Biomedical Signal Processing and Control 46 (2018) 94–103

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

A robust audio classification system for detecting pulmonary edema

K.J. Hong[a,b,*], S. Essid[a,b,**], W. Ser[a,b], D.C.-G. Foo[b,c]

[a] Nanyang Technological University, 50 Nanyang Ave, Singapore 639798, Singapore
[b] LTCI, Télécom ParisTech, Université Paris-Saclay, Paris 75013, France
[c] Tan Tock Seng Hospital, 11 Jalan Tan Tock Seng, Singapore 308433, Singapore

ARTICLE INFO

Article history:
Received 13 April 2018
Received in revised form 11 June 2018
Accepted 10 July 2018

Keywords:
Non-negative matrix factorization
Robust testing
Pulmonary edema
Biomedical signal processing
Feature learning

ABSTRACT

In this paper we present a robust audio classification system to efficiently detect pulmonary edema. The system uses a feature learning technique based on (NMF), then classified with logistic regression. A study was done to compare feature engineering approaches with feature selection techniques against NMF. Different NMF schemes were investigated and also compared with Principal Component Analysis. NMF scored 95% F1 score, which was superior to feature engineering techniques that had scores from 83% to 93%. Background noise collected from hospitals and speech from a speech corpus database was used to simulate noisy data. The system was then tested using noisy data. The best NMF scheme scored 74%, while other feature engineering techniques scored lower; from 66% to 71%. NMF was also used as a signal enhancement tool. It improved the F1 score to 77%. Lastly, only inhalations from breath sounds were considered and this further improved classification results to 86%. The proposed robust classification system using NMF thus proved to be an effective method for audio-based detection of pulmonary edema. If implemented in real-time, the proposed system can be used as a screening tool.

1. Introduction

Pulmonary edema is a condition where water engorges the alveolar beds. The water is pushed into air spaces and thus reduces normal oxygen movement. It can cause hemoptysis, difficulty in breathing, shorted of breath, and gurgling and wheezing sounds during breathing. It can be caused by congestive heart failure, kidney failure, high altitude exposure, lung damage, and major injuries.

Currently, physicians will carry out physical examinations before performing X-rays, CT-scans and electrocardiograms. Physicians mainly use auscultations to check for abnormal heart sounds, crackles, increased heart rate and rapid breathing. Most patients with pulmonary edema must face the disease and its implications for the rest of their lives. They must visit the physician for a check-up whenever they show symptoms, and if left untreated, they could suffer suffocation. These constant trips to the physician can add stress, worry and inconvenience. Physicians also in turn face a greater workload.

Nomenclature

ADC	analog-to-digital converter
CV	cross-validation
DAS	discontinuous adventitious sounds
ELW	excessive lung water
EN	elastic net
IS	Itakura–Saito
k-NN	k-nearest neighbour
KL	Kullback–Leibler
LPC	Linear predictor coefficients
LR	logistic regression
MFCC	Mel-frequency cepstral coefficients
MU	multiplicative update
NMF	non-negative matrix factorization
PAF	“popular” audio features
PCA	principal component analysis
RFE	recursive feature elimination
SN	sensitivity
SP	specificity
SVM	Support vector machine

Thus, there exists a need for an automatic, quick, accessible, and simple screening solution for the detection of pulmonary edema.

* Corresponding author at: Nanyang Technological University, 50 Nanyang Ave, Singapore 639798, Singapore.
** Principal corresponding author.
E-mail addresses: khong008@e.ntu.edu.sg (K.J. Hong), slim.essid@telecom-paristech.fr (S. Essid), ewser@ntu.edu.sg (W. Ser).

https://doi.org/10.1016/j.bspc.2018.07.004
1746-8094/

With this, patients need not solely rely on check-ups by physicians to ascertain excessive accumulation of water but will be able to carry out screening with greater convenience.

The paper proposes a more thorough study on the use of audio classification techniques for the automatic detection of excessive lung water. It introduces an original system which proves highly effective for this task. The system is tested and validated by augmented samples collected from a local hospital. This paper presents results and analysis of a proposed general framework that extracts, transforms, and classifies features for the detection of excessive lung water in real world sound recordings. Moreover, special care is taken to assess the robustness of the system in noisy recording conditions.

2. Previous work

Conventionally, auscultations are used to help diagnose a vast number of other conditions, as well as pulmonary edema. The physician pays attention to abnormal respiratory sounds called adventitious sounds. In the case of pulmonary edema, the physician listens to (DAS) called crackles. Thus, there has been much research developed to detect DAS found in lung sounds. Recently, Pramono et al. [23] presented a systematic review of automatic adventitious respiratory sound analysis. Regarding the analysis of crackles, they cited 36 papers that involve detection or classification and listed their data sources, amount of data, validation method, features used, classification method and performance. They concluded that performance of recent studies showed a high agreement with conventional non-automatic identification and suggested that automated adventitious sound detection or classification was a promising solution to overcome the limitations of conventional auscultation and to assist in the monitoring of relevant diseases. However, depending on the severity of the disease, these DAS might not manifest frequently or at all. Furthermore, studies such as [22,28,33,21,27,32] present event-based classification, where the DAS was isolated manually before classification. This approach would perform poorly if crackles do not appear in a particular segment of audio data. Also, the same DAS appear in various diseases. Fine crackles appear in patients with bronchiectasis, pulmonary edema, asthma, chronic bronchitis, severe sepsis, pneumonia, congestive heart failure, pulmonary fibrosis, and acute bronchitis. Consequently, the classification of DAS does not systematically translate to the classification of diseases. Also, the aim of the proposed study is to detect pulmonary edema in any segment of the breath sound, and not just segments with crackles.

On the other hand, there have been a few works that involved classifying diseases using lung sounds. Hernandez et al. [15] presented a classification system for recordings of lung sounds with diffused interstitial pneumonia against healthy subjects. The authors used a multivariate auto-regressive model with a supervised neural network as a classifier. [24] presented a system for automated diagnosis of pertussis using audio signals by analyzing cough and whoop sounds. Yang and Wee first proposed a signal processing approach for detecting excessive lung water using sound-based sensing [31]. Different features and classifiers were explored, and Mel-frequency cepstral coefficients (MFCC) together with k-nearest neighbor (k-NN) produced the best result. In our previous work [7], we proposed a feature extraction method that segmented the magnitude spectrum of lung sound recordings into sub-bands. The sub-band spectral coefficients were used as features that were ranked using principle component analysis and support vector machine-based recursive feature elimination. classification was done using k-NN and Support Vector Machine (SVM). *F*-measure of up to 99% was reported. However, although the sample size was small, the use of appropriate testing and validation procedures was lacking, which puts into question the statistical validity of the results and the generalization ability of the previously proposed system. Also, the system hyperparameters were not optimized. Lastly, there was no study of the robustness to noise.

Secondly, all studies cited use different kinds of feature engineering techniques to obtain features. For example, the standard choice for speech and audio classification task, MFCC [10,4], was used by [25,18,29,19,17]. All proposed feature engineering methods as part of their solutions had success of various degrees. However, in recent years feature learning has become the dominant trend in machine learning problems. Similar to learning in classifiers, feature learning entails algorithms automatically learning features determined by data. It is an alternative to feature engineering (as shown in Fig. 1). Deep learning [11] has been a popular choice for many classification and regression problems. It also has been used in audio related problems [2,13]. However, deep learning methods require huge amounts of data, and thus for applications with small datasets, signal specific feature engineering is preferred. Nevertheless, due to the nature of this problem, the sample size involved in this work is small and despite the popularity of deep learning, it could not be implemented successfully. Still, research such as [1] has shown that another popular method called non-negative matrix factorization (which has been successful in audio source separation and enhancement [12,26,20]) had success similar to deep learning methods in the audio classification domain when there is a lack of training data. Also, NMF was used in [14] for blind source separation of heart and lung sounds. Thus, we adopt NMF as a feature learning method in this paper.

Thus, we propose a robust system for the detection of pulmonary edema using classical classification techniques as well as state-of-the-art non-negative matrix factorization for feature learning. This is an alternative to indirectly detecting crackles to screen for pulmonary edema. The system was also tested under the introduction of environmental noise found in hospitals. The system hyperparameters were optimized, and cross-validation was used to reliably assess its performance.

3. Methodology

Fig. 1 shows the flow of data processing as a chart. First, the breathing sounds of the subjects were recorded using a stethoscope sensor. Next, synthetic data containing recordings of excessive water were generated so that the number of ill patient samples and healthy subjects were approximately equal. On top of that, synthetic test data was generated to assess the generalization ability of our system and improve the reliability of the evaluation metrics. A skewed ratio between the two classes would result in a biased sensitivity, specificity and *F*-measure. More details are in Section 4. Data was segmented into 250 ms segments and normalized. Features of the data were then extracted. The system was tested with and without feature selection and feature transformation techniques. The logistic regression (LR) classifier was trained using training data and the class probability was obtained using testing data. Finally, the class of each test data point was predicted.

It is common to transform or select some of the initial features to increase final classification accuracies. In the next subsection, a brief introduction of the techniques used in the feature extraction, transformation and classification steps are given.

3.1. Feature extraction

First, we consider various of popular audio features to compare with our NMF-based approach. Table 1 shows the list of these

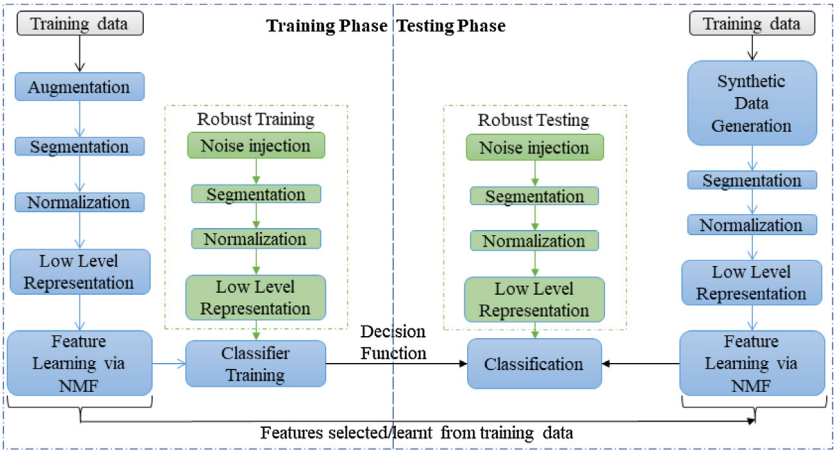

Fig. 1. Data flow chart for proposed classification system. The system was tested with and without feature transformation. The robustness testing block was used only when the system was tested for robustness. The system was also tested with and without NMF signal enhancement.

Table 1
Features extracted by Yaafe.

Name of feature	Features
Linear predictor coefficients	2
Line spectral frequency coefficients (LSF)	10
1st and 2nd derivatives of LSF	20
Mel-frequency cepstrum coefficients (MFCC)	13
1st and 2nd derivatives of MFCC	26
Loudness coefficients (energy in each Bark band)	24
Sharpness of the loudness coefficients	1
Spectral decrease	1
Global spectral flatness	1
Flux of magnitude spectrum	1
Centroid, spread, skewness and kurtosis of magnitude spectrum	4
Shape of spectral amplitude by linear regression	1
Centroid, spread, skewness and kurtosis of temporal waveform	4
Zero crossing rate of temporal waveform	1

features. The interested reader is referred to [16] for more information about the feature extraction methods. These features will be referred to as (PAF). All features given in Table 1 were extracted by Yaafe [16]. They are shortly described in the following paragraph.

Linear predictor coefficients (LPC) are parameters from a model of a signal as the output of an all-pole filter with a white spectral input.

Line spectral frequency coefficients are a transformation of the LPC to achieve greater interpolation properties and robustness to quantization.

MFCC is a popular transformation in speech and audio analysis that compactly represent the shape of the vocal tract manifestations in the envelope of a short time power spectrum. 40 triangular filters and 13 coefficients were chosen for the Mel-bank and MFCC coefficients respectively.

Loudness coefficients are the energies of 24 bands of the Bark scale. The Bark scale correspond to the first 24 critical bands of hearing. The bands are normalized such that the sum is 1.

Sharpness is a hearing sensation related to frequency and independent of loudness. It corresponds to the sensation of a sharp, painful, high-frequency sound and is the comparison of the amount of high frequency energy to the total energy.

The spectral decrease is weighted average of the decrease of the magnitude spectrum between consecutive frames.

The global spectral flatness is the ratio between the geometric mean and the arithmetic mean of the spectrum.

MFCC alone was also used as features for this system to compare with our previous work. All features were extracted from overlapping analysis windows of length 1024 samples using a hop-size of 512 samples (at a 8000 Hz sampling rate).

3.2. Feature selection methods

Two popular standard feature selection methods were used in this paper. The first was recursive feature elimination (RFE) [5]. With RFE, a decision function with weights is fitted with a classifier. RFE ranks features by recursively eliminating those that have the lowest weights in the decision function. The classifier is run repeatedly as features are eliminated until all the features have been ranked. In this paper, logistic regression was used as the classifier in the RFE procedure. After ranking, the best number of features were selected using k-fold cross validation.

The second method was the elastic net used with logistic regression (EN). It is an embedded feature selection and classification method. The estimate of the regression coefficients β is $\hat{\beta} = argmin_{\beta}\left[L_s\left(y, \mathbf{X}\beta\right) + \alpha\left(\lambda\left\|\beta\right\|^2 + (1-\lambda)\left\|\beta\right\|\right)\right]$ where $L_s$is the logistic loss, α is the regularization strength and λ is the ratio between ℓ_1 and ℓ_2 regularizations. The interested reader can refer to [6] for more information.

3.3. Principal component analysis

Principal component analysis (PCA) was used in this study as a benchmark to evaluate the effective of our NMF-based method. PCA [9] is a feature transformation technique that identifies a smaller number of uncorrelated variables, called *principal components*, from a large set of data. The goal of principal component analysis is to explain the maximum amount of variance with the fewest number of principal components. It achieves this by the diagonalization of the covariance matrix estimate of the data to transform data to orthogonal basis vectors. Features of the transformed data can then be ranked according to the eigenvalues along the transformed axis.

3.4. Non-negative matrix factorization

Non-negative matrix factorization (NMF) [12] decomposes non-negative data observations assembled as the N columns of the matrix $\mathbf{V}$, here magnitude spectrum of the audio signals, using non-negative dictionary elements and their expansion coefficients. Two non-negative matrices, $\mathbf{W}$ and $\mathbf{H}$, are estimated jointly so as to approximate the data matrix $\mathbf{V} \approx \mathbf{WH}$. $\mathbf{W}$ is called the dictionary and has dimensions $F \times K$, where F is the number of features and K is the number of basis vectors. Each column of $\mathbf{W}$, $\boldsymbol{w}_k \quad 1 \le k \le K$, are basis

vectors that are an alternative representation of data $\mathbf{V}$ in a newly projected space. $\mathbf{H}$, the activation matrix, has dimensions $K \times N$, where N is the number of samples. Each element of $\mathbf{H}$, $h_{kn;} 1 \le k \le K$, $1 \le n \le N$, contains the weight of $\boldsymbol{w}_k$ for the nth observation. Thus the activation matrix can be interpreted as a transformation of the data with a set of K features. Each column of $\mathbf{V}$ (each sample) can be written as a weighted sum of the basis vectors and represented by $\mathbf{H}$ and $\mathbf{W}$ such that $\boldsymbol{v}_n \approx \sum_{k=1}^{K} h_{kn}\boldsymbol{w}_k \quad s.t. \quad \succeq 0$.

NMF is calculated by minimizing an objective function:

$$\min_{\mathbf{W},\mathrm{H}} = D_\beta(\mathbf{V} \mid \mathrm{WH}) \quad s.t. \quad \mathrm{W}, \mathbf{H} \succeq 0, \tag{1}$$

where D_β is a separable divergence in the beta-family of divergences and $D\left(\mathbf{V} \mid \hat{\mathrm{V}}\right) = \sum_{f=1}^{F}\sum_{n=1}^{N} d\left(v_{fn} \mid \hat{v}_{fn}\right)$, where $\hat{\mathrm{V}} = \mathbf{WH}$ and v_{fn} is the f^{th} feature for sample n. In this paper, three divergences from the β family were investigated. The three divergences used were the ℓ_2, Kullback–Leibler (KL), and Itakura–Saito (IS) divergence.

The ℓ_2 divergence ($\beta = 2$) is defined as follows:

$$d_{\ell_2}(v_{fn}, \hat{v}_{fn}) := \left(v_{fn} - \hat{v}_{fn}\right)^2. \tag{2}$$

The KL divergence ($\beta = 1$) is:

$$d_{KL}(v_{fn}, \hat{v}_{fn}) := v_{fn} \log \frac{v_{fn}}{\hat{v}_{fn}} - v_{fn} + \hat{v}_{fn}. \tag{3}$$

The IS divergence ($\beta = 0$) is:

$$d_{IS}(v_{fn}, \hat{v}_{fn}) := \frac{v_{fn}}{\hat{v}_{fn}} - \log \frac{v_{fn}}{\hat{v}_{fn}} - 1. \tag{4}$$

The multiplicative update (MU) approach [12] is used to minimize the objective function. An iteration of the MU is

$$\mathbf{H} \longleftarrow \mathbf{H} \odot \frac{\mathbf{W}^T\left((\mathbf{WH})^{[\beta-2]} \odot \mathbf{V}\right)}{\mathbf{W}^T(\mathbf{WH})^{[\beta-1]}} \tag{5}$$

$$\mathbf{W} \longleftarrow \mathbf{W} \odot \frac{\left((\mathbf{WH})^{[\beta-2]} \odot \mathbf{V}\right)\mathbf{H}^T}{(\mathbf{WH})^{[\beta-1]}\mathbf{H}^T} \tag{6}$$

and the matrices $\mathbf{W}$ and $\mathbf{H}$ are initialized either randomly or using singular value decomposition of the data. In our paper, we initialize these matrices randomly 5 times and choose the final $\mathbf{W}$ and $\mathbf{H}$ with the lowest total divergence. The MU update rule is known to steadily decrease the cost although there is no guarantee of convergence as the problem is known to be non-convex. Hence in practice, the algorithm is run until a stopping criterion is met [12]. In this paper, the MU update rule ceases when the fractional change of the cost between each iteration is less than 10^{-4}. The activation matrix of a dataset can also be calculated with a given dictionary $\mathbf{W}$ by fixing it in the MU update step. Similarly, the dictionary of a dataset calculated with a given activation matrix. ℓ_2-norm can also be added as a regularization term as done in ridge regression. The objective function then becomes $\min_{\mathbf{W},\mathrm{H}} \left[D_\beta(\mathbf{V} \mid \mathrm{WH}) + \phi\|\mathbf{H}\|_2^2 + \phi\|\mathbf{W}\|_2^2\right] \quad s.t. \quad \mathrm{W}, \mathbf{H} \succeq 0$ where ϕ is the regularization strength. After applying NMF, the activation matrix is used as features for classification. Initial tests showed that ℓ_2-norm regularization consistently performed better than without regularization and thus in this paper, ℓ_2-norm regularization was used in all NMF calculations. To calculate the activation matrix of the test data, $\mathbf{W}$ and $\mathbf{H}$ are first jointly calculated from the training data. Using the pre-trained dictionary $\mathbf{W}$, the activation matrix $\check{\mathrm{H}}$ of the test data is computed by fixing the dictionary of the test data $\check{\mathrm{W}} = \mathbf{W}$ for every iteration in the MU approach and only using Eq. (6).

3.5. NMF feature learning for noisy training data

It is common practice to include noisy training data together with clean training data to increase classification accuracy of noisy test data. Conventionally, features are extracted or learnt using a combined set of clean and noisy training data and trained by the classifier. Fig. 2 shows the steps involved to learn features from a mix of clean and noisy training data. The dictionary of clean training data was used to obtain the activation matrix for noisy training and test data. The activation matrices were concatenated and standardized to train the classifier.

3.6. NMF as a signal enhancement tool

NMF can also be used for signal enhancement in the presence of noise [30]. Fig. 3 shows the steps involved for enhancing samples. First, $\tilde{\mathbf{W}}$ is learned using training noise $\tilde{\mathbf{V}}$ as described in Section 3.4. $\mathbf{W}$ is also learned using clean training data samples $\mathbf{V}$, where $\tilde{\mathbf{W}}$ and $\mathbf{W}$ are dictionaries for noise and clean data samples respectively. A new dictionary, $\bar{\mathbf{W}}$ representing the noisy data, is created by concatenating $\mathbf{W}$ and $\tilde{\mathbf{W}}$ such that $\bar{\mathbf{W}} = \left[\mathrm{W} \mid \tilde{\mathbf{W}}\right]$. Next, activation matrix $\bar{\mathbf{H}}$ for enhanced data is calculated using $\bar{\mathrm{V}} \approx \bar{\mathrm{W}}\bar{\mathrm{H}}$, fixing $\bar{\mathbf{W}}$ during multiplicative updates. In [30] only $\bar{\mathbf{H}}_{1:k}$ is used for denoising, but in our paper we propose using the entire $\bar{\mathbf{H}}$ as features for classification as $\tilde{\mathbf{W}}$ and $\mathbf{W}$ are not guaranteed to be independent of each other. The activation matrix of clean data for training the classifier is also obtained using the concatenated matrix. Activation matrices of both clean and enhanced data are then used for training the classifier, similarly to what is described in the previous subsection. In our paper, the number of basis vectors of noise were also optimized as another hyperparameter.

4. Experimental study

In order to assess the generalization ability of our proposed system, the evaluation framework consists of an inner and outer cross-validation (CV) loop. Each fold of the outer CV loop has a separate inner cross-validation loop. The inner CV loop optimizes the parameters of the feature selection/transformation and classification algorithms. A single fold in the outer CV loop uses the optimized parameters to test the left-out data using the optimized classifiers. The folds were divided in leave-one-patient-out fashion to reduce dependence with the evaluation scheme. There were 8 subjects who had excessive lung water (ELW) while 12 were healthy subjects. Thus, there were a total of 20 folds. For each fold in the outer framework, the test patient is excluded from the folds used for parameter optimization in the inner framework. The proposed evaluation framework was designed to minimize bias given the small sample size. Since the inner loop is always blind to the test samples, the optimization of parameters is unbiased.

Recordings from both left lung and right lung were taken. A stethoscope sensor was placed on various positions on the back of subjects. Subjects were asked to breathe, and an analog-to-digital converter ADC was used at 8000 Hz to send the signal to a computer where it was recorded. Recordings were done in the hospital consultation room. Each recording contained 4–16 breath cycles, depending on the stamina of the subject. Only recordings with distinct audible breathing were used. Specialist physicians validated the recordings by verifying subject data and listening to the recordings. Most of the recordings ranged from forty seconds to a minute. Data collection was approved by the Institutional Review Board of Tan Tock Seng Hospital and the subjects were patients that visited the hospital for medical checkups in the cardiology department pertaining to difficulty in breathing.

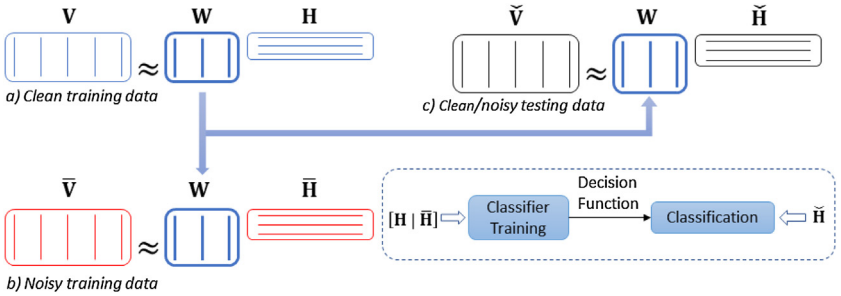

Fig. 2. Training noisy and clean data for classification of noisy test data.

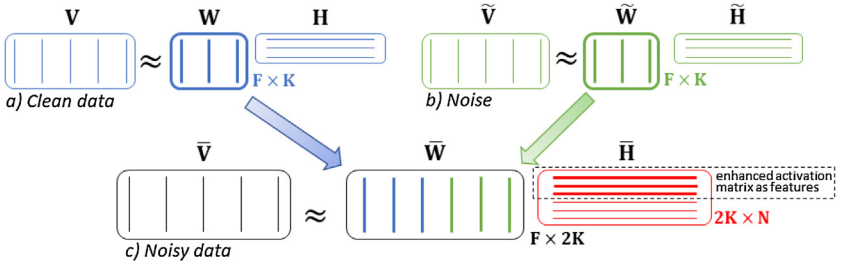

Fig. 3. Steps showing how NMF enhances samples by learning from noise samples.

Each patient sample was further segmented into 250-ms segments. This was done to maintain a high number of decisions to increase the statistical significance of the results. If longer lengths were chosen, the number of samples used for this work would been too low for any statistically significant conclusion. Because there were more healthy samples than patients with ELW, synthetic data was generated with audio software (Audacity) to produce time-altered and spectral envelope-altered variations to match the number of healthy data samples. These altered variations act as new samples from the original subjects, based on realistic audio signal modification. These synthetic data were kept in the same fold as the recording that it was generated from. This was to ensure the fold to be as independent as possible. To try and improve the statistical validity, we also attempt to generate synthetic data based on the same method as training data. Time-altered variations were produced by varying the length of the audio recording by 10%. Spectral envelope-altered variations were produced by changing the pitch by 10%. These altered variations were randomly picked so that patient data was increased to a suitable number for each patient. The resulting number of samples with ELW is 2574 (8 recordings) and the total number of samples from healthy persons is 2541 (12 recordings). These augmented samples were verified by a physician to be realistic.

Some authors considered the problem of the interference produced by heart sounds [8] but breath sounds were only audible in one of the recordings and were not significant enough to be a dominant sound.

The magnitude spectrum was used as baseline features. Short-term Fourier transform was done over 128 ms (1024-point) frames, with a hop size of 512 samples. The magnitude spectrum was averaged across all the frames for each 250 ms sample. This low-level representation was used for all features that were extracted from the magnitude spectrum. LR with the ℓ_2 regularization norm was used as the classifier. The interested reader can refer to [6] for more information.

The hyperparameters of the system were also optimized. Table 2 shows the parameters optimized as well as the range of those tested values.

Table 2
Hyperparameters optimized by the system.

Algorithm	Parameter	Range
PCA	No. of components	$2^5 - 2^8$
NMF of data and noise	No. of basis vectors	$2^5 - 2^8$
	ℓ_2 regularization	0–1000
Elastic net	Alpha parameter	$2^{-7} - 2^{11}$
	ℓ_1 ratio	0–1
Logistic regression	Reg. parameter	$2^{-7} - 2^{11}$

Table 3
List of noises used for robustness improvement, and their sources.

No. of samples	Type of noise	Tot. Dur. (s)
6	Mobile phone ring	19.0
6	Hospital machines beep with intermittent chatter	33.1
8	Background chatter	36.6
2	Television and chatter	11.3
14	Near-microphone speech	34.2

Performance was compared using $F1$-measure. Sensitivity and specificity were also calculated alongside. The $F1$-measure was calculated using $F1 = \frac{2TP}{2TP+FP+FN} \times 100$. Sensitivity (SN) was calculated using $SN = \frac{TP}{FN+TP} \times 100$ and specificity (SP) was calculated using $SP = \frac{TN}{FP+TN} \times 100$, where TP are the true positives, FP are the false positives, TN are the true negatives, and FN are the false negatives. Patients with excessive lung water were regarded as positive and individuals without excessive lung water were regarded as negative.

A mixture of background noises recorded in hospital rooms, speech signals, and other environmental noises were also added to the test recordings in order to test its robustness. Table 3 shows the list of noises used, and their sources. Speech signals were extracted from the 2008 NIST Speaker Recognition Evaluation Test Set and the rest of the noises were recorded from hospital rooms. Noise recordings were separated into half as training and testing sets. Different sets of noise samples were used for training and testing sets. Only noise recordings from the training sets were used to calculate the basis vectors of the noise and noisy training data and only noise

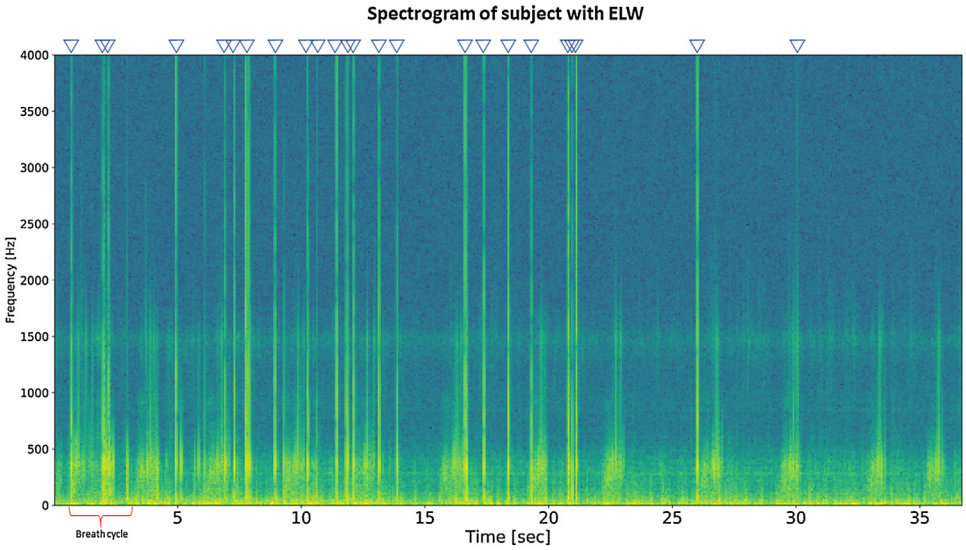

Fig. 4. Spectrogram of a sample with ELW showing a breath cycle. Triangle shows the sections with appearance of crackles.

recordings from the testing sets were used to generate noisy test data.

For each recording sample, noise recordings were picked randomly. Next, noises were arranged consecutively without pause. Noise was also recorded at 8000 Hz. In order to generalize to other conditions such as a noisy room, we added recorded background noise to the recording. Physicians validated that 6 dB SNR was reflective of a (sometimes) noisy hospital room. The chosen SNR was an estimation of the most realistic (by the physicians) among a few variants of SNR.

When testing robustness (noisy test data), noisy recordings were added to clean recordings to train the system. To minimize bias, only noisy training data that came from the same patient as clean training data was included in each iteration of the CV loop.

This study uses the leave-one-patient-out CV paired t-test to measure the level of significance when comparing classification results. Because each fold is either completely positive (ELW) or negative (healthy), the results of the statistical tests do not directly reflect comparisons between two F1-scores. Nonetheless they still do indicate if one method is better than the other. The confidence interval chosen was 90%. *p*-values were only reported when useful.

5. Typical spectrogram of samples and noise

In order to better understand the nature of the problem at hand, we plot the spectrograms of ELW and healthy samples, and noise. Fig. 4 shows the spectrogram from one subject with ELW. The recording was 34 s long with 10 breath cycles. Many crackles were present throughout the recording. These crackles manifested themselves on the spectrogram as sharp transient spikes that had a bandwidth across the whole frequency range, shown as triangles at the top of the graph. Otherwise, the dominant frequencies were mainly from 0 Hz to 50 Hz and 250 Hz to 500 Hz. One breath cycle is shown in the figure. Inhalations had relatively larger amplitudes, while exhalations were mainly quiet. In the areas with more crackles, the breath cycles were harder to visually identify. Other samples with ELW do not have as many crackles, or the crackles were softer. Some had very few spikes in the spectrogram.

Fig. 5 shows the spectrogram of a subject with healthy lungs. The recording was 52 s long, with 15.5 breath cycles. Since, it was a recording of a patient with healthy lungs, there was an absence of crackles. The dominant frequencies were mainly from 0 Hz to 500 Hz as well. However, compared to the ELW sample, there were

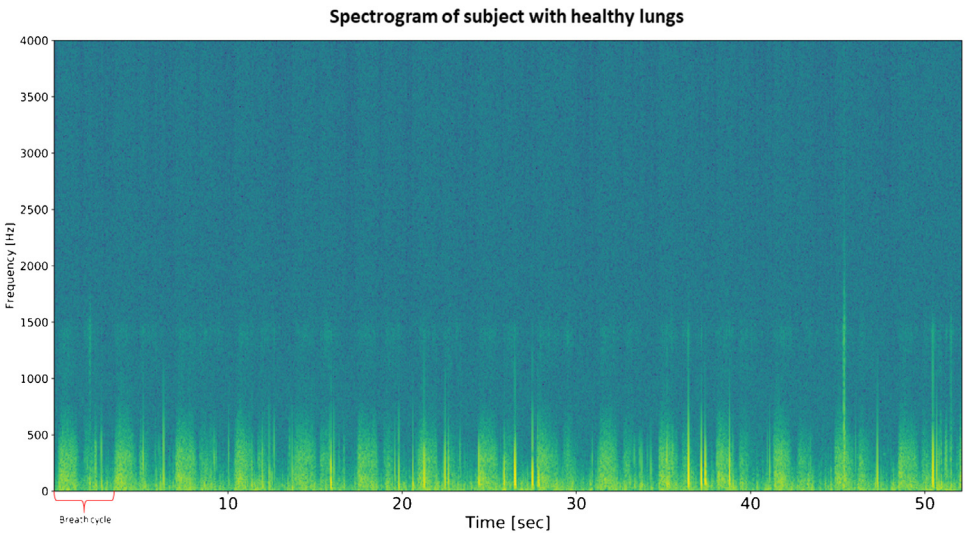

Fig. 5. Spectrogram of a sample with healthy lungs showing a breath cycle.

Table 4
Sensitivity (sn), specificity (sp) and *F*1 scores for lr on magnitude spectrum, paf, and mfcc for clean and noisy test samples.

	Feature selection	MS		PAF		MFCC		MS	PAF	MFCC
		SN	SP	SN	SP	SN	SP	F1		
Clean data	–	74	95	91	95	88	93	83	93	90
	RFE[a]	–		91	94	–		–	93	–
	EN[b]	–		89	93	–		–	91	–
Noisy Data	–	75	63	56	84	69	72	71	66	70
	RFE[a]	–		56	91	–		–	68	–
	EN[b]	–		65	80	–		–	70	–

[a] Average 57 features.
[b] Average 53 features.

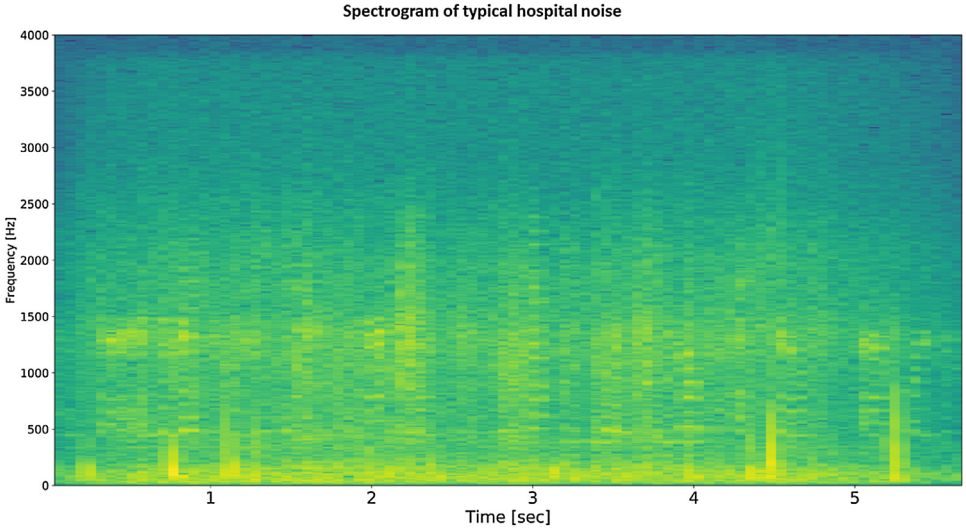

Fig. 6. Spectrogram of a noise sample containing background noise from a television set and speech.

less energy in 0–50 Hz range. The breath cycles, with one shown in the figure, were much easier to identify than the subjects with ELW. Some other samples with healthy lungs were even easier to identify, and in such cases, the recordings had virtually no noise present. Again, inhalations had relatively larger amplitudes, while exhalations were mainly quiet.

Lastly, Fig. 6 shows the spectrogram of a noise sample containing background noise from a television set and speech. The recording was 6 s. The distribution of the frequencies was more spread out, from the 0 Hz to 1500 Hz range, with slightly dominant frequencies from 0 Hz to 250 Hz and 1200 Hz to 1500 Hz. The lower dominant frequencies of this particular noise overlap the dominant frequencies of both types of lung recording, making the noisy data potentially harder to classify. Other kinds of noise samples as stated in Table 3 have also similar overlaps in dominant frequencies except for the ringing of the mobile phone.

6. Results and discussion

6.1. Initial results with clean data

First, initial tests were run to obtain *F*1-score using Logistic Regression (LR) on the magnitude spectrum (MS), PAF, and MFCC. MFCC was chosen as a benchmark because, as mentioned in Section 2, many other works use MFCC as their chosen feature. The test results are presented in Table 4. PAF had the best F1 at 93%. MFCC scored 90% and MS scored 83%. Hence, using engineered features PAF and MFCC helped to increase the F1 score from the baseline MS. In addition to that, feature selection algorithms were employed for PAF. This was because there could be a possibility of highly correlated or irrelevant features and using feature selection algorithms could eliminate or reduce those features and possibly increase classification performance. RFE produced an F1 score of 93% while EN produced a score of 91%. This was similar to when no feature selection was used. However, RFE lowered the number of features to an average of 57 features out of 100 across the outer folds, and EN lowered the number of feature to 53%. RFE produced better *F*1 scores and had roughly the same number of features as EN. SP also performed better than SN consistently. For MS, SN was 21% lower than SP. When PAF was used, the SN increased to 91% while the SP stayed the same at 95%.

Table 5
*F*1 score of nmf with varying basis and divergences, and pca.

K basis	128		256		128	256
	SN	SP	SN	SP	F1	
ℓ_2	68	67	68	68	68	68
KL	92	92	93	93	92	93
IS	95	95	94	94	95	94
PCA	87	86	87	87	87	87

6.2. Feature learning with NMF

Next, NMF was used as the feature learning method, using MS as the input representation instead of feature engineering methods like PAF. Results are presented in Table 5. The number of basis vectors *K* and the three types of divergences mentioned in

Table 6
F1, sn and sp test score for nmf with noisy and clean training data.

Test data		SN	SP	F1
Clean	KL	91	77	85
	IS	90	82	87
Noisy	KL	70	62	67
	IS	71	78	74

Table 7
F1, sn and sp test score for nmf with enhanced and clean training data.

Test data		SN	SP	F1
Clean	KL	89	92	90
	IS	93	91	92
Noisy	KL	72	71	72
	IS	76	79	77
Inter. Noisy	IS	84	65	77

Section 3.4 were investigated. ℓ_2 regularization was used for all NMF calculations. The regularization term was included as one of the parameters to be tuned. The scores reported were the average test scores from the inner folds. PCA, another popular feature learning method, was also investigated with the same number of components to compare with NMF. 32 and 64 basis vectors were investigated in our study but were systematically underperforming compared to 128 and 256 basis vectors. Thus, they are not reported in this section. Across all number of basis vectors, the ℓ_2 divergence scored the lowest while the Itakura–Saito (IS) scored the highest. Kullback–Leibler (KL) scores were in between. The ℓ_2 divergence scored the highest $F1$ score at 68% when 128 or 256 basis vectors were used, KL divergence scored the highest $F1$ score at 93% when 256 basis vectors were used, and IS scored the highest at 95% when 64 or 128 basis vectors were used. On the contrary, PCA merely scored 87% when 128 or 256 components were used.

One reason that the KL and IS divergences fared better PCA could be that KL and IS NMF gave better part-by-part representations. PCA, being a least-squares representation, placed more emphasis on larger features than smaller features. Both KL and IS divergences were not as scale-sensitive as PCA, and thus gave better representation. ℓ_2 too, most probably fared poorly because, similarly to PCA, it is a least-squared divergence, placing emphasis on larger features.

Lastly, only the IS divergence scored higher than engineered features PAF. Also, using 256 basis vectors produced the highest SN and SP for ℓ_2 and KL divergences at both 68% (SN and SP) and 93% (SN and SP) respectively. Using 128 or 256 basis vectors produced the highest SN and SP for IS divergence at 94% and 95% (SN and SP).

IS divergence scored the highest among other divergences. This is in line with previous research, as IS divergence is scale invariant and is a popular choice for audio signal processing problems [3]. KL comes close behind. Thus, features with higher amplitudes had equal emphasis as features with lower amplitudes, producing better dictionaries, which in turn produced better test metrics. Also, NMF with the IS divergence scored better than engineered features PAF or MFCC. This shows that for this audio classification problem, the NMF method is superior to conventional engineered features and feature learning PCA.

6.3. Robust classification with realistic noisy data

Robustness to noise was tested next. This is reported in Table 4. Noise samples were used to generate noisy test samples as explained in Section 4. First, algorithms described in Section 6.1 were tested. $F1$ scores for PAF had dropped by 27% to 66%, whereas MS dropped by 12% to 71% and MFCC dropped by 20% to 70%. Feature selection on PAF was also used. RFE scored 68%, while EN scored 70%. RFE managed to increase the F1 score of the PAF by 2% while EN managed to increase it by 4%. As MS had better F1 score than PAF, feature engineering and selection, specifically PAF, was unsuccessful in improving classification results for noisy test data. Using PAF also degraded the SN greatly.

Next, the NMF system that used activation matrices from both clean and noisy training data was tested. The results are presented in Table 6. NMF with both KL and IS divergence were used. 128 and 256 basis vectors were used in the inner CV to optimize the number of basis vectors and the better K was chosen for the outer CV. Firstly, for clean test data, KL scored 85% while IS scored 87% F1-measure. Secondly, for noisy test data, the F1 score for NMF for KL was 67% while IS had 74%. Taking in account the p-values, IS scored higher than KL for noisy test data ($p = 0.001$) but not for clean test data ($p = 0.58$). IS scored 3% higher than using MS as baseline features ($p = 0.0135$). Thus, the proposed system that used feature learning was more robust than MFCC or PAF, that relied on feature engineering.

Finally, the NMF system that used activation matrices from both clean and enhanced noisy training data was tested. Table 7 shows the result. Firstly, for clean test data, $F1$ scores of IS and KL did not show any statistical difference. Again, IS scored higher than KL for noisy test data ($p = 0.003$). This suggests that non-relevant signal components for classification get captured by the noise dictionary. One possible reason would be the contribution of noise recordings from the hospital environment. These matched low-level environmental noises found even in clean recordings.

Next, we will compare results between with and without NMF enhancement using IS divergence. By using NMF enhancement, the $F1$ score for clean test data increased by 5% ($p = 0.0358$). SN and SP also increased. The $F1$ score for noisy test data increased by 3% ($p = 0.0708$). Again, SN and SP both increased. Thus, we conclude that NMF signal enhancement succeeded in learning the noise dictionary to further improve the SN, SP and $F1$ scores.

Next, we shall look at the test cases where our proposed method succeeded. The two recordings with the highest accuracy score were both subjects with healthy lungs (95% accuracy for noisy test data). Upon visual inspection of their spectrograms, those two recordings had the clearest breath cycle identification, with distinct breath cycles, and obvious inhalation and exhalation sections. These implied that these recordings had less variations in breathing and helped in the classification of samples from these recordings.

The test cases with the two worst accuracies were also from subjects with healthy lungs (31% and 54% accuracy for noisy test data). Upon visual inspection of their spectrograms, these two recordings had less distinct breath cycles (although not as indistinct as ELW samples), with higher amplitudes in the exhalation sections than average. The subjects could have exhaled harder than other subjects, or incidentally, performed forced exhalations. Also, these spectrograms contained some dominant frequencies in the 0–50 Hz range, similar to subjects with ELW, as discussed in Section 5. These two observations could have been the reasons why they performed the worst in the proposed method. However, overall, the proposed method still fared much better than feature engineering techniques such as MFCC.

6.4. Robust classification with intermittently noisy data

The proposed system is designed to handle more extreme noise situations, hence should perform similarly when noise is intermittent. To test the robustness of the system under a different configuration of noise, intermittently noisy data was generated. In this case, noises were arranged similarly to what was described in Section 4, except that there were pauses with random length of distribution $N(3, 0.5)$ seconds between noise samples. The SNR was kept the same at 6 dB. The IS divergence was chosen for this test,

Table 8
*F*1 score for nmf with inhalations only with enhanced and clean training data.

Test data	SN	SP	F1
Clean	88	98	93
Noisy	77	97	86
Inter. noisy	81	96	87

and the noisy data was enhanced by NMF. The result is reported in Table 7 under 'Inter. Noisy'. Compared to the non-intermittent noisy data, the SN improved by 10% to 84%. However, the SP decreased from 80% to 62%, and the *F*1 score remained unchanged.

6.5. Further discussion

In order to gain further insight into the physiological nature of the problem at hand, breath sounds were manually divided to inhalation and exhalation, and only inhalation breath segments were considered. The MS was then extracted from these breath segments and used as the new baseline features for IS NMF. To compare with *F*1 scores from undivided breath sounds, the robustness against noise was tested as reported in Section 6.3. IS-divergence was chosen as it scored consistently better. The results are reported in Table 8 as 'Inter. Noisy'. Using inhalations alone increased the *F*1 score to 86% for enhanced data (though not statistically significant at $p = 0.237$) and 87% for enhanced intermittent noisy data ($p = 0.0955$). This demonstrates that certain segments of the breath contain features that could better differentiate healthy and lungs and lungs with ELW in noisy conditions, and in turn increase the robustness of the system. However, more studies can be done to investigate this effect. Also, more work can be done in automatically accounting for these segments, such as applying latent space classification models.

7. Conclusion

In summary, a thorough study was done on the classification of breath sound from subjects with excessive lung water against subjects with healthy lungs using the proposed system. Various engineered features such as popular audio features including MFCC were tested using logistic regression and their effectiveness was studied. The proposed system using NMF on magnitude spectrum together with NMF signal enhancement proved to be a robust system under the effect of environmental and speech noise, proving superior to other popular audio features, and feature selection techniques.

If implemented in real-time, the proposed system can be useful in a number of ways. In a hospital setting, it could aid physicians and nurses in screening patients who might have excessive lung water. In a home setting, home-care practitioners as well as subjects themselves could use the system to assess if a trip to the physician is necessary.

Acknowledgment

The authors would like to express their thanks to Mr. Yu JuFeng for his effort in designing the data collection software and conducting the data collection. They would also like to thank the staff at Tan Tock Seng hospital for allowing and assisting in collection of data. Also, they would like to thank Dr. Shahrzad for validating the audio recordings. Lastly, they would like to thank NTU scholarship for providing opportunity to conduct this research.

References

[1] V. Bisot, R. Serizel, S. Essid, G. Richard, Feature learning with matrix factorization applied to acoustic scene classification, IEEE/ACM Trans. Audio Speech Lang. Process. 25 (6) (2017) 1216–1229.
[2] L. Deng, D. Yu, et al., Deep learning: methods and applications, Found. Trends® in Signal Process. 7 (3–4) (2014) 197–387.
[3] C. Févotte, N. Bertin, J.-L. Durrieu, Nonnegative matrix factorization with the Itakura–Saito divergence: with application to music analysis, Neural Comput. 21 (3) (2009) 793–830.
[4] Z. Fu, G. Lu, K.M. Ting, D. Zhang, A survey of audio-based music classification and annotation, IEEE Trans. Multimed. 13 (2) (2011) 303–319.
[5] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (1) (2002) 389–422.
[6] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning. Springer Series in Statistics, Springer New York Inc., New York, NY, USA, 2001.
[7] K.J. Hong, W. Ser, Z. Lin, D.C.-G. Foo, Acoustic detection of excessive lung water using sub-band features, in: Circuits and Systems Conference (DCAS), IEEE Dallas. IEEE, 2015, pp. 1–4.
[8] S. İçer, S. Gengeç, Classification and analysis of non-stationary characteristics of crackle and rhonchus lung adventitious sounds, Digital Signal Process. 28 (2014) 18–27.
[9] I. Jolliffe, Principal Component Analysis, Wiley Online Library, 2002.
[10] Y.E. Kim, E.M. Schmidt, R. Migneco, B.G. Morton, P. Richardson, J. Scott, J.A. Speck, D. Turnbull, Music emotion recognition: a state of the art review, Proc. ISMIR. Citeseer (2010) 255–266.
[11] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[12] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (6755) (1999) 788.
[13] H. Lee, P. Pham, Y. Largman, A.Y. Ng, Unsupervised feature learning for audio classification using convolutional deep belief networks, Advances in Neural Information Processing Systems (2009) 1096–1104.
[14] C. Lin, E. Hasting, Blind source separation of heart and lung sounds based on nonnegative matrix factorization, in: International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS), IEEE, 2013, pp. 731–736.
[15] H. Martinez-Hernandez, C. Aljama-Corrales, R. Gonzalez-Camarena, V. Charleston-Villalobos, G. Chi-Lem, Computerized classification of normal and abnormal lung sounds by multivariate linear autoregressive model, in: 27th Annual International Conference of the Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005, IEEE, 2006, pp. 5999–6002.
[16] B. Mathieu, S. Essid, T. Fillon, J. Prado, G. Richard, Yaafe, an easy to use and efficient audio feature extraction software, ISMIR (2010) 441–446.
[17] S. Matsunaga, K. Yamauchi, M. Yamashita, S. Miyahara, Classification between normal and abnormal respiratory sounds based on maximum likelihood approach, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009, IEEE, 2009, pp. 517–520.
[18] S. Matsutake, M. Yamashita, S. Matsunaga, Abnormal-respiration detection by considering correlation of observation of adventitious sounds, Signal Processing Conference (EUSIPCO), 2015 23rd European. IEEE (2015) 634–638.
[19] P. Mayorga, C. Druzgalski, R. Morelos, O. Gonzalez, J. Vidales, Acoustics based assessment of respiratory diseases using gmm classification, in: Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2010, pp. 6312–6316.
[20] N. Mohammadiha, P. Smaragdis, A. Leijon, Supervised and unsupervised speech enhancement using nonnegative matrix factorization, IEEE Trans. Audio Speech Lang. Process. 21 (10) (2013) 2140–2151.
[21] A. Mondal, P. Bhattacharya, G. Saha, Detection of lungs status using morphological complexities of respiratory sounds, Sci. World J. (2014).
[22] R. Naves, B.H. Barbosa, D.D. Ferreira, Classification of lung sounds using higher-order statistics: a divide-and-conquer approach, Comput. Methods Prog. Biomed. 129 (2016) 12–20.
[23] R.X.A. Pramono, S. Bowyer, E. Rodriguez-Villegas, Automatic adventitious respiratory sound analysis: a systematic review, PLOS ONE 12 (5) (2017), e0177926.
[24] R.X.A. Pramono, S.A. Imtiaz, E. Rodriguez-Villegas, A cough-based algorithm for automatic diagnosis of pertussis, PLOS ONE 11 (9) (2016), e0162128.
[25] N. Sengupta, M. Sahidullah, G. Saha, Lung sound classification using cepstral-based statistical features, Comput. Biol. Med. 75 (2016) 118–129.
[26] D. Seung, L. Lee, Algorithms for non-negative matrix factorization, Adv. Neural Inf. Process. Syst. 13 (2001) 556–562.
[27] M.A. Tocchetto, A.S. Bazanella, L. Guimaraes, J. Fragoso, A. Parraga, An embedded classifier of lung sounds based on the wavelet packet transform and ann, IFAC Proc. 47 (3) (2014) 2975–2980.
[28] S. Ulukaya, G. Serbes, I. Sen, Y.P. Kahya, A lung sound classification system based on the rational dilation wavelet transform, in: IEEE 38th Annual International Conference of the Engineering in Medicine and Biology Society (EMBC), IEEE, 2016, pp. 3745–3748.

[29] S. Umeki, M. Yamashita, S. Matsunaga, Classification between normal and abnormal lung sounds using unsupervised subject-adaptation, in: Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific, IEEE, 2015, pp. 213–216.
[30] K.W. Wilson, B. Raj, P. Smaragdis, A. Divakaran, Speech denoising using nonnegative matrix factorization with priors, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008, IEEE, 2008, pp. 4029–4032.
[31] F. Yang, W. Ser, J. Yu, D.C.-G. Foo, D.P.S. Yeo, P.-L. Chia, J. Wong, Lung water detection using acoustic techniques, in: Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2012, pp. 4258–4261.
[32] M. Yeginer, Y. Kahya, Modeling of pulmonary crackles using wavelet networks, in: 27th Annual International Conference of the Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005, IEEE, 2006, pp. 7560–7563.
[33] K. Zhang, X. Wang, F. Han, H. Zhao, The detection of crackles based on mathematical morphology in spectrogram analysis, Technol. Health Care 23 (s2) (2015) S489–S494.