature can be attributed to its ability to model the features and characteristics of the human auditory system. This was the reason why it was chosen as a candidate feature, since adventitious sounds can often be heard with auscultation. Thirteen coefficients were extracted along with their differential coefficients, also known as delta coefficients. The delta term measures the rate of change of the coefficients. Hence, its inclusion can allow for more information about the signal dynamics to be included for analysis and training. The library [8] was used in the extraction of MFCC features.

2) Discrete Wavelet Transform (DWT): From the characteristics of respiratory disease recordings, the presence of certain frequencies can indicate the existence of adventitious sounds. Lung sounds contain frequencies ranging from 100 Hz to 2500 Hz, while adventitious sounds such as wheezes and crackles are generally between 100 and 1000 Hz [7].

The wavelet transform is a method that allows analysis of signals in the time-frequency domain and the spectral components in the recordings can be mapped with time. This is useful in breath sound recordings as it allows abnormal sounds to be localized.

The wavelet transform coefficients indicates the similarity between a signal and its analyzing wavelet, the greater the resemblance the higher the value of coefficients [9]. The analyzing wavelet selected was hence based on the likeness to the adventitious sounds that are present in the breath recordings. As adventitious breath sounds only occur periodically, and depending on whether these sounds are present, the resulting coefficients will fluctuate over time. It is because of this that the wavelet transform coefficients were considered as a candidate features in this study. The discrete wavelet transform, in particular, was used, as the result is less redundant than the continuous form.

A 6-level decomposition was performed on the pre-processed signal using a range of different analyzing wavelets (Daubechies 1 to 10, Symlet 2 to 10 and Coiflet 1 to 5). The detail coefficients (cD) of levels 2 to 5 were retained for further analysis as they represented frequency ranges of lung sounds, as described earlier. Each level corresponds to frequencies 1000-2000 Hz, 500-1000 Hz, 250-500 Hz and 125-250 Hz, respectively. To quantify the changes in the coefficients, statistical features describing the retained coefficients were calculated. They included energy, entropy, bounded variation, and variance.

3) Time Domain Features: In diseased patient recordings with no adventitious sounds, the coefficients from the wavelet transforms would be similar to those from a healthy subject with no abnormal sounds. However, it is still possible to tell them apart due to the differences in their amplitudes in the time domain. To explore the differences between the two subject groups, the power, mean, variance, skewness and kurtosis were measured. The power of the acoustic signals in the diseased group was expected to be lower than in the healthy subjects group. The variance, skewness and kurtosis are the second, third and fourth higher order statistical features that are used to describe the signal with respect to the mean.

B. Feature Selection

The Wilcoxon Sum of Rank test (also known as the Mann-Whitney U-Test) was used to determine if and which of the features extracted were statistically significant. This test was suitable because the data were not normally distributed and had a small number of samples in one of the classes [10]. The null hypothesis for the test was that there is no difference in the extracted features between healthy subjects and patients diagnosed with respiratory disease. Features with a p-value of less than 5% were considered statistically significant and the null hypothesis could be rejected in favour of the alternative hypothesis.

As shown in Table II, eight of the most significant (i.e. lowest p-values) features extracted were the MFCC Delta coefficients, followed by the energy and variance in the 5th level DWT coefficient. The coefficients from this level corresponded to the frequency range of 125 Hz to 250 Hz, which is within the range where most of the frequency components of adventitious lung sounds are located. All the features that were found to be statistically significant were then tested with a minimum Redundancy Maximum Relevance (mRMR) test [11]. The algorithm would find the most relevant feature and each additional feature would be ranked according to their redundancy. The top 20 features from the mRMR test are shown in Table III.

TABLE II

15 MOST SIGNIFICANT FEATURES FROM RANK SUM TEST IN ASCENDING ORDER

Features Names	p–Values
5th MFCC Delta Coefficient	1.68E-13
1st MFCC Delta Coefficient	1.00E-12
3rd MFCC Delta Coefficient	2.02E-12
2nd MFCC Delta Coefficient	4.61E-11
11th MFCC Delta Coefficient	1.05E-09
4th MFCC Delta Coefficient	2.45E-09
8th MFCC Delta Coefficient	1.21E-08
12th MFCC Delta Coefficient	2.41E-08
sym7 cD_5 Energy	2.68E-08
sym7 cD_5 Variance	2.68E-08
db8 cD_5 Energy	2.75E-08
db8 cD_5 Variance	2.76E-08
sym10 cD_5 Energy	3.84E-08
sym10 cD_5 Variance	3.86E-08
coif3 cD_5 Variance	3.87E-08

IV. CLASSIFICATION

The classification task was performed using the RUSBoost algorithm [12]. RUSBoost is a combination of random under sampling (RUS) and Boosting technique. This algorithm was chosen because the combination of these two techniques is particularly useful for dealing with datasets that have imbalanced classes.

The number of iterations (T) used for all the trained classifiers in this paper was 30. Two minority class proportions (N) were examined, specifically 35% and 50%. The base classifier used was a Decision Tree.

The performance of the trained classifier was evaluated using the accuracy, sensitivity and specificity metrics. For the purpose of these metrics, a True positive (TP) was defined as the correct identification of a recording corresponding to respiratory disease, and True Negative (TN) as the correct identification of a recording corresponding to no respiratory disease. A healthy recording classified as diseased was a false positive (FP), and a diseased recording wrongly classified as healthy was a False Negative (FN). These metrics are mathematically defined as:

TABLE III

FIRST 20 FEATURES RETURNED BY THE MIN-REDUNDANCY MAX-RELEVANCE (MRMR) ALGORITHM

Features Names	rank
1st MFCC Delta Coefficient	1
sym4 cD_3 Energy	2
Kurtosis	3
11th MFCC Coefficient	4
db3 cD_5 Energy	5
Arithmetic Mean	6
db2 cD_2 Coefficient	7
db9 cD_5 Coefficient	8
8th MFCC Delta Coefficient	9
db9 cD_2 Bounded Variation	10
db6 cD_3 Coefficient	11
4th MFCC Delta Coefficient	12
6th MFCC Delta Coefficient	13
coif1 cD_4 Energy	14
13th MFCC Delta Coefficient	15
db4 cD_2 Coefficient	16
8th MFCC Coefficient	17
sym7 cD_5 Bounded Variation	18
2nd MFCC Delta Coefficient	19
sym10 cD_5 Energy	20

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

V. RESULTS AND DISCUSSION

The classification results from the trained models are presented in this section. The first part presents the results obtained with the individual group of features (i.e. MFCC, DWT and Time Domain) whereas the second presents the results from models using the significant features, as selected using the Wilcoxon Rank Sum test and mRMR algorithm.

The models trained with MFCC and its derivative performed the best among the groups of features; achieving accuracy, sensitivity and specificity of 87.5%, 88.1% and 87.3% respectively. The comparison of different feature types is presented in Table IV and Table V. The trained models using features from the mRMR test are better than those using individual types of features, as presented in Table VI and Table VII, particularly in the cases of time domain and DWT coefficient features. With MFCC however, there is only a marginal improvement when twenty of the top mRMR ranked features were used in classification.

The models trained with minority class proportion (N) set at 35% resulted in a higher sensitivity than those using 50%. As the dataset contained a significantly larger number of recordings from patients with respiratory disease, a lower N means more samples from the respiratory disease group were used in the training of individual base classifiers. The results showed that, in general, when the value of sensitivity increases, specificity decreases (or vice versa) as a consequence. As a result, the decision on the choice of the classifier would very much be linked to the final intended clinical use of the algorithm.

The proposed algorithm can potentially be of use for early or timely diagnosis of respiratory diseases. Chronic respiratory diseases are generally slow-developing and symptoms in the early stages are infrequent and less pronounced. It is also at this stage that the patient is more likely to be underdiagnosed. In COPD for example, the misdiagnosis rate is higher when done by general practitioners (GPs) as compared to experts [3]. For these misdiagnosed patients, the implications are severe. They would not have access to the appropriate care, education and treatment to control the disease progression. With GPs being the first point of contact, it is hence crucial that they are supported with tools to conduct more accurate diagnosis.

TABLE IV
CLASSIFICATION RESULT (MINORITY CLASS RATIO, N = 0.5)

	MFCC	DWT	TD
Accuracy (%)	85.5	63.1	62.9
Sensitivity (%)	85.4	62.8	62.8
Specificity (%)	87.3	70.9	64.6

TABLE V
CLASSIFICATION RESULT (MINORITY CLASS RATIO, N = 0.35)

	MFCC	DWT	TD
Accuracy (%)	87.5	72.9	74.5
Sensitivity (%)	88.1	73.7	75.5
Specificity (%)	74.6	55.5	51.8

TABLE VI
CLASSIFICATION RESULT BASED ON THE NUMBER OF RANKED MRMR FEATURES USED (MINORITY CLASS RATIO, N = 0.5)

	Num. Features			
	5	10	15	20
Accuracy (%)	81.7	82.9	82.6	87.1
Sensitivity (%)	81.2	82.6	82.3	86.8
Specificity (%)	91.8	89.1	89.1	93.6

VI. CONCLUSIONS

This paper presented an algorithm that uses breath sound recordings to determine whether or not a particular recording originates from a patient with respiratory disease. The best trained model achieved sensitivity of 86% and specificity of 93% respectively, showing the potential of using breath recordings to determine if a person might be suffering from respiratory diseases. An area for improvement that would further increase the accuracy of the proposed algorithm would be to explore the optimum length of breath recordings that would maximize the detection accuracy. The recordings used in this paper had an average duration of 21.5 seconds. Although this is much longer than individual breath segments, other recording duration might work better. Another aspect that should be investigated is the conditions in which breath signals are recorded. Different conditions may affect the results since factors such as ambient noises, speech and the position of the patient during the recordings could affect both the characteristics, as well as the quality of the recorded signals.

TABLE VII
CLASSIFICATION RESULT BASED ON THE NUMBER OF RANKED MRMR FEATURES USED (MINORITY CLASS RATIO, N = 0.35)

	Num. Features			
	5	10	15	20
Accuracy (%)	84.7	86.7	86.9	89.4
Sensitivity (%)	84.8	87.1	87.4	89.7
Specificity (%)	80.0	75.5	73.6	81.8

REFERENCES

[1] Forum of International Respiratory Societies, *The Global Impact of Respiratory Disease Second Edition*. Sheffield: European Respiratory Society, 2017.

[2] A. Bohadana, G. Izbicki, and S. S. Kraman, "Fundamentals of lung auscultation," *N Engl J Med*, vol. 370, no. 8, pp. 744–51, 2014.

[3] S. Hangaard, T. Helle, C. Nielsen, and O. K. Hejlesen, "Causes of misdiagnosis of chronic obstructive pulmonary disease: A systematic scoping review," *Respir Med*, vol. 129, pp. 63–84, 2017.

[4] R. Palaniappan, K. Sundaraj, and S. Sundaraj, "A comparative study of the SVM and k-nn machine learning algorithms for the diagnosis of respiratory pathologies using pulmonary acoustic signals," *BMC Bioinformatics*, vol. 15, no. 1, p. 223, Jun 2014.

[5] R. L. Murphy, "In defense of the stethoscope," *Respiratory Care*, vol. 53, no. 3, pp. 355–369, 2008.

[6] B. Rocha, D. Filos, L. Mendes, I. Vogiatzis, E. Perantoni, E. Kaimakamis, P. Natsiavas, A. Oliveira, C. Jácome, A. Marques, *et al.*, "A respiratory sound database for the development of automated classification," in *Precision Medicine Powered by pHealth and Connected Health*. Springer, 2018, pp. 33–37.

[7] S. Reichert, R. Gass, C. Brandt, and E. Andres, "Analysis of respiratory sounds: state of the art," *Clin Med Circ Respirat Pulm Med*, vol. 2, pp. 45–58, 2008.

[8] D. P. W. Ellis. (2005) PLP and RASTA (and MFCC, and inversion) in Matlab. Online web resource. [Online]. Available: http://www.ee.columbia.edu/ dpwe/resources/matlab/rastamat/

[9] The MathWorks, Inc. (1994-2019) Interpreting Continuous Wavelet Coefficients - MATLAB & Simulink. [Online]. Available: https://www.mathworks.com/help/wavelet/gs/interpreting-continuous-wavelet-coefficients.html

[10] N. Nachar, "The mann-whitney u: A test for assessing whether two independent samples come from the same distribution," *Tutorials in Quantitative Methods for Psychology*, vol. 4, no. 1, pp. 13–20, 2008.

[11] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans Pattern Anal Mach Intell*, vol. 27, no. 8, pp. 1226–38, 2005.

[12] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "Rusboost: A hybrid approach to alleviating class imbalance," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 40, no. 1, pp. 185–197, 2010.