An audio classification system for detecting pulmonary edema

Hong, Kah Jun

2019

Hong, K. J. (2019). Thermal failure of carbon nanostructures. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/90282>

<https://doi.org/10.32657/10220/48518>

AN AUDIO CLASSIFICATION SYSTEM FOR DETECTING
PULMONARY EDEMA

HONG KAH JUN

School of Electrical & Electronic Engineering
Nanyang Technological University

A thesis submitted to the Nanyang Technological University in partial fulfillment of the requirement
for the degree of Doctor of Philosophy

October 2018

Hong Kah Jun: *An Audio Classification System for Detecting Pulmonary Edema*, October 2018

SUPERVISORS:

Assoc. Prof. Ser Wee

Prof. Slim Essid

STATEMENT OF ORIGINALITY

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

06/03/2018

Date

Hong Kah Jun

SUPERVISOR DECLARATION STATEMENT

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

06/03/2018

Date

Dr. Ser Wee

AUTHORSHIP ATTRIBUTION STATEMENT

This thesis contains material from 3 papers published in the following peer-reviewed journal(s) / from papers accepted at conferences in which I am listed as an author.

Chapter 4 is published as Kah Jun Hong et al. "Acoustic detection of excessive lung water using sub-band features." In: *Circuits and Systems Conference (DCAS), 2015 IEEE Dallas*. IEEE. 2015, pp. 1–4

and

Kah Jun Hong, Wee Ser, and David Chee-Guan Foo. "An analysis of sounds for lungs with excessive water." In: *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*. IEEE. 2016, pp. 3662–3665.

The contributions of the co-authors for "Acoustic detection of excessive lung water using sub-band features" are as follows:

- Dr. David Foo provided the initial project direction.
- Yu JuFeng designed the data collection software and conducted the data collection.
- I processed the data, designed the experimental framework, and performed all coding and calculations.
- I interpreted the results and provided the discussion.
- I wrote the draft of the manuscript.
- A/Prof Ser Wee and A/Prof ZhiPing Lin provided guidance in the interpretation of results.
- A/Prof Ser Wee and A/Prof ZhiPing Lin edited and revised the manuscript drafts.

The contributions of the co-authors for "An analysis of sounds for lungs with excessive water" are as follows:

- I provided the initial project direction.
- Yu JuFeng designed the data collection software and conducted the data collection.
- I processed the data, designed the experimental framework, and performed all coding and calculations.

- I interpreted the results and provided the discussion.
- I wrote the draft of the manuscript.
- A/Prof Ser Wee provided guidance in the interpretation of results.
- A/Prof Ser Wee edited and revised the manuscript drafts.

Chapter 5 is published as K. J. Hong et al. "A robust audio classification system for detecting pulmonary edema." In: *Biomedical Signal Processing and Control* 46 (Sept. 2018), pp. 94–103.

The contributions of the co-authors for "A robust audio classification system for detecting pulmonary edema" are as follows:

- Prof Slim Essid provided the initial project direction.
- Yu JuFeng designed the data collection software and conducted the data collection.
- I processed the data, designed the experimental framework, and performed all coding and calculations.
- I interpreted the results and provided the discussion.
- I wrote the draft of the manuscript.
- Prof Slim Essid and A/Prof Ser Wee provided guidance in the interpretation of results.
- Prof Slim Essid and A/Prof Ser Wee edited and revised the manuscript drafts.

06/03/2018

Date

Hong Kah Jun

*Every good gift and every perfect gift is from above,
coming down from the Father of lights,
with whom there is no variation or shadow due to change.*

— James 1:17

ACKNOWLEDGMENTS

Firstly, I would like to express my sincere gratitude to my supervisor Prof. Ser Wee, and my co-supervisor Prof. Slim Essid. Their continuous support for my PhD. studies helped me overcome this challenge. Their patience and kind encouragement motivated me when I was discouraged, and their insightful comments helped me widen my perspectives in approaching my research problems.

I would also like to express my thanks to Mr. Yu JuFeng for his effort in designing the data collection software and conducting the data collection, the staff at Tan Tock Seng hospital for allowing and assisting in collection of data, and Dr. Shahrzad for validating the audio recordings.

My appreciation also goes out to my ex-colleagues at Nanyang Technological University and Télécom ParisTech who have helped in one way or another, and taught me many things, especially Tom Dupré la Tour.

I would like to thank my friends and family who have supported me through this journey, and given me the strength to press on.

Lastly, I would like to give thanks to the Lord, who is the giver of all good things.

CONTENTS

1	INTRODUCTION	1
1.1	Background and Motivation	1
1.2	Scope and Contributions	1
1.3	Organization of Thesis	3
2	LITERATURE REVIEW	5
2.1	Conventional Approaches to Detecting Pulmonary Edema	5
2.2	Lung Sound Models and Studies	5
2.3	Machine Learning Approaches to Symptom and Pathology Detection	7
2.4	Machine Learning Approaches to Feature Learning in Audio Classification	8
2.5	Conclusion	9
3	EXISTING TECHNIQUES USED FOR BENCHMARK TESTS	11
3.1	Existing Audio Feature Engineering Techniques	11
3.2	Feature Selection Techniques	13
3.3	Principal Component Analysis	14
3.4	K-means Clustering	15
3.5	Support Vector Machine as Classifier	15
3.6	Conclusion	15
4	DATA PROCESSING, EVALUATION FRAMEWORK, AND STATISTICAL TESTING	17
4.1	Data Processing	17
4.2	Evaluation Framework	18
4.3	Conclusion	21
5	FEATURE ENGINEERING WITH IMPROVEMENT USING EMD	23
5.1	Typical Spectrogram of Samples	23
5.2	Feature Engineering Techniques	25
5.3	Empirical Mode Decomposition	26
5.4	Breath Phases	29
5.5	Robustness Testing	29
5.6	Experimental Study	31
5.7	Results and Discussion	32
5.8	Conclusion	35
6	ROBUST ELW DETECTION USING NMF	37
6.1	Introduction	37
6.2	The Methodology for Proposed System	38
6.3	Non-Negative Matrix Factorization	39
6.4	Experimental Study	42

6.5	Results and Discussion	43
6.6	Conclusion	47
7	LATENT SSVM AND ROBUST CASCADING DICTIONARY LEARNING	49
7.1	Experimental Study	49
7.2	Experiments with Latent SSVM	50
7.3	Experiments with Cascading Dictionary Learning . . .	52
7.4	Conclusion	63
8	CONCLUSION	67
8.1	Thesis Summary	67
8.2	Further Work	68
9	PUBLICATIONS	71
A	CALCULATION DETAILS FOR FEATURE SELECTION	73
	BIBLIOGRAPHY	77

LIST OF FIGURES

Figure 2.1	Left, A: Ultrasound image of a control patient with no ULC. Right, B: Ultrasound image of ULC in a patient with high altitude pulmonary edema.	6
Figure 4.1	Stethoscope sensor used for recording data . .	17
Figure 4.2	Evaluation framework consisting of inner CV for optimization and outer CV for testing. . .	19
Figure 4.3	Histogram plot for class distribution across folds	20
Figure 5.1	Spectrogram of a sample with ELW showing a breath cycle. Triangles show the sections with appearance of crackles.	24
Figure 5.2	Spectrogram of a sample with healthy lungs showing a breath cycle.	24
Figure 5.3	Box plot of top 20 features from PAF. "" and "" indicates 1 st and 2 nd derivatives	27
Figure 5.4	Example of IMFs with the original signal in red. There are 32 IMFs including the residue. . . .	30
Figure 5.5	Spectrogram of a noise sample containing background noise from a television set and speech	32
Figure 6.1	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. .	38
Figure 6.2	Training noisy and clean data for classification of noisy test data	41
Figure 6.3	Steps showing how NMF enhanced samples by learning from noise samples	42
Figure 7.1	Breath segmentation of a sample	53
Figure 7.2	Latent variables of a clean sample	53
Figure 7.3	Latent variables of a noisy sample	54
Figure 7.4	Steps showing how to obtain cascaded dictionaries.	56
Figure 7.5	Training error across increasing $\mathbf{W}_c$	59
Figure 7.6	Errors and sample count across each sift. The green partitions represent the sifts just before a new subset	64

LIST OF TABLES

Table 5.1	Features extracted by Yaafe.	25
Table 5.2	List of noises used for robustness improvement and their sources.	31
Table 5.3	SN, SP and F1 scores for SVM on MS, PAF, and MFCC.	33
Table 5.4	SN, SP and F1 scores for SVM on FEE.	34
Table 5.5	SN, SP and F1 scores for SVM on inhalation and exhalation.	35
Table 5.6	SN, SP and F1 scores for SVM on MS, PAF, & MFCC for noisy test samples.	36
Table 6.1	Hyperparameters optimized by the system . .	43
Table 6.2	List of noises used for robustness improvement and their sources.	43
Table 6.3	F1 score of NMF with varying basis & divergences, and PCA.	44
Table 6.4	F1, SN and SP for NMF with noisy and clean training data.	46
Table 6.5	F1, SN and SP for NMF with enhanced and clean training data.	46
Table 6.6	F1, SN and SP for NMF with enhanced and clean training data on k-SVM & GBC	48
Table 7.1	SN, SP and F1 for LSSVM on clean and noisy training data	51
Table 7.2	SN, SP and F1 for SVM on clean and noisy training data	51
Table 7.3	Hyperparameters optimized by the system . .	57
Table 7.4	Hyperparameters optimized for TDNMF. . . .	58
Table 7.5	F1, SN and SP for NMF with k-means clustering	60
Table 7.6	F1, SN and SP for NMF with cascaded dict. (left) and best results from Chapter 6 (right). .	61

ACRONYMS

DAS Discontinuous Adventitious Sounds
ULC Ultrasound Lung Comets
k-NN k-Nearest Neighbor

LPC Linear Predictive Coefficients
LSF Line Spectral Frequency
MFCC Mel-frequency Cepstral Coefficients
RFE Recursive Feature Elimination
EN Elastic Net
SVM Support Vector Machine
PCA Principal Component Analysis
IS Itakura-Saito
KL Kullback-Leibler
ELW Excessive Lung Water
GBC Gradient Boosting Classifier
CV Cross-Validation
SN Sensitivity
SP Specificity
PAF Popular Audio Features
EMD Empirical Mode Decomposition
IMF Intrinsic Mode Functions
MS Magnitude Spectrum
FEE Feature Engineering with EMD improvement
NMF Non-negative Matrix Factorization
EN Elastic Net
MU Multiplicative Update
SNR Signal-to-Noise Ratio
CDL Cascading Dictionary Learning
TDNMF Task Driven Non-negative Matrix Factorization
k-MC k-Means Clustering
LSSVM Latent Structural SVM

ABSTRACT

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.

This thesis proposes new insights and methods to detect pulmonary edema using auscultation recordings. The first contribution makes use of feature engineering to gain new insights into the characteristics of these audio recordings. Empirical mode decomposition is used to improve the classification performance.

In the first contribution, it is shown that conventional feature selection algorithms are not designed to handle external interfering hospital noises. In the second contribution, improvements are made on the robustness with the introduction of a new system. The second contribution uses non-negative matrix factorization (NMF) to develop a robust audio classification system to detect pulmonary edema. A study was done to compare feature engineering approaches with feature selection techniques against NMF. Different NMF schemes were investigated and compared with Principal Component Analysis. Background noise collected from hospitals and speech from a speech corpus database were used to simulate noisy data. The system was then tested using noisy data. NMF was also used as a signal enhancement tool, which improved the classification scores.

Contributions in the third chapter are divided into two main parts. Inspired by the analysis of spectrograms of breath recordings, the first section investigates the use of Latent SSVM to show that breath sounds contain latent information that could potentially be exploited. To further improve the proposed NMF system, the second part of the contribution uses NMF to develop a cascading dictionary learning algorithm that allows the use of sample subsets to increase classification accuracy. The SVM classifier is used to group subsets of data

and calculate dictionaries on each subset of data. These subsets can be thought of as latent variables that have been grouped into these subsets. The method is described and we compare results with benchmarks, including results from the previous chapter.

In conclusion, this thesis focuses on the automatic classification of pulmonary edema with the use of breath recordings through the chest wall. In particular, feature engineering was used to gain new insights. NMF was next used to develop a robust classification system. Lastly, Latent SSVM was used to show that breath sounds contain latent information that could be exploited. A cascading dictionary learning algorithm was proposed for that purpose and improved classification results in noisy conditions.

INTRODUCTION

In this chapter, the background and motivations of this thesis are first described in [Section 1.1](#). In [Section 1.2](#), the scope and contributions of thesis are described. Finally, the organization of the thesis is presented in [Section 1.3](#).

1.1 BACKGROUND AND MOTIVATION

Pulmonary Edema occurs when the alveolar beds get engorged by water in the lungs. The excessive lung water is pushed into the airspaces and thus reduces normal oxygen movement in the lungs. It can cause coughing up of blood or bloody froth, difficulty in breathing, shortness of breath, and grunting, gurgling or wheezing sounds with breathing. It may also lead to anxiety or restlessness, decrease in level of alertness, pale skin and leg, or abdominal swelling [4]. Pulmonary Edema have different causes such as congestive heart failure, kidney failure, high altitude exposure, lung damage, and major injuries .

Currently, physicians will carry out a physical examination before performing X-ray, CT-scans and electrocardiograms. The physician uses auscultations to check for abnormal heart sounds, crackles, increased heart rate, and rapid breathing. Crackles are caused by explosive opening of the alveoli during both inspiration and expiration phase [5]. Patients with congestive heart failure and other permanent injuries and diseases have to face excessive lung water throughout their lives. They have to see the physicians whenever they show symptoms of excessive lung water, which if left untreated can lead to suffocation. These constant trips to the physician can add stress, worry and inconvenience. Physicians also have to handle a greater workload, especially when dealing with patients who are unsure if their lungs have accumulated too much water.

Thus, the motivation for the our thesis is the need to aid patients and physicians in screening for excessive lung water. This thesis presents methods developed during the course of the author's PhD project involving classification of excessive lung water that is robust against realistic hospital noise.

1.2 SCOPE AND CONTRIBUTIONS

This thesis focuses on the automatic classification of pulmonary edema with the use of breath recordings through the chest wall. We focus on offline feature extraction and classification. We also concentrate on

developing algorithms that are robust to realistic hospital noise. The contributions made can be summarized into three parts:

1. In the first part, we study the effectiveness of using known feature engineering in classifying pulmonary edema signals. Our preliminary studies on feature engineering techniques can be found in [1]. We introduce the use of empirical mode decomposition and find that using 17 intrinsic mode functions together with various other engineered features improves the performance on the detection of pulmonary edema. Preliminary studies using empirical mode decomposition can be found in [2]. We also perform experiments on the segregation of breath phases– inhalation and exhalation. We show that segregating the breath recordings in such a way does not improve classification performance, thereby concluding that the entire breath cycle is preferred in detecting pulmonary edema. We also study the effect of adding realistic hospital noise to the data, thereby testing the robustness of the popular engineered features as well as the improvements made by empirical mode decomposition. We show that the classification performance (F1 measure) drops by at least 23%. Details can be found in [Chapter 5](#).
2. For the second part, we introduce the use of non-negative matrix factorization for feature learning and present a study on the use of various divergences for estimating the factorization. We discovered that the itakura-saito divergence is the most helpful for our task. Using the itakura-saito divergence, we find that our proposed system improves F1 measure by 11% for the detection of pulmonary edema in the presence of hospital noise. This work is summarized in a journal publication [3]. The details are presented in [Chapter 6](#).
3. In the final part, we perform two main tasks. Firstly, we investigate the use of Latent SSVM to show that breath sounds contain latent information that could potentially be exploited. Secondly, we propose a new cascading dictionary learning algorithm that takes advantage of latent information. The algorithm learns a series of dictionaries based on subsets of training data. We propose a method to best choose the number of dictionaries used to classify both clean and noisy test data. We show that although there is a disimprovement in the classification of clean test data by 2% F1 measure, there is an improvement in the classification of noisy test data by 2% F1 measure. We are currently consolidating the findings to submit to a conference for publication. Details of this work are reported in [Chapter 7](#).

If such systems are adopted and implemented, they can be useful in a number of ways. In a hospital setting, they 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.

1.3 ORGANIZATION OF THESIS

The rest of this thesis is organized as follows:

[Chapter 2](#) presents an extensive literature review on current approaches to detecting pulmonary edema, some studies on lung sound models, and machine learning approaches to other symptom and pathology detection.

[Chapter 3](#) gives a general overview of all the existing techniques used in the thesis. They include feature engineering, selection, transformation and classification techniques.

[Chapter 4](#) describes the data collection and processing procedure, the evaluation framework to assess the generalization ability of our proposed systems, and methods of statistical testing.

[Chapter 5](#) presents a study on the effect and improvement of various feature engineering techniques in the detection of pulmonary edema as described in the previous section.

[Chapter 6](#) presents a robust audio classification system to efficiently detect pulmonary edema as described in the previous section.

[Chapter 7](#) present a new method of feature transformation that learns dictionaries based on subsets of data called cascading dictionary learning as described in the previous section.

Lastly, [Chapter 8](#) gives a summary of the work done in this thesis as well as possible future work.

LITERATURE REVIEW

In this chapter, we outline existing literature on pulmonary edema in [Section 2.1](#), studies on lung sound models in [Section 2.2](#), and machine learning approaches to related diseases and lung sounds in [Section 2.3](#). We illustrate that there is a need to develop algorithms that can give insight on pulmonary edema using real data, and to develop systems to detect pulmonary edema.

2.1 CONVENTIONAL APPROACHES TO DETECTING PULMONARY EDEMA

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 Discontinuous Adventitious Sounds ([DAS](#)) called crackles. Physicians also analyze X-rays and CT-scans. X-rays that contain blacked-out sections of the lung indicate water deposits [\[4\]](#). However, water accumulation would have been already severe at that point in time. Recently, there have been multiple studies on ultrasound as a way of determining the quantity of lung water due to pulmonary edema [\[6–9\]](#). The appearance of Ultrasound Lung Comets ([ULC](#)) is indicative of an increase in lung water level. [Figure 2.1](#) shows the comparison the ultrasound image of a normal lung and a lung with excessive lung water. The right figure shows a lung with 2 comets. Studies [\[9\]](#) and [\[8\]](#) both concluded that having no less than 5 comets was the indicator of diagnosis of pulmonary edema; and as the lung water level receded in the lungs, so did the number of comets. Auscultations and ultrasound detection both suggest that the acoustic properties of the lung can be utilized to extract features for detecting pulmonary edema.

2.2 LUNG SOUND MODELS AND STUDIES

Lung acoustic models were first developed in 1989 in [\[10\]](#) by Wodicka *et al.*. They theorized a model of sound transmission from within the respiratory tract to the chest walls. A model of the respiratory tract was first proposed. It was made of interconnected airway segments each represented by a T-network. The authors, using the model, calculated the attenuation of sound in the parenchyma, as a function of composite density and volumetric compliance of gas, and tissue found in the

Figure 2.1: Left, A: Ultrasound image of a control patient with no ULC. Right, B: Ultrasound image of ULC in a patient with high altitude pulmonary edema.

parenchyma. The resulting conclusion was that the attenuation was frequency dependent, and also dependent on the water-to-air ratio, h . Wodicka *et al.* simulated the attenuation of pressure wave propagation through 10cm of parenchyma model. They also simulated plural effusion by varying h from 0.25 (normal lung) to 0.35 (effused lung). An effused lung resulted in a lower attenuation which attenuated less as the frequencies increased. Through theoretical modeling, it was shown that excessive water in lungs changed the acoustic transfer function of the modeled lung, and so the magnitude spectrum of sample signals could be used as features for detecting pulmonary edema.

In a more engineering-oriented approach, Mulligan *et al.* developed an instrument for measuring changes in the distribution of lung fluid in the respiratory system [11] in 2009. A chest phantom model was constructed. Using an finite impulse response adaptive filter, a transfer function of that chest phantom model was determined. Water was injected into that model, and the transfer function was calculated at certain water levels. Frequency spikes were observed at 117.2 Hz and 242.2 Hz, and the spikes became sharper as more water entered the model. Zaeim *et al.* [12] improved the chest phantom model and plural effusion was modeled using two latex plastic bags. One was filled with 600 ml of water while the other contained 250 ml of water. Measurements were taken at eight angular positions and three axial positions. The healthy lung and effused lung were both analyzed. Results showed that for both cases, attenuation at frequencies above 500 Hz were high. For the healthy lung, the highest amplitude frequency response varied with the axial measurement location. For the effused lung, highest amplitude frequency response was similar for each axial measurement location. The authors concluded that the locations of the fluid filled bags were clearly detectable. Thus, this work showed that lungs filled with water were distinguishable from healthy lungs at least at the locations that were filled with water. However, there have

not been many studies on the acoustic characteristics of pulmonary edema from recordings of subjects with the disease. More research can be done to obtain insights on recordings of breath sounds on pulmonary edema.

2.3 MACHINE LEARNING APPROACHES TO SYMPTOM AND PATHOLOGY DETECTION

As mentioned above, 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.* [13] 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. All studies cited use different kinds of feature engineering techniques to obtain features. For example, the standard choice for speech and audio classification tasks, Mel-frequency Cepstral Coefficients (MFCC) [14, 15], was used by [16–20]. All proposed feature engineering methods as part of their solutions had success of various degrees and they concluded that the 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 [21–26] 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.* [27] 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. [28] 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 [29]. Different features and classifiers were explored, and MFCC together with k-Nearest Neighbor (k-NN) produced the best result. 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.

2.4 MACHINE LEARNING APPROACHES TO FEATURE LEARNING IN AUDIO CLASSIFICATION

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. Deep learning certainly is the main representative of this trend today. Deep learning [30] has been a popular choice for many classification and regression problems. It also has been used in audio-related problems [31]. Specifically, deep learning has been applied to speech recognition [32], audio sensing [33], speech denoising [34], and speech/music generation with WaveNet [35] with good results. However, deep learning methods require huge amounts of data, and the cited research all use large-scale datasets with large number of tuned parameters. Thus for applications with small datasets, signal-specific feature engineering is usually preferred, or else the model will overfit [36].

The idea of transfer learning [37, 38] has been popular in some areas of deep learning, mainly computer vision [39] and in natural language processing [40, 41], where the model is trained on a large set of data in one domain and fine-tuned on a smaller target dataset. Following this trend, the audio research community have also trained neural networks for many tasks such as audio event detection [42, 43], spoofing detection [44], music information retrieval [45], audio sensing [33], robust speech recognition [46] and even bird species identification [47]. There also have been at least one attempt at transfer learning [48]. The large-scale pre-trained model WaveNet was trained on speech and music data. They are significantly different from breathing sounds in terms of frequency bands and statistical distributions (spectrograms will be discussed in Chapter 5). We will not be able to fine-tune these models and build on their representations to suit our own models. Data augmentation is also limited as long as the signal of interest are limited in quantity.

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 will most likely not be implemented successfully. We are thus motivated to adopt and focus on developing other methods of feature learning for audio in small datasets. Still, research such as [49] has shown that another popular method called non-negative matrix factorization has been successful in audio source separation and enhancement. [50–52] had success similar to deep learning methods in the audio classification domain where there is a lack of training data. Also, Non-negative Matrix Factorization (NMF) was used in [53] for blind source separation of heart and lung sounds.

2.5 CONCLUSION

Conventionally, physicians have used auscultations to detect crackles in the process of diagnosing pulmonary edema. There have also been studies on ultrasound images, showing ‘comets’ that denote the severity of pulmonary edema. We also presented several studies on effused lung models, showing that effused lungs have different acoustic properties from normal lungs. From an engineering standpoint, there have been many methods developed in detecting DAS. However, DAS occurrences are temporally located, and previously proposed event-based classification would perform poorly in segments without DAS. On top of that, crackles appear in other various diseases, and these methods cannot make such distinctions. In the next chapter, we describe studies we have done analyzing breathing sounds from subjects with pulmonary edema and comparing them with subjected with healthy lungs. We also propose an empirical mode decomposition-based algorithm that can help provide insight into these breath sounds.

EXISTING TECHNIQUES USED FOR BENCHMARK TESTS

This chapter investigates the usefulness of various audio feature engineering techniques. These features are first described in [Section 3.1](#). We also describe feature selection techniques found in [Section 3.2](#), and transformation techniques in [Section 3.3](#). In [Section 3.4](#) we introduce a clustering technique, and in [Section 3.5](#) we describe the classifier adopted such that our proposed methods could be fairly compared with existing methods.

3.1 EXISTING AUDIO FEATURE ENGINEERING TECHNIQUES

As mentioned in [Chapter 2](#), there have been feature engineering techniques developed to classify [DAS](#). To qualify the performance of our proposed systems, we compare existing feature engineering techniques on audio recordings of Excessive Lung Water ([ELW](#)) and healthy lungs. In this section, we describe these techniques and give reasons why we have chosen them. Calculation details for these features are described [Appendix A](#).

3.1.1 *Linear Predictive Coefficients*

Linear Predictive Coefficients ([LPC](#)) [[54](#)] has been an effective tool for speech coding applications. It is used due to its all-pole characteristics of vowel spectra [[55](#)], which in turn is better because the human ear is more sensitive to spectral poles than zeros. [LPC](#) is also a short-term estimate and thus is suitable as a feature for the classification of pulmonary edema. It has been used for other various classification problems [[56](#), [57](#)].

3.1.2 *Line Spectral Frequency Coefficients*

Line Spectral Frequency ([LSF](#)) [[58](#)] was first developed for the robust representation of the [LPC](#). This is useful for our application because we will be testing the robustness of the features in the presence of noise. [LSF](#) also produces greater interpolation properties and thus can be effective in the presence of temporal features which are found in pulmonary edema. They also have been used for other various classification problems [[59](#), [60](#)]. We also use the 1st and 2nd derivatives of [LSF](#) as features.

3.1.3 *Mel-Frequency Cepstral Coefficients*

The MFCC [61] was originally used in speech recognition to extract features, but now are the standard choice for speech and audio classification task[16–20]. In general, MFCC were designed to represent English phonemes being produced by the human vocal tract. We also use the 1st and 2nd derivatives of MFCC as features.

3.1.4 *Loudness Coefficients*

Loudness coefficients are the energy in each Bark band, normalized by the overall sum [62]. They are energy measures in alternative spectrum bands and can be used as features that correspond to human’s perception of loudness, for example in [63].

3.1.5 *Sharpness of Loudness*

The sharpness is the 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. Because of the presence of crackles and other temporal artifacts in pulmonary edema, sharpness can be useful as a feature for our classification task. They have been used for other audio classification tasks [64–67].

3.1.6 *Spectral Decrease*

The spectral decrease is a weighted average of the decrease of the magnitude spectrum between consecutive frames. It can be useful for the classification of pulmonary edema because the magnitude spectrum of ELW samples tends to contain temporal artifacts that change in the magnitude quickly. It has been used in audio classification tasks such as [68][69].

3.1.7 *Spectral Flatness*

The spectral flatness is the ratio between the geometric mean and the arithmetic mean of the spectrum. It was designed to give a quantitative measure of the ‘whiteness’ of a spectrum [70], and is robust to common signal distortions [71]. It can be useful in our classification task because ELW samples tend have a wider spectrum and thus could be more “white”. It has also been used in other audio classification tasks such as [72–74].

3.1.8 *Spectral Flux*

Spectral flux is a measure of how quickly the power spectrum of a signal is changing, calculated by comparing the power spectrum for one frame, t , against the power spectrum from the previous frame [75]. Similarly to spectral decrease, it can be useful for the classification of pulmonary edema because temporal artifacts in ELW samples change their power spectrum more rapidly than healthy lung samples. It has also been used in other audio tasks such as [76, 77].

3.1.9 *Spectral Slope*

The spectral slope is a linear regression of the magnitude spectrum. This gives an overall linear trend of the magnitude spectrum, and is able to describe a linear relationship among the amplitudes of the magnitude spectrum. This feature could be useful in classifying pulmonary edema if there are different types of emphasis on the bands on the magnitude spectrum. There are other audio classification tasks that use spectral slope as a feature [68, 78].

3.1.10 *Spectral Moments*

The spectral moments have been a popular method in retrieving information directly from the magnitude spectrum. They provide a description of the overall shape, rather than the fine structure of the magnitude spectrum [79]. Thus they can be used as features for our classification tasks. In our cases, we use up to the fourth spectral moments— the centroid, spread, skewness, and kurtosis of the magnitude spectrum. They have also been used in other audio classification tasks [68, 80].

3.1.11 *Zero Crossing Rate*

The zero crossing rate is also a popular feature to measure the weighted average of the spectral distribution of the waveform [81]. The dominant frequency principle [82] shows that when a certain frequency band carries more power than other bands, it tends to indicate the normalized expected zero crossing per unit time. It is also computationally cheap and as a result used frequently as a feature for audio classification [69, 83].

3.2 FEATURE SELECTION TECHNIQUES

One way of utilizing these features is to present of all them as the input of a chosen classifier. However, such features could be noisy,

or features could be correlated with each other, and implementing feature selection techniques can increase classification accuracy [84–86]. We use two feature selection techniques, the Recursive Feature Elimination (RFE) [87] and elastic net with hinge loss [88].

3.2.1 Recursive Feature Elimination

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. For our work, Support Vector Machine (SVM) was used as the classifier in the RFE procedure. After ranking, the best number of features were selected using k-fold cross validation.

3.2.2 Elastic Net

The second method was the elastic net used with SVM, optimizing the hinge loss. It is an embedded feature selection and classification method. The estimate of the SVM weights β is:

$$\hat{\beta} = \arg \min_{\beta} \left[L_s(y, X\beta) + \alpha \left(\lambda \|\beta\|^2 + (1 - \lambda) \|\beta\| \right) \right]$$

where L_s is the hinge loss, α is the regularization strength and λ is the ratio between ℓ_1 and ℓ_2 regularization. The interested reader can refer to [88] for more information.

3.3 PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) was used as a benchmark to evaluate the effectiveness of our proposed transformation methods in Chapter 6 and Chapter 7. PCA [89] 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 PCA 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.

To calculate the transformation matrix used in PCA, the eigenvectors and eigenvalues of the data X are first calculated. Next, the first p number of eigenvectors are chosen that correspond to the eigenvalues that represent the p^{th} percentile of the sum of eigenvalues. Lastly, the transformation matrix U is generated by taking these eigenvectors as its columns. Then, the transformed features is obtained using $Y = UX$.

3.4 K-MEANS CLUSTERING

k-Means Clustering (**k-MC**) was used as a benchmark to evaluate the effectiveness of our improved transformation method in [Chapter 7](#). **k-MC** is an unsupervised learning algorithm that attempts to solve the clustering problem. k number of centroids are fixed a priori and defined, one for each cluster. The centroids are calculated with the aim to minimize an objective function $J = \sum_{j=1}^k \sum_{i=1}^n \|x_i^j - c_j\|^2$ where $\|x_i^j - c_j\|^2$ is a chosen distance measure between data point x_i^j and the cluster centre c_j . The objective function is minimized with an iterative loop that shifts the initialized centroid towards the minimum of the objective function. More information can be found in [\[90\]](#). We use the k-means++ for initial seeding [\[91\]](#).

3.5 SUPPORT VECTOR MACHINE AS CLASSIFIER

SVM[\[88\]](#) is used in all our proposed methods as a classifier. Many works use **SVM** as the chosen classification method as it is reliable and effective [\[68, 72, 92–97\]](#). **SVMs** make a decision by first drawing a separating margin between classes such that the distance from it to the nearest data point on each class is maximized. It is calculated by minimizing the optimization function:

$$\arg \min_w \left[\frac{1}{n} \sum_{i=1}^n \max(0, 1 - y_i (w \cdot x_i - b)) + \lambda \|w\|^2 \right],$$

where w is the separating hyperplane coefficient vector, b is the bias, y_i is the class in the form of either -1 or $+1$, and x_i is its corresponding data point. λ affects the trade-off between increasing the margin-size and correct classification of x_i . Minimization of the optimization function is solved typically by gradient descent or quadratic programming.

3.6 CONCLUSION

In this chapter, we have described existing feature engineering and feature selection techniques used commonly in audio classification tasks. These techniques are used in our proposed systems as benchmarks to measure the system's performance. We have also described **SVM**, the chosen classifier in all our proposed systems. In the next chapter, we describe the data collection and processing procedure, evaluation framework for classification, and the statistical significance testing method.

DATA PROCESSING, EVALUATION FRAMEWORK, AND STATISTICAL TESTING

In this chapter, we describe the data collection and processing procedure in [Section 4.1](#). We also present the assessment framework to assess the generalization ability of our proposed systems, and method of statistical testing in [Section 4.2](#).

4.1 DATA PROCESSING

The works in this thesis use recordings of breathing sounds as data. There were 8 subjects who had [ELW](#) while 12 were healthy subjects. Recordings from both left lung and right lung were taken. A stethoscope sensor was placed on various positions on the backs of subjects and the subjects were asked to breathe. Next, those breath sounds were recorded using a stethoscope sensor (Addest Technovation) as shown in [Figure 4.1](#). An analog-to-digital converter was used at 8000Hz 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. The subjects were patients who visited the hospital for medical checkups in the cardiology department pertaining to difficulty in breathing.

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 have been too

Figure 4.1: Stethoscope sensor used for recording data

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 attempted 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 was 2574 (8 recordings) and the total number of samples from healthy persons was 2541 (12 recordings). These augmented samples were verified by a physician to be realistic.

The magnitude spectrum was used as baseline features. Short-term Fourier transform was done over 128ms (1024-point) frames, with a hop size of 512 samples. The magnitude spectrum was averaged across all the frames for each 250ms sample. This low-level representation was used for all features that were extracted from the magnitude spectrum.

4.2 EVALUATION FRAMEWORK

In order to assess the generalization ability of our 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 a leave-one-patient-out fashion to reduce dependence with the evaluation scheme. There were 8 subjects who had ELW while 12 were healthy subjects. Thus, there were a total of 20 folds. For each fold in the outer framework, the test patient was 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 as unbiased as possible. Figure 4.2 shows the evaluation framework. Figure 4.3 shows the class distribution of samples per fold for a total of 2574 samples for subjects with ELW and 2541 samples for healthy subjects

Figure 4.2: Evaluation framework consisting of inner CV for optimization and outer CV for testing.

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.

4.2.1 Statistical Testing

A modified approximate randomization test is used throughout to measure the statistical significance when comparing classification results from one system with another. The modified approximate randomization test is based on the approximate random test [98]. The approximate randomization test is a randomization test in which the significance of an obtained statistic is determined by approximating the distribution of the test statistic. We use the two-tailed paired-test variant. The method is as follows:

Figure 4.3: Histogram plot for class distribution across folds

1. Given the outputs of two systems A and B to be in the form of $(o_A^1, o_B^1), \dots, (o_A^F, o_B^F)$ for F number of folds.
2. For R times:
 - a) Randomly exchange o_A^f and o_B^f for all F with probability 0.5. We call the new output of these systems A' and B' .
 - b) Calculate the test statistic of A' and B' . If one is testing mean difference of outputs across folds, then the test statistic will be: $t(A', B') = \frac{1}{F} \sum^F |o_{A'}^f - o_{B'}^f|$.
3. For a large value of R, $\frac{r+1}{R+1}$ approaches the significance level, where r is the number of times that $t(A', B') \geq t(A, B)$. In our experiments, $R = 200000$ is large enough for the reduction of variance due to randomness to insignificant levels.

Thus, if the test statistic of both systems are different, then the observed $t(A', B') \geq t(A, B)$ should be unlikely under this distribution.

Because of the nature of the CV framework, test data for each fold is entirely made up of ELW or healthy data. Thus, each fold would only contain true positives and false negatives, or true negatives and false positives. In this way, the F1-measure for each fold cannot be calculated and the mean difference in F1-measure across folds cannot be used as a test statistic. In order to overcome that limitation, we choose the test statistic to be the absolute difference of the F1-score calculated from the sum of all folds: $t(A', B') = |F1(A') - F1(B')|$.

4.3 CONCLUSION

In this chapter, we have presented the method in which we collect and process the data. Next, we described the framework we used to assess our proposed systems and lastly, the statistical testing method to determine the significance of our proposed methods.

FEATURE ENGINEERING WITH IMPROVEMENT USING EMD

In this chapter we present a study on the effect and improvement of various feature engineering techniques in the detection of pulmonary edema. [Section 5.1](#) compares spectrograms of samples from the different classes. In our study, we employ various feature engineering techniques that have been previously used in many other problems. We also employ feature selection techniques and analyze their effect on the dataset. These are briefly explained in [Section 5.2](#). Next in [Section 5.3](#), we use empirical mode decomposition to analyze the recordings into their intrinsic mode functions. We also investigate the segregation of breath recordings into inhalation and exhalations and compare results with full breath recordings in [Section 5.4](#). In [Section 5.5](#), we investigate the robustness of the system, describe the framework to generate noisy data, and show a spectrogram of a typical noise recording. In [Section 5.6](#), we give a brief description of the experimental study. We show that by using some intrinsic mode functions, we can further improve the classification score using some of the engineered features. The results and discussions are finally presented in [Section 5.7](#).

5.1 TYPICAL SPECTROGRAM OF SAMPLES

In order to better understand the nature of the problem at hand, we plot the spectrograms of ELW and healthy samples, and noise. [Figure 5.1](#) shows the spectrogram from one subject with [ELW](#). The recording was 34-second 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 from 0Hz to 2000Hz frequency range, and they can be easily seen on the graph. Otherwise, the dominant frequencies were mainly from 0Hz to 50Hz and 250Hz to 500Hz. One breath cycle is shown on 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.

[Figure 5.2](#) shows the spectrogram of a subject with healthy lungs. The recording was 52 seconds 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 0Hz to 500Hz

Figure 5.1: Spectrogram of a sample with ELW showing a breath cycle. Triangles show the sections with appearance of crackles.

Figure 5.2: Spectrogram of a sample with healthy lungs showing a breath cycle.

as well. However, compared to the [ELW](#) sample, there was less energy in the 0Hz to 50Hz range.

The breath cycles for healthy recordings, 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 artifacts present. Again, inhalations had relatively larger amplitudes, while exhalations were mainly quiet.

Table 5.1: Features extracted by Yaafe.

NAME OF FEATURE	FEATURES
LPC	2
LSF	10
1st and 2nd derivatives of LSF	20
MFCC	13
1st and 2nd derivatives of MFCC	26
Loudness coefficients	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

5.2 FEATURE ENGINEERING TECHNIQUES

5.2.1 Feature Extraction, Selection and Classification

First, we consider various popular audio features to compare and rank. Table 5.1 shows the list of these features. These features will be referred to as Popular Audio Features (PAF). All features given in Table 5.1 were extracted by Yaafe [99]. They have been described in detail in Chapter 3. A total of 109 features are generated as a result. MFCC alone was also used as features for this system to compare with our previous work.

Two popular standard feature selection methods were used to compare the effectiveness of the proposed system. The two methods are RFE and Elastic Net (EN) used with hinge loss. Details of these methods were explained in Chapter 3. This was used 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 and EN were chosen because they work well when the data size is small. They are less prone to overfitting, and have relatively few hyperparameters to tune. [84] provides a comprehensive overview of feature selection methods and objectives to substantiate our decision. We use SVM as our classifier. We chose SVM because it is a tried-and-tested classifier which is also known to generalize well even when trained on small datasets [100].

5.2.2 Box Plot Visualizations for PAF

In total, there are 109 features extracted for PAF. In order to visualize the potentially more useful features clearly, we show box plots for the top twenty features that have the greatest difference between the medians of each class. We also standardized the features for more direct comparison, such that they had zero mean and one standard deviation. Figure 5.3 shows these box plots. The top twenty features are shown as the x-axis labels. Overall, they show good differentiation between the two classes, especially LSF' 8 and MFCC'' 8. Most of these features are made up of LPC, LSF, and MFCC with its first and second derivatives. Half of the twenty features are MFCC based, and is in line with literature showing that it is a standard feature for audio classification tasks. Seven features are LSF coefficients and its first derivatives. As mentioned in Chapter 3, LSFs are known to be robust to noise and quantization.

5.3 EMPIRICAL MODE DECOMPOSITION

Empirical Mode Decomposition (EMD) empirically identifies the intrinsic oscillatory modes of the signal by particular time scales [101]. These modes are called Intrinsic Mode Functions (IMF) and they characterize the data through zero or near zero local mean. Due to the sifting process, EMD is by nature orthogonal, adaptive, and local. In the original method proposed by Huang *et. al.*, each IMF would satisfy two conditions. The first condition states that the number of extrema and zero crossings can only differ by at most one. The second condition states that the local mean of the IMF must be zero. The locality of the mean then can be defined by the stationarity of the signal. The signal can be represented by the sum of its IMF:

$$s(t) = \sum_n IMF_n(t) + r_N(t) \quad (5.1)$$

where N is the total number of IMFs, and $r_N(t)$ is the residue signal which is constant or monotonic.

Figure 5.3: Box plot of top 20 features from PAF. “” and “” indicates 1st and 2nd derivatives

The sifting process is as follows:

1. Generate an upper envelope of the given signal $s(t)$ by a cubic spline using the local maxima.
2. Generate a lower envelope of the given signal $s(t)$ by a cubic spline using the local minima.
3. Find the average of the envelopes to attain the mean envelope $m(t)$.
4. Compute the difference $h(t) = s(t) - m(t)$. This is regarded as one sift.
5. Check if $h(t)$ meets IMF conditions, then
 - a) if $h(t)$ is an IMF, calculate the remainder $r(t) = s(t) - h(t)$
 - b) if $h(t)$ is not a IMF, replace $s(t)$ with $h(t)$, and repeat step 1 to step 5.

Once the first IMF is generated, the second IMF can be generated by using the same sifting procedure on the first residue $r_1(t) = s(t) - \text{IMF}_1(t)$. When the residue $r_N(t)$ is either monotonic or constant, the algorithm stops.

There are also different measures of goodness of the decomposition that either control the number of sifts to get an IMF, or the number of IMFs generated [101]. In our study, we use four such parameters. The first is the standard deviation test. If the standard deviation falls below a certain threshold, the sifting ends. This is calculated using

$$\text{SD} = \sum_t \left[\frac{h_i(t) - h_{i-1}(t)}{h_{i-1}(t)} \right]^2 \quad (5.2)$$

where $h_i(t)$ is the i^{th} IMF. The second is a scaled variance test. Likewise, if the scaled variance falls below a certain threshold, the sifting ends. This is calculated using

$$\text{SV} = \frac{\sum_t [h_i(t) - h_{i-1}(t)]^2}{\max[h_{i-1}(t)] - \min[h_{i-1}(t)]} \quad (5.3)$$

The third is a total power test. This is calculated by $\text{TP} = \sum_t |r_n(t)|$. The fourth is an absolute amplitude test. This is calculated by $\text{AA} = \max[r_n(t)] - \min[r_n(t)]$. If the total power or absolute amplitude after each IMF fall below a certain threshold, the EMD algorithm stops.

There are many ways to use the IMFs for feature analysis, such as the Hilbert Huang Transform [102], and as a filter bank [103]. In our preliminary experiments, we discovered that extracting features from each IMF and combining them together as one feature set performed better than various other methods. Thus, for example, if we choose

13 MFCCs as features from each sample, then a sample with 10 IMFs would give 130 features in total. Figure 5.4 shows an example of EMD on a ELW signal and a healthy (HL) signal. For this signal, there are 32 IMFs including the residue signal. One can see that the frequency components are separated into different IMFs, and as the IMFs progress, the dominant frequency of the local oscillations become lower. For example, IMF 1 to 8 contain mainly high frequency components that arise from transient artifacts. As mentioned previously, ELW samples tend to contain more transient artifacts because of the presence of crackles. We can see that there are more IMFs that contain high frequency components for the ELW recording. Also, the bottom 5 IMFs for both ELW and healthy recordings look similar. Thus, we potentially need only to consider some of the IMFs since most likely, not all the frequency components will be useful in classifying pulmonary edema, especially lower frequency components. We can do that by choosing only features extracted from the first M number of IMFs, where M is a parameter that is to be optimized.

5.4 BREATH PHASES

As mentioned in the literature review, the most common way to detect pulmonary edema is through the identification of crackles. These crackles are found in the late inspiration and early inspiration phases. Thus, we investigate the effect of only either inhalations or exhalation breath segments in order to see if it can increase classification scores. We segregated the inhalation and exhalation segments. We performed data augmentation similarly to what we have described in Chapter 4 to equalize the number of samples per class. The segmentation was validated by a physician.

5.5 ROBUSTNESS TESTING

In order to generalize our system to other conditions such as a noisy room, we added recorded background noise to the recordings. A mixture of background noises were recorded in hospital rooms, speech signals, and other environmental noises were also added to the test recordings in order to test its robustness. Table 5.2 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 recordings from the testing sets were used to generate noisy test data.

Figure 5.4: Example of IMFs with the original signal in red. There are 32 IMFs including the residue.

Table 5.2: 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

For each recording sample, noise recordings were picked randomly. Next, noises were arranged consecutively without pause. Noise was also recorded at 8000Hz. Physicians validated that 6dB Signal-to-Noise Ratio (SNR) was reflective of a 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 fold.

Figure 5.5 shows the spectrogram of a noise sample containing background noise from a television set and speech. The recording was 6-seconds long. 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 overlapped the dominant frequencies of both types of lung recordings, making the noisy data potentially harder to classify. Other kinds of noise samples as stated in Table 5.2 have also similar overlaps in dominant frequencies except for the ringing of the mobile phone.

5.6 EXPERIMENTAL STUDY

In order to assess the ability of our proposed system fairly with as little bias as possible, we adopted the evaluation framework described in Chapter 4. The data used was collected and processed as described in Chapter 4. The resulting number of samples with ELW is 2574 (8 recordings) and the total number of samples from healthy persons is 2541 (12 recordings).

Figure 5.5: Spectrogram of a noise sample containing background noise from a television set and speech

Some authors considered the problem of the interference produced by heart sounds [104] 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 128ms (1024-point) frames, with a hop size of 512 samples. The magnitude spectrum was averaged across all the frames for each 250ms sample. This low-level representation was used for all features that were extracted from the magnitude spectrum. Linear soft-margin SVM was used as the classifier. Details can be found in Chapter 3. We also investigated MFCC features and PAF features.

The hyperparameters of the system were also optimized. The regularization parameter of SVM was optimized within the range of $2^{-7} - 2^{11}$. Performance was compared using F1-measure, SN and SP. Chapter 3 shows how they were calculated. This study also uses the non-parametric approximated randomization test to measure the level of significance of the F1-score when comparing classification results, with 200000 trials as described in Chapter 4. The minimum p-value was 0.000005.

5.7 RESULTS AND DISCUSSION

5.7.1 Initial results with MS, MFCC and PAF

First, we ran tests to obtain F1-score using SVM on the Magnitude Spectrum (MS), PAF, and MFCC. MFCC was chosen as a benchmark because, as mentioned in Chapter 3, many other works use MFCC as their chosen feature. The test results are presented in Table 5.3. PAF

Table 5.3: SN, SP and F1 scores for SVM on MS, PAF, and MFCC.

FEATURE SELECTION	MS		PAF (109)		MFCC (13)	
	SP	SP	SN	SP	SN	SP
-	90	93	90	95	88	93
RFE [*]	89	93	88	94	89	93
EN [#]	87	93	89	94	89	94

FEATURE SELECTION	MS	PAF	MFCC
	F1		
-	92	92	90
RFE [*]	91	91	91
EN [#]	90	91	91

^{*} average 273 features for MS, 37 for PAF, 11 for MFCC

[#] average 384 features for MS, 66 for PAF, 12 for MFCC

and MS had the best F1 at 92% while MFCC scored slightly lower at 90%. Hence, using engineered features PAF and MFCC did not help increase the F1 score from the baseline MS. The other features found in PAF other than MFCC also seemed to help increase the F1 score by 2%. In addition to using a classifier, feature selection algorithms were employed. This was no better than when no feature selection was used, indicating that the features selected manually were relevant for our task. However, RFE lowered the number of selected features to an average of 273 for MS, 37 for PAF and 11 for MFCC across the outer folds. EN lowered the number of selected features to an average of 384 for MS, 66 for PAF and 12 for MFCC across the outer folds. Thus, RFE selected fewer features than EN. Also, SP scored higher than SN; it was harder to classify a sample with ELW correctly than a healthy sample.

5.7.2 Improvements with EMD

MFCC from each IMF were extracted and used as features. MFCC was chosen in our experiments because of the comparable performance it had with PAF and MS despite its small dimensionality (for clean environments). Through preliminary studies, we discovered that including PAF features from non-decomposed samples improved classification performance. Also, using only the first 17 IMFs gave the best inner-CV

Table 5.4: SN, SP and F1 scores for SVM on FEE.

FEATURE SELECTION	FEE (331)		PAF		FEE	PAF
	SN	SP	SN	SP	F1	
-	91	96	90	95	94	92
RFE [*]	94	97	88	94	95	91
EN [#]	88	93	89	94	91	91

^{*} average 103 for FEE

[#] average 220 for FEE

accuracies. Thus, there were 331 features. We will reference these features as Feature Engineering with EMD improvement (FEE). Table 5.4 shows the test results. The F1 measure for FEE without feature selection techniques did not change as compared with PAF without any feature selection ($p = 0.215$). However, after feature selection by RFE, FEE increased F1 measure score by 3% ($p = 0.0664$). One reason why feature selection with RFE helped could be because some IMFs within the 17 IMFs were correlated with one another. Another reason would be that since EMD is adaptive, the same IMF number across different samples may have oscillations of different dominant frequencies. So RFE helped to remove or at least reduce those mismatched features and this improved F1 measure results. EN however, did not improve classification performance. It seems that overall, RFE was a better feature selection method than EN.

5.7.3 Breath Phases

We also ran similar tests on inhalation phases and exhalation phases separately. The results are reported in Table 5.5. We only tested with MS as features because these results already show that segmenting the breaths to either inhalations or exhalations did not help improve classification results. In fact they scored poorer than when full breath cycle recordings were used. For example, exhalation scored better than inhalations, but even then the F1 measure was 4% lower than results reported in Section 5.7.1 for MS. Thus, there exist features in both inhalation and exhalation phases that can increase performance on the detection of pulmonary edema. However, future work can include more complex algorithms to extract latent information found in both breath phases, such as latent dynamic systems.

Table 5.5: SN, SP and F1 scores for SVM on inhalation and exhalation.

FEATURE SELECTION	INHAL.		EXHAL.		INHAL.	EXHAL.
	SN	SP	SN	SP	F1	
-	88	65	93	76	79	86
RFE*	87	62	93	76	77	86

* average 352 features for inhalation, 175 for exhalation

5.7.4 Robust Testing

Robustness to noise was tested next. We combined the clean training data with noisy training data so that the classifier could learn both clean and noisy training data at the same time. Classification performance was calculated and reported in Table 5.6. Noise samples were used to generate noisy test samples as explained in Section 5.5. F1 scores for PAF without feature selection techniques had dropped by 23%, to 69%, whereas MS dropped by 31%, to 61% and MFCC dropped by 26%, to 64%. FEE dropped by 24%, to 70%. Feature selection was also used. Overall there was no change to the F1 score and they were both unhelpful for all feature sets to improve performance under noisy test data. However, there were reductions in RFE and EN number of features used. We can thus conclude that although feature engineering improved the classification performance of clean data, it was ineffective with the addition of realistic noise.

5.8 CONCLUSION

In this chapter we presented spectrograms of both ELW and healthy lung sample recordings. We also studied the potential of using feature engineering for classification and showed that feature engineering did only to reduce the number of features, but not increase classification performance. Feature selection also did not increase the classification performance but also further reduced the number of features. We also introduced the use of EMD to increase classification performance by engineering features from its IMF, as well as using various other engineered features. We managed to increase the F1 measure by 3%. We investigated the segregation of breath phases and showed that using the entire breath cycle is more effective than using either inhalations or exhalations separately. Lastly, various engineered features such as popular audio features including MFCC and FEE were tested using SVM, and their effectiveness under noise was studied. We showed that feature engineering cannot perform well in the presence of realistic

Table 5.6: SN, SP and F1 scores for SVM on MS, PAF, & MFCC for noisy test samples.

FEATURE SELECTION	MS		PAF		MFCC		FEE	
	SN	SP	SN	SP	SN	SP	SN	SP
-	65	51	66	74	67	59	69	73
RFE ^{**}	66	55	65	75	65	63	70	74
EN ^{##}	55	81	68	61	57	31	65	65

FEATURE SELECTION	MS	PAF	MFCC	FEE
	F1			
-	61	69	64	70
RFE ^{**}	63	68	64	71
EN ^{##}	63	66	51	65

^{**} average 411 features for MS, 72 for PAF, 11 for MFCC, 103 for FEE

^{##} average 202 features for MS, 75 for PAF, 13 for MFCC, 220 for FEE

noise. In the following chapter, we will propose a system that improves robustness under realistic data.

In this chapter, we present a robust audio classification system to efficiently detect pulmonary edema. This chapter follows the work of the previous chapter. Previously, we showed that conventional feature selection algorithms are not designed to handle external interfering hospital noises. In this chapter, we seek to improve the robustness of our proposed systems with the introduction of a new system. The system uses a feature learning technique based on NMF, then classified with SVM. In Section 6.1, we first give a background of NMF for classification tasks. In Section 6.2, we describe the methodology of our system. In Section 6.3, we give a comprehensive explanation of non-negative matrix factorization. NMF was also used as a signal enhancement tool. A study was done to compare feature engineering approaches with feature selection techniques against NMF. Different NMF schemes were investigated and also compared with PCA. The experimental study is found in Section 6.4. The results and discussions are found in Section 6.5.

6.1 INTRODUCTION

As discussed in Chapter 2, previous works mainly focus on classifying DAS. In the case of pulmonary edema, the physician listen to crackles. The few works that involve classification of diseases use autoregressive models [27], LPC and MFCC [29], as features for classification. Specifically, MFCC has been the standard choice for many speech and audio classification tasks [14, 15]. However, although all proposed feature engineering methods had various degrees of success, they are also problem specific. We have explained it thoroughly in Section 2.4. Thus, in this chapter we adopt NMF as a feature learning method.

In this chapter, we also show that feature engineering techniques are insufficient in extracting useful features for classification under hospital environmental noise. Instead, 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.

Figure 6.1: 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.

6.2 THE METHODOLOGY FOR PROPOSED SYSTEM

Figure 6.1 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 SN, SN and F1-measure. More details are in Chapter 4. Data was segmented into 250ms segments and normalized. Features of the data were then extracted. The system was tested with and without feature selection and feature transformation techniques. SVM 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. We will give a brief description of the techniques used in the feature extraction, transformation and classification steps as mentioned in Chapter 3.

First, we consider various popular audio features to compare with our NMF-based approach. We use the same PAF audio features found in Chapter 3. A total of 109 features are generated as a result.

13 MFCCs 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 8000Hz sampling rate).

Two popular standard feature selection methods were used to compare the effectiveness of the proposed system. The two methods are [RFE](#) and [EN](#) used with hinge loss. Details of these methods were explained in [Chapter 3](#).

[PCA](#) was used in this study as a benchmark to evaluate the effective of our [NMF](#)-based method. [PCA](#) is a feature transformation technique that maximizes the variance of the transformed features in their new feature space. A detailed explanation can be found in [Chapter 3](#).

In order to assess the ability of our proposed system fairly with as little bias as possible, we adopted the evaluation framework described in [Chapter 4](#).

6.3 NON-NEGATIVE MATRIX FACTORIZATION

[NMF](#) [50] decomposes non-negative data observations assembled as the N columns of the matrix $\mathbf{V}$ using non-negative dictionary elements and their expansion coefficients. Here, the matrix $\mathbf{V}$ is the magnitude spectrum of the audio signals. 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. Columns of $\mathbf{W}$, $\mathbf{w}_k$; $1 \leq k \leq K$, 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} ; $1 \leq k \leq K$, $1 \leq n \leq N$, contains the weight of $\mathbf{w}_k$ for the n^{th} 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

$$\mathbf{v}_n \approx \sum_{k=1}^K h_{kn} \mathbf{w}_k \text{ s.t. } \geq 0.$$

[NMF](#) is calculated by minimizing an objective function:

$$\min_{\mathbf{W}, \mathbf{H}} D_{\beta}(\mathbf{V} | \mathbf{WH}) \text{ s.t. } \mathbf{W}, \mathbf{H} \geq 0, \quad (6.1)$$

where D_{β} is a separable divergence in the beta-family of divergences and $D(\mathbf{V} | \hat{\mathbf{V}}) = \sum_{f=1}^F \sum_{n=1}^N d(v_{fn} | \hat{v}_{fn})$, where $\hat{\mathbf{V}} = \mathbf{WH}$ and v_{fn} is the f^{th} feature for sample n . In this chapter, 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:

$$d_{\ell_2}(v_{fn}, \hat{v}_{fn}) := (v_{fn} - \hat{v}_{fn})^2. \quad (6.2)$$

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

$$d_{\text{KL}}(v_{fn}, \hat{v}_{fn}) := v_{fn} \log \frac{v_{fn}}{\hat{v}_{fn}} - v_{fn} + \hat{v}_{fn}. \quad (6.3)$$

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

$$d_{\text{IS}}(v_{fn}, \hat{v}_{fn}) := \frac{v_{fn}}{\hat{v}_{fn}} - \log \frac{v_{fn}}{\hat{v}_{fn}} - 1. \quad (6.4)$$

The Multiplicative Update ([MU](#)) [[50](#)] approach is used to minimize the objective function. An iteration of the [MU](#) is

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^T \left((\mathbf{WH})^{[\beta-2]} \odot \mathbf{V} \right)}{\mathbf{W}^T (\mathbf{WH})^{[\beta-1]}} \quad (6.5)$$

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\left((\mathbf{WH})^{[\beta-2]} \odot \mathbf{V} \right) \mathbf{H}^T}{(\mathbf{WH})^{[\beta-1]} \mathbf{H}^T} \quad (6.6)$$

and the matrices $\mathbf{W}$ and $\mathbf{H}$ are usually initialized either randomly or using singular value decomposition of the data. In our study, 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 [[50](#)]. In this paper, the [MUs](#) cease 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 can be calculated with a given activation matrix. An ℓ_2 -norm penalty can also be added as a regularization term as done in ridge regression. The objective function then becomes

$$\min_{\mathbf{W}, \mathbf{H}} \left[D_{\beta}(\mathbf{V} \mid \mathbf{WH}) + \phi \|\mathbf{H}\|_2^2 + \phi \|\mathbf{W}\|_2^2 \right] \text{ s.t. } \mathbf{W}, \mathbf{H} \succeq 0$$

where ϕ is the regularization strength. After applying [NMF](#), the activation matrix is used as a set of 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 $\check{\mathbf{H}}$ of the test data is computed by fixing the dictionary of the test data $\check{\mathbf{W}} = \mathbf{W}$ for every iteration in the [MU](#) approach, and only using [Equation 6.6](#).

Figure 6.2: Training noisy and clean data for classification of noisy test data

6.3.1 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. Figure 6.2 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.

6.3.2 NMF as a Signal Enhancement Tool

NMF can also be used for signal enhancement in the presence of noise [105]. Figure 6.3 shows the steps involved for enhancing samples. First, the noise dictionary $\tilde{W}$ is learned using training noise $\tilde{V}$ as described in Section 6.3. W is also learned using clean training data samples V , where $\tilde{W}$ and W are dictionaries for noise and clean data samples respectively. A new dictionary, $\bar{W}$ representing the noisy data, is created by concatenating W and $\tilde{W}$ such that $\bar{W} = [W | \tilde{W}]$. Next, activation matrix $\bar{H}$ for enhanced data is calculated using $\tilde{V} \approx \bar{W}\bar{H}$, fixing $\bar{W}$ during multiplicative updates. In [105] only $\bar{H}_{1:k}$ is used for denoising, but in our paper we propose using the entire $\bar{H}$ as features for classification as $\tilde{W}$ and 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 study, the number of basis vectors of noise were also optimized as another hyperparameter.

Figure 6.3: Steps showing how NMF enhanced samples by learning from noise samples

6.4 EXPERIMENTAL STUDY

In order to assess the generalization ability of our proposed system, we adopted the evaluation framework described in Chapter 4. The data use was collected and processed as described in Chapter 4. The resulting number of samples with ELW is 2574 (8 recordings) and the total number of samples from healthy persons is 2541 (12 recordings). In order to generalize to other conditions such as a noisy room, we added recorded background noise to the recordings. A mixture of background noises were recorded in hospital rooms, speech signals, and other environmental noises were also added to the test recordings in order to test its robustness. Table 6.2 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 two: half as the training set and half as the 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 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 8000Hz. Physicians validated that 6dB SNR was reflective of a 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 fold.

The hyperparameters of the system were also optimized. Table 6.1 shows the parameters optimized as well as the range of those tested values.

Performance was compared using F1-measure, SN and SP. Chapter 3 shows how they were calculated. This study also uses the non-parametric approximated randomization test to measure the level of

Table 6.1: Hyperparameters optimized by the system

ALGORITHM	PARAMETER	RANGE	NO. OF VALUES
PCA	No. of components	$2^5 - 2^8$	4
NMF of data and noise	No. of basis vectors	$2^5 - 2^8$	4
	ℓ_2 -regularization	0 – 1000	11

Table 6.2: 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

significance of the F1-score when comparing classification results, with 200000 trials as described in [Chapter 4](#). The minimum p-value was 0.000005.

6.5 RESULTS AND DISCUSSION

6.5.1 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 6.3](#). The number of basis vectors K and the three types of divergences mentioned in [Section 6.3](#) were investigated. ℓ_2 -regularization was used for all [NMF](#) calculations. The regularization term was included as one of the parameters to be tuned. [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 under-performing compared to 128 and 256 basis vectors. Thus, they are not reported in this section. Across all number of basis vectors, the

Table 6.3: F1 score of NMF with varying basis & divergences, and PCA.

k BASIS	128		256		128	256
	SP	SP	SN	SP	F1	
ℓ_2	46	74	49	69	53	54
KL	92	92	90	95	92	93
IS	96	97	96	96	96	94
PCA	90	93	89	93	92	91

ℓ_2 divergence scored the lowest while the **IS** scored the highest. **KL** and **PCA** scores were in between. ℓ_2 divergence scored poorly for both 128 and 256 basis vectors at 53% and 54. **KL** divergence scored similarly for both number of basis vectors, with 92% when 128 basis vectors were used and 93% for 256 basis vectors. **IS** scored the highest at 96% when 128 basis vectors were used and 94% when 256 basis vectors were used. **PCA** performed similarly to **KL** divergence, scoring 92% when 128 basis vectors were used and 91% when 256 basis vectors were used.

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 [106]. One reason that the **IS** divergences fared better than **KL** and **PCA** could be that **IS** divergence gave better part-by-part representation. **PCA**, being a least-squares representation, placed more emphasis on larger features than smaller features. **KL**, although not as scale sensitive, still gave similar representation to **PCA**. **IS** divergence is scale invariant and thus gave better representation. Hence, features with higher amplitudes had equal emphasis as features with lower amplitudes, producing better dictionaries, which in turn produced better test metrics. ℓ_2 too, most probably fared poorly because it is a least-squared divergence, placing quadratic emphasis on larger features. Lastly, **NMF** with the **IS** divergence scored similarly with **FEE** ($p = 0.689$). Previous results with engineered features are re-shown as Table 6.4 for ease of comparison.

6.5.2 Robust Classification with Realistic Noisy Data

Next, Robustness was tested. The **NMF** system that used activation matrices from both clean and noisy training data was first studied. The results are presented in Table 6.4. **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 79% while IS scored 92% F1-measure. Secondly, for noisy test data, the F1 score for NMF for KL was 55% while IS had 75%. Taking in account the p-values, IS scored higher than KL for clean test data ($p = 0.000005$) and noisy test data ($p = 0.000005$). When we compared results of FEE and RFE with IS NMF with noisy testing data, F1 score for our proposed system was higher than FEE by 6% ($p = 0.0171$), and when we compared our proposed system with PCA, our system scored 11% higher ($p = 0.000954$).

Finally, the NMF system that used activation matrices from both clean and enhanced noisy training data was tested. Table 6.5 shows the result. Firstly, for clean test data, KL scored 90% while IS scored 94% F1-measure. Secondly, for noisy test data, the F1 score for NMF for KL was 73% while IS had 81%. IS scored higher than KL for clean test data ($p = 0.0996$) and for noisy test data ($p = 0.00127$). 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. When we compared results of FEE with IS NMF with enhanced training data, F1 score for our proposed system was higher than FEE by 13% ($p = 0.000520$). When we compared our proposed system with PCA, our system scored 17% higher ($p = 0.000560$). Thus, by using NMF we could improve the robustness of the system while feature selection techniques and PCA could not.

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 did not increase ($p = 0.482$). The F1 score for noisy test data improved by 6%. ($p = 0.0537$). Thus, we conclude that NMF signal enhancement succeeded in learning the noise dictionary to further improve the SN, SP and F1 scores for noisy test data, and thus better handled classification overall when test data is noisy or clean.

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. This implied that these recordings had less variations in breathing and helped in the classification of samples.

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

Table 6.4: F1, SN and SP for NMF with noisy and clean training data.

TEST DATA		SN	SP	F1
Clean	KL	74	87	79
	IS	93	90	92
	PCA	74	78	76
	FEE	94	97	95
Noisy	KL	54	59	55
	IS	76	74	75
	PCA	67	57	64
	FEE	70	74	71

Table 6.5: F1, SN and SP for NMF with enhanced and clean training data.

TEST DATA		SN	SP	F1
Clean	KL	85	95	90
	IS	93	94	94
Noisy	KL	75	68	73
	IS	84	76	81
Inter. Noisy	IS	86	82	85

some dominant frequencies in the 0Hz to 50Hz range, similar to subjects with [ELW](#), as discussed in [Section 5.1](#). 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](#).

6.5.3 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 [Section 6.4](#), except that there were pauses with random length of distribution $\mathcal{N}(3, 0.5)$ seconds between noise samples. The [SNR](#) was kept the same at 6dB. The [IS](#) divergence was chosen for this test, and the noisy data was enhanced by [NMF](#). The

result is reported in Table 6.5 under 'Inter. Noisy'. Compared to the continuously noisy data, the SN improved by 2% to 86%, SP increased from 76% to 81%, and the F1 score increased from 81% to 86%. Our proposed system is able to handle intermittent noise.

6.5.4 Further Work with Gradient Boosting Classifier

So far, SVM has been used as the classifier. Linear SVM is a linear classifier and has limited classification capabilities when classifying non-linear data. Hence, Gaussian-kernel SVM and Gradient Boosting Classifier (GBC) were chosen to improve the classification results of enhanced data. We chose the Gaussian-kernel SVM because it is a popular non-linear extension of the linear SVM. Although the focus of our research is not choosing the best classifier but learning features, we have also performed experiments to show that one can potentially increase performance by using non-linear classifiers. Gradient boosting classifier was also chosen because it is a tree-based classifier that is flexible and has given relatively good performance in classification tasks [107]. The interested reader can refer to [88] for more information on the classifiers. Just as how we tuned the hyperparameters of the SVM, we also tuned the hyperparameters of the Gaussian-kernel SVM and GBC. In particular, we tuned the gamma and regularization parameters of Gaussian-kernel SVM. We tuned learning rate, number of estimators, maximum number of features and the maximum depth of each tree for GBC.

The results are reported in Table 6.6. Firstly, for Gaussian-kernel SVM, there was no statistically significant improvement over linear SVM for all three types of test data. Next, GBC is reported in the same table. For clean test data, GBC scored 96%, which was 3% ($p = 0.099$) higher than SVM. Secondly, for noisy test data, there was no statistically significant improvement ($p = 0.469$). Lastly, for intermittently noisy data, the F1 score for GBC was 88%, 3% higher than SVM ($p = 0.0611$).

Possibly, the Gaussian kernel was the non-optimal choice to represent the data. Nonetheless, the results also suggest that the tree-based classifier GBC could be a good choice for specific conditions such as clean and intermittently noisy environments. Although not the main focus of this chapter, we show that we can potentially improve classification results further in specific conditions with tree-based models.

6.6 CONCLUSION

In this chapter, 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. The proposed system using NMF on magnitude spectrum together with NMF signal enhancement proved to be a robust system under the effect of environmental

Table 6.6: F1, SN and SP for NMF with enhanced and clean training data on k-SVM & GBC

	TEST DATA	SN	SP	F1
k-SVM	Clean	97	94	96
	Noisy	87	71	81
	Inter. Noisy	89	78	85
GBC	Clean	97	97	97
	Noisy	87	76	83
	Inter. Noisy	91	85	88

and speech noise, proving superior to other popular audio features and feature selection techniques.

LATENT SSVM AND ROBUST CASCADING DICTIONARY LEARNING

In this chapter, we present a new method of feature transformation that learns dictionaries based on subsets of data called Cascading Dictionary Learning (CDL). In the previous chapter, we showed that NMF could improve the robustness of our proposed system. In this chapter, we attempt to explore different methods to assign latent variables to each sample on the discoveries made in Section 5.1. We investigate the use of Latent Structural SVM (LSSVM) and show that it is not successful in improving results. We also propose a new method that improves classification in noisy conditions. The method selects subsets of data based on a classifier, and learns dictionaries on those subsets. The method is repeated over subsets of data of decreasing size, and the final dictionary is a concatenation of all the cascaded dictionaries. In Section 7.1, we briefly describe the experimental study used in this entire chapter. Then in Section 7.2 we describe experiments on Latent SSVM. With these experiments, we reported results and analysis on the generated latent variables. In Section 7.3, we introduce and elaborate on the CDL method. We also present the results and discussions of our experiments.

7.1 EXPERIMENTAL STUDY

In order to assess the ability of our proposed system fairly with as little bias as possible, we adopted the evaluation framework described in Chapter 4. The data use was collected and processed as described in Chapter 4. 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. Data was segmented into 250ms segments and normalized. The resulting number of samples with ELW is 2574 (8 recordings) and the total number of samples from healthy persons is 2541 (12 recordings). A mixture of background noises were recorded in hospital rooms. Speech signals were also added to the test recordings in order to test its robustness. The method in which the noise was processed was the same as in Chapter 6. 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. The IS divergence was used for this study as it proved to be the best divergence for the classification of ELW as seen in Chapter 6. The soft-margin SVM was used as the classifier to ensure a fair

comparison with the previous experiments. This study also uses the non-parametric approximated randomization test to measure the level of significance of the F1-score when comparing classification results, with 200000 trials as described in [Chapter 4](#). The minimum p-value was 0.000005.

7.2 EXPERIMENTS WITH LATENT SSVM

In our previous work, we investigated the use of only using either segments of inhalations or exhalations and found that it did not help improve classification results. In fact they scored poorer than when full breath cycle recordings were used. However, we can still consider the use of such segmentation by considering latent variables in the classification process to detect excessive lung water in clean and noisy conditions. In this chapter we applied [LSSVM](#) to features learnt by [NMF](#) to account for some latent information relating to hidden signal dynamics, and report some results and analysis.

7.2.1 Methodology

In our previous work in [Chapter 6](#), we proposed a system that can detect pulmonary edema in noisy conditions. The system used a feature learning technique based on [NMF](#), then performs classification with [SVM](#). In this section, we performed some experiments that apply [LSSVM](#) on those learnt features in a bid to improve classification performance by making use of potential hidden signal dynamics.

7.2.2 Latent Structural Support Vector Machine

To understand [LSSVMs](#), SSVMs must first be introduced.

SSVMs learn a prediction rule on a given training set of input-output structure pairs $(x_n, y_n) \in (\mathcal{X} \times \mathcal{Y})^n$ in the form $f_{\mathbf{w}}(x) = \arg \max_{y \in \mathcal{Y}} [\mathbf{w} \cdot \Phi(x, y)]$ where Φ is a joint feature vector that describes the relationship between input x and structured output y , with $\mathbf{w}$ being the parameter vector. $\mathbf{w}$ is determined by maximizing the regularized risk on the training set. The risk is measured by a loss function $\Delta(y, \hat{y})$ that determines how the prediction $\hat{y}$ differs from the truth y . The convex optimization problem to be solved is thus

$$\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \left[\max_{\hat{y} \in \mathcal{Y}} [\Delta(y_i, \hat{y}) + \mathbf{w} \cdot \Phi(x_i, \hat{y})] - \mathbf{w} \cdot \Phi(x_i, y_i) \right],$$

where C is the regularization parameter. It can be solved with cutting-plane or stochastic gradient methods.

Table 7.1: SN, SP and F1 for LSSVM on clean and noisy training data

TEST DATA	SN	SP	F1
Clean	89	91	90
Noisy	78	75	77

Table 7.2: SN, SP and F1 for SVM on clean and noisy training data

TEST DATA	SN	SP	F1
Clean	93	90	92
Noisy	76	74	75

In latent SSVM, the input-output relationship cannot be completely characterized by x and y alone, but is also dependent on a set of latent variables $h \in H$. Thus, the prediction rule is now $f_w(x) = \arg \max_{(y,h) \in \mathcal{Y} \times \mathcal{H}} [w \cdot \Phi(x, y, h)]$. The loss function becomes the extended loss function that measures the difference between the pair $(\hat{y}_i, \hat{h}_i)$ given by the prediction rule and the best latent variable h_i^* that explains the (x_i, y_i) pair. Thus, the optimization problem to be solved is:

$$\min_w \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \left(\max_{(\hat{y}, \hat{h}) \in \mathcal{Y} \times \mathcal{H}} [\Delta(y_i, \hat{y}, \hat{h}) + w \cdot \Phi(x_i, \hat{y}, \hat{h})] \right) - C \sum_{i=1}^n \left[\max_{h \in \mathcal{H}} w \cdot \Phi(x_i, y_i, h) \right].$$

The optimization problem can be solved with the Concave-Convex Procedure. For more information, the interested reader may refer to [108]. Here, the number of latent variables was chosen to be two for each class in view of the inhalation and exhalation cycle.

7.2.3 Results with LSSVM

Table 7.1 shows the test results of LSSVM when two latent variables were used for each class. Table 7.2 shows the test result of using SVM. There were no statistical significance when LSSVM was used over SVM for both clean test data ($p = 0.426$) and noisy data ($p = 0.327$).

7.2.4 Analysis of Latent Variables

Next, the decoded latent variable values were extracted and analyzed. Figure 7.1 shows the breath segmentation of a sample with ELW. Figure 7.2 shows the latent variable values of the same sample. Comparing the breath segments and latent variable values of the clean sample, we can see that the latter was able to differentiate early exhalations from the rest of the recording. These segments are labeled 1, 3 and 5 in Figure 7.2. There was differentiation in other parts of the

breath as well. Upon listening, these segments, however, did not correspond to any unique transient sounds. Nevertheless, these segments still corresponded to distinct segments in the time series of the data. Segments 2 and 4 mark these segments.

Figure 7.3 shows latent variable values when noise is added to the sample. In this case, latent variable values were still able to differentiate early exhalations. Blue segments 2, 4, and 7 show early exhalations. Other latent variables were able to identify segments of inhalations, especially late inhalations. Segments 1, 3 and 6 in the figure reflect late inhalations while segment 8 corresponded to the entire inhalation.

There were also some latent variables (green and red) that corresponded to the wrong class (healthy). Upon listening, they turned out to be parts of the recording with speech noise present. The green segments were speech from NIST database while the red segment was singing from background music. This showed that speech signals were mistaken by the LSSVM as segments of a healthy recordings. Despite that, it also means that LSSVM could potentially be used to differentiate speech noises from other hospital noises, and a more robust latent variable classifier could be used to classify ELW in the presence of speech noise. Another interesting point to note was that latent variables could potentially differentiate speech from singing. The latent variable model could also be used to take into account similar types of noises to improve robustness. However, more investigation is necessary to verify this.

Since the latent variables were able to identify early exhalations, late inhalations and speech/singing noises, future work may include choosing more latent variables to represent more distinctly the different latent information found in both clean and noisy data. This could in turn improve classification results.

7.3 EXPERIMENTS WITH CASCADING DICTIONARY LEARNING

Although analysis by LSSVM showed that latent variables could differentiate early exhalations, late inhalations as well as certain distinct segments from the rest of the breath sounds, as well as decoded latent variable values were able to identify segments of recording that had speech noises, they did not improve classification results. Nonetheless, breath sounds contain latent information that could potentially be exploited to increase classification performance. We next attempted to generate latent variables in a supervised fashion. We used SVM to group subsets of data and calculate dictionaries on each subset of data. These subsets can be thought of as latent variables that have been grouped into these subsets.

Figure 7.1: Breath segmentation of a sample

Figure 7.2: Latent variables of a clean sample

Figure 7.3: Latent variables of a noisy sample

7.3.1 Methodology of Proposed Method

As described in previous chapters, previous works mainly focus on classifying DAS. In our previous work in Chapter 6, we proposed a system that can detect pulmonary edema in noisy conditions. The system used a feature learning technique based on NMF, then performs classification with SVM. In this chapter, we seek to improve that method, and show that learning dictionaries from subsets of data can improve classification results. Similarly to Chapter 6, the system was also tested under the introduction of environmental noise found in hospitals, and a noise dictionary was concatenated during the learning of those sub dictionaries.

7.3.2 Cascading Dictionary Learning

The proposed method learns dictionaries from subsets of data, then repeatedly learns dictionaries from smaller subsets of data over decreasing size. Figure 7.4 shows a general overview of the algorithm, with an example where three dictionaries are obtained using the CDL algorithm. The algorithm is as follows: for a given dataset X with targets Y in the form of $(x_1, y_1), \dots, (x_n, y_n)$, we learn the dictionary W and activation matrix H by optimizing the divergence found in Equation 6.1. We train the classifier $\alpha(h_n, y_n)$ to produce a prediction $\hat{y}_n$ for each h_n . We aim to learn dictionaries based on samples that are the most easily classifiable. Thus, we remove all corresponding x_n where $y_n \neq \hat{y}_n$. We call the removed samples $X_{1,1}^r$ and the remaining samples $X_{1,1}^q$. We learn a new W and H with $X_{1,1}^q$, and train the classifier to again get $\hat{y}_n$. We retrain the classifier because the new separating hyperplane may produce predictions for the remaining samples that

are wrong. We then remove x_n where $y_n \neq \hat{y}_n$, and merge them with $X_{1,1}^r$ to get $X_{1,2}^r$. The remaining samples are $X_{1,2}^q$. We iterate this until the difference of the classification error ε between each iteration is within a certain tolerance τ , if $X_{1,t}^q$ has only one class, or if $\varepsilon = 0$. The resulting $\mathbf{W}_1$ is the first cascaded dictionary. This first dictionary thus represent the first cluster of samples which produces the lowest training error. A new $\mathbf{W}$ and $\mathbf{H}$ is learnt from remainder $X_{1,t}^r$ and the process is iterated to get $\mathbf{W}_2$. We end the iterative behavior when $X_{c,t}^r$ is smaller than a number σ , if $Y_{c,t}^r$ is of one class, or if c number of dictionaries are obtained. Thus a series of $\mathbf{W}_c$ is obtained that each represent their own cluster of data that in themselves produce the minimum training error. The dictionaries are concatenated together to get the final dictionary. A subset of dictionaries can also be used instead of all C dictionaries lest later dictionaries represent outliers or hard-to-classify training samples. In our preliminary studies, we found that 0.005 was a good value of τ . If the sub-sample sets had an unbalanced class distribution, we re-sampled the class with fewer samples using random oversampling. The summary of the algorithm is summarized in [algorithm 1](#).

In our experiments, we have both clean training data, and noisy training data generated from the clean training data. Thus instead of learning new dictionaries with $X_{c,t}^q$ from the noisy training dataset X^N , we learn the dictionaries from samples of the clean training dataset X^C corresponding to $X_{c,t}^q$ from the noisy training dataset. We then concatenate the noise dictionary $\mathbf{W}_{\text{noise}}$ similarly to [Section 6.3.2](#) and learn $\mathbf{H}$ from the new concatenated dictionary and training data.

Result: $\mathbf{W} = [\mathbf{W}_1 \mid \mathbf{W}_1 \mid \dots \mid \mathbf{W}_C]$
Initial: $X^C, X_{0,0}^r = X^N, Y, \ell_2, K, \mathbf{W}_{\text{noise}}, c$;
while $X_{c,t}^r < \sigma \vee Y_{c,t}^r \in \{0 \parallel 1\} \vee c \text{ dicts.}$ **do**
 $X_{c,0}^r = \emptyset$;
 $\varepsilon_{c,0} = 0$;
 while $\varepsilon_{c,t} - \varepsilon_{c,t-1} > \tau \vee Y_{c,t}^q \in \{0 \parallel 1\}$ **do**
 Calculate $\mathbf{H}_{c,t}^C$ from $X_{c,t}^r$ that corresponds to $X_{c,t}^{r,N}$;
 Calculate $\alpha(h_{c,t}^N, y_{c,t}, c)$ to get $\hat{y}_{c,t}^N$;
 $X_{c,t}^r = X_{c,t-1}^r \cup \{x_{c,t}^N\}$, where $\hat{y}_{c,t}^N \neq y_{c,t}$;
 end
 $X_{c,t}^q = X_{c-1,t}^r \setminus X_{c,t}^r$;
 Calculate $\mathbf{W}_c$ from $X_{c,t}^q$
end

Algorithm 1: CDL Algorithm

7.3.2.1 NMF as Benchmark

We compared results of the proposed method with results from [Chapter 6](#), mainly the [NMF](#) system. We compare the best result of that

Figure 7.4: Steps showing how to obtain cascaded dictionaries.

Table 7.3: Hyperparameters optimized by the system

ALGORITHM	PARAMETER	RANGE
NMF of data and noise	No. of basis vectors	Optimized from Chapter 6
	ℓ_2 regularization	0 – 1000
	Divergence	Itakura-Satio

chapter, with the IS divergence. We re-tuned the ℓ_2 regularization by a grid search, just as how we used the same hyperparameter-tuning in Chapter 6 to fit it to our problem. However, to reduce computational time, the numbers for basis vectors used for both data and noise for this study were the optimized number of basis vectors found in Chapter 6. We tested the cascading dictionary learning using the IS divergence. Table 7.3 shows the parameters optimized as well as the range of those tested values.

7.3.2.2 *k*-Means Clustering as Benchmark

We also compare results of our proposed method with results with *k*-MC. This is because the CDL method makes use of sub-sets of data which can also be seen as clusters of data. In addition to that, *k*-MC has been a popular clustering algorithm due to its simple and fast procedure. Specifically, it has been used in various audio segmentation and classification tasks [109–112]. In the chapter, we choose the number of centroids to be equal to the number of dictionaries learnt in the CDL method. After forming these clusters, we learn dictionaries from them and concatenate them together to form one long dictionary. We also use the IS divergence, and we concatenate the noise dictionary with the final dictionary, similarly to the proposed system in Chapter 6. We also tune the final number of dictionaries used for testing, similarly to CDL, which will be described in detail in the next section.

7.3.2.3 Task Driven NMF as Benchmark

Lastly, we compare results from our proposed method with results with Task Driven Non-negative Matrix Factorization (TDNMF). TDNMF jointly optimizes the classifier and dictionary learning, and is thus a supervised variety of NMF. CDL is also semi-supervised; hence we decided to use TDNMF as a comparison. Influenced by the classifier, the basis vectors are encouraged to explain the discriminative information in the data while keeping a low reconstruction cost. The algorithm was used to improve acoustic scene classification [49]. The divergence used was the ℓ_2 -norm, combined with a logistic regression classifier. In this chapter, we use TDNMF without temporal integration with a

Table 7.4: Hyperparameters optimized for TDNMF.

ALGORITHM	PARAMETER	RANGE
TDNMF	No. of basis vectors	128, 256
	NMF ℓ_2 regularization	0, 50, 100, 200, 400, 800, 1000
	NMF ℓ_1 regularization	0, 50, 100, 200, 400, 800, 1000
	LR ℓ_2 regularization	0.001, 0.01, ..., 10000

binomial logistic regression classifier. We also tune hyperparameters of TDNMF. Table 7.4 shows the range of hyperparameters optimized.

7.3.3 Results and Discussion

7.3.3.1 Behavioral Analysis

First, we analyzed the behavior of the algorithm and studied how to optimize its performance. We calculated the training error across increasing number of concatenated $\mathbf{W}_c$ until the stopping criterion was met. We observed that the training error consistently decreased as more cascaded dictionaries were added. Figure 7.5 shows the training error over one such fold. The dashed line shows the training error from using one dictionary over the whole training set. As one can observe, a total of 3 cascaded dictionaries were learnt. The number of dictionaries learnt are dependent on the data. The training error does not guarantee a monotonic decrease as more dictionaries are added. However, the change in testing error generally follow the change in training error, and thus the choice of cascaded dictionaries for testing data were chosen according to the cascaded dictionaries that gave the best training score. There were also cases where cascading dictionary learning did not produce better test scores. For these cases, we chose to use the standard NMF with IS divergence over the whole dataset. This was decided by using the inner-CV framework, and was part of the system optimization procedure. Using again the same evaluation framework and data, the SN, SP and F1 score of clean, noisy, and intermittent noisy test samples were calculated.

7.3.3.2 k-Means Clustering

First as a baseline, we tested the clustered NMF with k-MC. To do this, we used k-MC to find clusters of noisy data, then learnt separate dictionaries from clean data that correspond to the noisy data in these clusters. We concatenated the dictionaries with the noise dictionary to form our final dictionary and used it to learn features for classification. We finally trained the classifier and calculate predictions for the test

Figure 7.5: Training error across increasing W_c

Table 7.5: F1, SN and SP for NMF with k-means clustering

	SN	SP	F1
Clean	97	93	95
Noisy	84	75	80
Inter. Noisy	87	82	85

data. The results are reported in Table 7.5. We next compared them with the best results of Chapter 6 found in Table 7.6 (right). There is no improvement for either clean test data ($p = 0.573$), noisy test data ($p = 0.742$) or intermittently noisy test data ($p = 0.712$). We can conclude that *k-MC* with NMF does not help improve classification performance.

7.3.3.3 Task Driven NMF

Next, we tested TDNMF. We compare the average accuracy across the inner CVs and that of NMF with IS divergence. The average accuracy for TDNMF was $16 \pm 1\%$ lower than the average accuracy. Thus, we can conclude, without comparing the F1 score of the outer CV, that TDNMF is unsuitable for our task. One reason that TDNMF performed poorer could be that when Bisot *et al.* used basic NMF for acoustic scene classification, they reported that the choice of β -divergence has almost no impact on the result. However, in the previous chapter we have shown that the IS divergence fared better than KL or ℓ_2 -norm, especially with noisy data. The divergence chosen for TDNMF was the ℓ_2 -norm, and was unsuitable for our task. This could have contributed to the poorer result, since the ℓ_2 -norm for our task fared the poorest, as shown in the previous chapter.

7.3.3.4 Cascading Dictionary Learning

We next tested CDL. The SN, SP and F1 score are shown in Table 7.6 (left). We will be comparing them with the best results of Chapter 6 found in Table 7.6 (right). The F1 score for clean test data decreased by 2% ($p = 0.0341$). The F1 score for noisy test data increased by 2% ($p = 0.0759$). Lastly, F1 score for intermittently noisy test data remained the same. Although the cascading dictionary learning algorithm comprised clean test scores for the sake of increasing noisy test scores, it can be fair to say that it is an advantageous trade-off. Firstly, one can assume that noisy conditions are more common than noise-free conditions in hospital rooms, especially public hospitals where there is higher traffic. Second, the F1 measure for clean test data is already very high at 92%, and the decrement from 94% is less significant than an increment from 81% to 83%. These results are promising but require further

Table 7.6: F1, SN and SP for NMF with cascaded dict. (left) and best results from Chapter 6 (right).

	SN	SP	F1		SN	SP	F1
Clean	92	93	92	Clean	93	94	94
Noisy	87	78	83	Noisy	84	76	81
Inter. Noisy	86	83	85	Inter. Noisy	86	82	85

investigation to improve its effectiveness in both classifying clean test data and noisy test data.

One limitation is that [CDL](#) finds a local minimum for clusters of data that produce the minimum training error for resulting $\mathbf{H}$ features. It does not guarantee a global minimum, similarly to basic [NMF](#). More work can be done to improve the optimization method in finding better local minima. Another limitation is that the supervision is based on the noisy training set, and it is not generalized to clean test data. This could have accounted for the nature of the reported results. An improvement can be made in future work in developing a [CDL](#) that can generalize to both clean and noisy test data.

We have also performed preliminary studies of introducing more non-linearity by using a [GBC](#) instead of an [SVM](#). However, the [GBC](#) results in overfitting as it returns a perfect training score, which leads to a single dictionary rather than cascaded dictionaries. We also performed preliminary studies with naive classifiers like Gaussian Naive Bayes and Linear Discriminant Analysis to encourage underfitting so that [CDL](#) can focus on learning better dictionaries in compensation. Results showed that those classifiers performed too poorly for [CDL](#) to work effectively. Further work can be done on adapting a suitable classifier for the [CDL](#) so that performance can be improved on.

7.3.3.5 Computational Complexity

Lastly, we have attempted to measure the computational complexity in terms of time taken for our computer to generate all dictionaries for all folds in a given set of hyperparameters. For a 32-core CPU with limitless RAM, conventional [NMF](#) took a total of 176 minutes while [CDL](#) took a total of 351 minutes. [CDL](#) definitely has a higher computational complexity, but it is also dependent on the sample size of each subset and the number of subsets generated.

7.3.4 Additional Experiments with DCASE16 Audio Scene Classification

To test the generalization ability of the [CDL](#), we benchmarked [CDL](#) against [NMF](#) on the DCASE16 audio scene classification dataset [113].

7.3.4.1 Dataset Description

The TUT Acoustic scenes 2016 dataset was used for the task. The dataset consists of recordings from various acoustic scenes, all having distinct recording locations. For each recording location, 3-5 minute long audio recording was captured. The original recordings were then split into 30-second segments for the challenge. They were a mix of noisy and quiet acoustic scenes, with a total of 15 different scenes (classes). Recordings were made using a Soundman OKM II Klassik/studio A3, electret binaural microphone and a Roland Edirol R-09 wave recorder using 44.1 kHz sampling rate and 24 bit resolution. In our benchmark experiment, we converted the stereo recordings to mono. While the dataset was chosen because it consisted of audio data which was similar to our original task, it was also bigger in size and more challenging to classify. Noises and transient sounds appear in multiple types of scenes. For example, bird noises could be heard in a forest but also in a park. Also, some travelling scenes are similar, for example in the train and tram. This classification problem could thus benefit from cascading dictionaries where each dictionary could represent a specific latent variable of a subset of data, where within-subset samples were more easily classifiable.

7.3.4.2 Experimental Study

Since Bisot *et. al.* discovered that using CQT coefficients were effective as features for this dataset [114], we adopted the same data pre-processing procedure and low-level feature extraction procedure. Each sample was two second long, with CQT averaged over a sliding window of length 4096 points and a hop-size of 100%. The number of bins for the CQT was 144, resulting in 144 features and 15 samples per recording. Evaluation and hyper-parameter tuning was done on a train-validation split, with the split following the first fold as given in the dataset meta-descriptions. Training of the model was done on the train split, hyper-parameter tuning was done on the validation split, and performance evaluation was done on a separate test set provided. There were a total of 1170 recordings, with 78 recordings per class. 880 recordings were used for the training split, 290 recordings were used for the validation split, and 390 recordings were used for testing. The hyper-parameters tuned were the same as before, with additional tuning on the final number of dictionaries chosen when CDL was used. For each recording, the activation matrix of the corresponding 15 samples were averaged, and the classifier was trained on these final activation matrices. The approximate randomization test was used to test for statistical significance, and 1 million iterations were chosen for 390 test recordings. Multi-class F1-measure was used as the performance metric.

7.3.4.3 Modifications to CDL

Previously, a linear classifier was used to iteratively remove wrongly classified training samples. For this experiment, we discovered that a tuned SVM with NMF gave a training accuracy score of 98%. Thus, we chose to use a weak classifier to allow more misclassifications and enable the CDL to form subsets that can produce better accuracy under a weaker classifier. In this regard, we chose a decision tree classifier with a maximum depth of 8, giving a training accuracy of 82% with NMF. A constraint was also placed on the sifting procedure, where the samples removed were class-balanced. Thus, the number of samples removed from each class per sift followed the lowest number of samples out of all classes. For cases where a particular class was not misclassified at all, we randomly removed samples from that class equal to the lowest number of samples out of all classes. Sifting ended for one subset when the error dropped by at least half compared to the start of the sifting procedure for one subset.

7.3.4.4 Results and Discussion

Figure 7.6 shows the change in error and sample count across sifts. The green partitions represent the sifts just before a new subset. The error starts out to be around 35%. For the first sift, the error eventually drops to around 17% before one subset is calculated. The first subset also consists of half the samples. 4 more subsets were generated and the last subset were made up of samples that were difficult to classify accurately. Subset size also decreased as each subset was generated. Across all subsets, the error managed to decrease further from the original error. CDL managed to improve classification accuracy for each subset using a weak classifier.

The F1-measure for both NMF and CDL are reported next. NMF produced an F1 score of 67.3% while CDL produced an F1 score of 69.7%. CDL managed to perform better than NMF by 2.4% ($p = 0.177$).

Since a weak tree-based classifier was used to cluster the data into subsets, we lastly performed experiments using the GBC instead of SVM. NMF produced an F1 score of 70.4% while CDL produced an F1-score of 74.1%, which was 3.8% higher ($p = 0.0826$). Although the p-values are borderline significant, it is reasonable as the CDL was proposed as a robust dictionary learning method, specifically to deal with realistic noise (no free lunch theorem). Nonetheless, with further work CDL can be improved to better generalize across different kinds of data, and eventually be a model that is both more robust and diversified.

7.4 CONCLUSION

In this chapter we first presented a study on using latent variables to classify breath recordings between subjects with excessive lung water

Figure 7.6: Errors and sample count across each sift. The green partitions represent the sifts just before a new subset

and with healthy subjects. [NMF](#) was used as a feature learning technique to learn features from the magnitude spectrum. Using [LSSVM](#) did not improve classification performance for both clean and noisy test data compared to when [SVM](#) was used as classifier. The latent variables were analyzed in time domain and analysis shows that the latent variables could differentiate early exhalations, late inhalations as well as certain distinct segments from the rest of the breath sounds. Lastly, the decoded latent variable values were able to identify segments of recording that had speech noises. Thus, we showed that breath sounds contain latent information that could potentially be exploited to increase classification performance in the future.

In the second section, we proposed a cascading dictionary learning method that learns dictionaries from subsets of training data. We have applied the proposed method on classification of breath sound from subjects with excessive lung water against subjects with healthy lungs. Additionally, we have tested the robustness of the algorithm using clean and noisy test data. The proposed system proved to be more robust under the effect of environmental and speech noise than using only standard [NMF](#) with [IS](#) divergence over the whole training dataset. It also proved to be more robust than classic clustering algorithm [k-MC](#). Even though the proposed method scored poorer for clean testing data, it improved results for noisy test data and can be considered a preferable exchange in performance capabilities. These results are promising but require further investigation to improve the [CDL](#) approach's effectiveness in both classifying clean test data and noisy test data.

CONCLUSION

In this chapter we present a summary of this thesis and suggest some future work directions. Section [Section 8.1](#) briefly summarizes this thesis with its main contributions, while [Section 8.2](#) points out some possible future work.

8.1 THESIS SUMMARY

In the thesis, we focused on the classification of pulmonary edema with the use of breath recordings through the chest wall. Specifically, we worked with offline features extraction and classification. We also concentrated on developing algorithms that are robust to realistic hospital noise, where the noises were prerecorded and noisy recordings were generated. The contributions made can be summarized into three parts:

1. In the first part, we studied the effectiveness of using known feature engineering in classifying pulmonary edema signals. The best F1 measure was 92% without the use of feature selection. We then introduced the use of [EMD](#) and found that using 17 intrinsic mode functions together with various other engineered features improves the performance on the detection of pulmonary edema to 95%. We also performed experiments on the segregation of breath phases– inhalation and exhalation. We showed that segregating the breath recordings in such a way does not improve classification performance, thereby concluding that the entire breath cycle is preferred in detecting pulmonary edema.
2. For the second part, we study the effect of adding realistic hospital noise to the data, thereby testing the robustness of the popular engineered features as well as the improvements made by empirical mode decomposition. We showed that the classification performance (F1 measure) dropped by at least 23%. We next introduced the use of [NMF](#) for feature learning and presented a study on the use of various divergences. We discovered that the [IS](#) divergence was the most helpful for our task, and postulated that it was due to the scale invariance of that divergence. Using the [IS](#) divergence, we found that our system improved F1 measure by 11% for the detection of pulmonary edema in the presence of hospital noise.
3. In the final part, we used [LSSVM](#) to show that breath sounds have latent information that could be exploited. The latent variables

were analyzed in time domain and analysis shows that the latent variables could differentiate segments of both clean and noisy audio signals. We also proposed a new [CDL](#) algorithm. The algorithm learnt a series of dictionaries based on subsets of training data. We proposed a method in which to best choose the number of dictionaries used to classify both clean and noisy test data, basing it on both training error and inner cross validation test error. We showed that although there was a disimprovement in the classification of clean test data by 2% F1 measure, there was an improvement in the classification of noisy test data by 2%.

8.2 FURTHER WORK

The works presented in this thesis are at the forefront of pulmonary edema detection. Thus, different aspects of our work can be improved in the future to obtain more accurate and reliable algorithms. These improvements can be also split into three main parts.

1. For the first part, more studies can be made on automatic selection of most useful [IMFs](#). [IMFs](#) are adaptive and there is no guarantee that the same [EMD](#) procedure on different signals can produce, for the same [IMF](#) level, similar bandwidth in its frequency components. Thus it can be difficult to determine which [IMFs](#) are more useful. Rather than choosing [IMFs](#) based on classification performance as we have, more studies can be made to propose adaptive selection techniques that are based on frequency bandwidth of the [IMFs](#), or even other parameters of the decomposition. There also exist other modified [EMD](#) algorithms that are more robust to noise and give more stable decomposition [115]. However, they take a much longer time to process. If given more time to apply these alternatives, we could improve the decomposition process and choose the [IMFs](#) that contain more useful measures further increasing classification performance. Other than improvements in the [EMD](#) area, we can also make further improvements in feature selection. Our box-plot ([Figure 5.3](#)) showed that the top features were related to [LSFs](#) and derivatives of [MFCC](#). This suggests that other engineered features might have been less useful or even end up as noise. More studies could have been made to eliminate those features other than using [RFE](#). We can, for example, study the correlation of the features. If improvements in feature selection can be made, we can potentially improve the classification performance of pulmonary edema both in clean and noisy conditions. We also did not answer the question of why features such as [LSFs](#) and derivatives of [MFCC](#) were more useful in the classification of pulmonary edema. If new insights are gained from the relevance of

these features, we can potentially propose new features specific to the classification of pulmonary edema based on the more useful features in [PAF](#).

2. For the second part, we separated the feature learning and classification tasks. Logically speaking, due to the fact that the [MU](#) rule can only guarantee local convergence, it is highly possible that there exist dictionaries with the Itakura-Saito divergence that can produce better training loss with its produced features. In relation to that, there exist supervised [NMF](#) variants that co-learn the dictionary and classifiers parameters, which have shown to be effective in acoustic scene classification [114]. Bisot proposed a method that co-learns the dictionary and parameters of logistic regression to learn dictionaries that also minimize the logistic loss of the classifier with the features they produce. However, this supervised variant used the ℓ_2 divergence, and in our experiments we have shown that the ℓ_2 divergence produced poor results. Thus, a similar approach adopted from Bisot's work can be used in future work to develop a supervised dictionary learning method that can be used with the Itakura-Saito or Kulback-Leibler divergence. Apart from improving dictionary learning, future work can also include improvements on enhancements in the presence of noise. In that regard, probabilistic [NMF](#) methods are popular in source separation and have shown success [116], and the enhancement problem can be modeled as a pseudo-source separation model, separating noise and data elements. Since we have shown the [IS](#) divergence to be the best divergence, we can study the use of the Gaussian composite model [106], which is closely related to [NMF](#) with the [IS](#) divergence. The Gaussian composite model essentially models the magnitude spectrum as a sum of uncorrelated centered Gaussian components structured through their variance. Using probabilistic models can provide better estimations for factorization under noise since we have prior knowledge of distributions of both data and also different types of noise. Also, we can modify the use of the Gaussian composite model by training a classifier on both separated sources, thus making full of use of any features that could have been separated wrongly. We can lastly investigate modeling different types of hospital noise as separate noise dictionaries to better enhance noisy samples or separate hospital noise from respiratory sounds.
3. For the third part, more work can be done to optimize the sample subsets. Incorporating more sophisticated techniques can help improve the dictionary learning, and in turn increase classification performance. Another way of looking at the problem would be assigning latent states to the sample subsets. That would

then be similar to latent SVM [117, 118], where latent states were assigned to different image samples based on their features and samples for each latent state had their own SVM classifier trained. The appendix contains some preliminary experiments we conducted with Latent Structural SVM. We can thus extrapolate that idea by assigning latent states to samples such that sample for each latent state are represented by their own dictionary and trained on separate classifiers. That is one direction we can take for future work for the third part. An alternative approach would be to consider the temporal dynamics as a basis for latent states hidden in the data. As explained in [116], we can potentially use probabilistic modeling to allow the activation of certain dictionary basis vectors for a given state. Also, if we use hidden markov models structures for such models as found in [119–121], we can incorporate other types of temporal structures like breath segmentation or noise differentiation. Likewise we can still train a classifier on both separated sources.

Overall, more data can be collected to increase the generalization ability of our proposed systems. We recognize that a small data size is a significant limitation to our work. However, data collection is difficult as physicians have to double-up during the process, having to juggle both data collection and consultations.

PUBLICATIONS

- [1] K. J. Hong, S. Essid, W. Ser, and D.C.-G. Foo. "A robust audio classification system for detecting pulmonary edema." In: *Biomedical Signal Processing and Control* 46 (Sept. 2018), pp. 94–103.
- [2] Kah Jun Hong, Wee Ser, and David Chee-Guan Foo. "An analysis of sounds for lungs with excessive water." In: *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*. IEEE. 2016, pp. 3662–3665.
- [3] Kah Jun Hong, Wee Ser, Zhiping Lin, and David Chee-Guan Foo. "Acoustic detection of excessive lung water using sub-band features." In: *Circuits and Systems Conference (DCAS), 2015 IEEE Dallas*. IEEE. 2015, pp. 1–4.

CALCULATION DETAILS FOR FEATURE SELECTION

In this chapter we describe in detail the calculations to obtain these features.

LINEAR PREDICTIVE COEFFICIENTS

LPC are coefficients resulting from auto-regressive modeling and are obtained by solving the Yule-Walker equations:

$$\begin{bmatrix} r(1) & r(2)^* & \cdots & r(p)^* \\ r(2) & r(1) & \ddots & \vdots \\ \vdots & \ddots & \ddots & r(2)^* \\ r(p) & \cdots & r(2) & r(1) \end{bmatrix} \begin{bmatrix} a(2) \\ a(3) \\ \vdots \\ a(p+1) \end{bmatrix} = \begin{bmatrix} -r(2) \\ -r(3) \\ \vdots \\ -r(p+1) \end{bmatrix}$$

where $\begin{bmatrix} r(1) & r(2) & \cdots & r(p+1) \end{bmatrix}$ is autocovariance estimate for the signal w modeled as

$$\mathbf{W}\mathbf{a} = \mathbf{b}$$

where

$$\mathbf{W} = \begin{bmatrix} w(1) & 0 & \cdots & 0 \\ w(2) & w(1) & \ddots & \vdots \\ \vdots & w(2) & \ddots & 0 \\ w(m) & \vdots & \ddots & w(1) \\ 0 & w(m) & \ddots & w(2) \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & w(m) \end{bmatrix}, \mathbf{a} = \begin{bmatrix} 1 \\ a(2) \\ \vdots \\ a(p+1) \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

The vector $\mathbf{a}$ is used as the features for classification.

LINE SPECTRAL FREQUENCY COEFFICIENTS

LSF decomposes the p^{th} order linear predictor $A(z)$ into a symmetrical and anti-symmetrical part denoted by the polynomials $P(z)$ and $Q(z)$ respectively.

For a given

$$A(z) = 1 + \sum_{i=1}^p a_i z^{-i}$$

$P(z)$ and $Q(z)$ are

$$P(z) = A(z) + z^{-(p+1)} A(z^{-1})$$

$$Q(z) = A(z) - z^{-(p+1)} A(z^{-1})$$

$P(z)$ and $Q(z)$ are thus used as features for classification. We also use the 1st and 2nd derivatives of [LSF](#) as features.

MEL-FREQUENCY CEPTRAL COEFFICIENTS

[MFCC](#) is done in two steps:

1. Calculate a Mel-spectrogram by using a bank of log triangular band-pass filters in the Mel-scale, where the formula to convert f Hz to m mels:

$$m = 1127 \times \log \left(1 + \frac{f}{700} \right)$$

2. Apply discrete cosine transform to the Mel-spectrogram to produce the [MFCC](#) features.

$$X_k = \sum_{n=0}^{N-1} x_n \cos \left[\frac{\pi}{N} \left(n + \frac{1}{2} \right) k \right] \quad k = 0, \dots, N-1$$

where x_n is the n^{th} mel magnitude with N number of mel-spectrum coefficients, and X_k is the k^{th} [MFCC](#) feature.

LOUDNESS COEFFICIENTS

The bark scale is a frequency scale on which equal distances correspond with perceptually equal distances, ranging from 1 to 24 Barks, corresponding to the first 24 critical bands of hearing. The Bark band edge and center frequencies given in Hz can be found in [\[122\]](#). The loudness for each band is calculated using

$$N'(k) = \frac{E(k)^{0.23}}{N}$$

where k is the index of a band, $E(k)$ is the energy in the k^{th} band, and $N = \sum N'(k)$

SHARPNESS OF LOUDNESS

The sharpness is calculated with the formula [122]:

$$A = 0.11 \times \frac{\sum_k k \cdot g(k) \cdot N'(k)}{N} \quad \begin{cases} g(k) = 1 & k < 15 \\ g(k) = 0.066e^{0.171 \cdot k} & k \geq 15 \end{cases}$$

SPECTRAL DECREASE

The spectral decrease is weighted average of the decrease of the magnitude spectrum between consecutive frames. It can be useful for the classification of pulmonary edema because the magnitude spectrum of ELW samples tends to contain temporal artifacts that change in the magnitude quickly. The spectral decrease can be computed by [122]:

$$S_{\text{decrease}} = \frac{1}{\sum_{f=2:F} a(f)} \sum_{f=2:F} \frac{a(f) - a(1)}{f - 1}$$

where $a(f)$ is the magnitude of frequency f

SPECTRAL FLATNESS

The formula for spectral flatness is [122]:

$$S_{\text{flatness}} = \frac{e^{\frac{1}{N} \sum_f \log(a_f)}}{\frac{1}{N} \sum_f a_f}$$

SPECTRAL FLUX

Spectral flux is calculated by [122]

$$S_{\text{flux}} = \frac{\sum_f [a_f(t) - a_f(t-1)]^2}{\sqrt{\sum_f a_f(t)^2} \sqrt{\sum_f a_f(t-1)^2}}$$

SPECTRAL SLOPE

The spectral slope can be calculated with the following:

$$S_{\text{slope}} = \frac{F \cdot \sum_f f a_f - \sum_f f \sum_f a_f}{F \cdot \sum_f f^2 - (\sum_f a_f)^2}$$

SPECTRAL MOMENTS

The spectral moments are defined as follows:

$$S_1 = \mu_1$$

$$S_2 = \sqrt{\mu_2 - \mu_1^2}$$

$$S_3 = \frac{2(\mu_1)^3 - 3\mu_1\mu_2 + \mu_3}{(S_2)^3}$$

$$S_4 = \frac{-3\mu_1^4 + 6\mu_1\mu_2 - 4\mu_1\mu_3 + \mu_4}{(S_3)^4} - 3$$

where S_1 is the centroid, S_2 is the spread, S_3 is the skewness, S_4 is the kurtosis, and $\mu_i = \frac{\sum_{k=0}^{N-1} k^i \cdot A(k)}{\sum_{k=0}^{N-1} A(k)}$ where $A(k)$ is the magnitude of the k^{th} component of the magnitude spectrum.

BIBLIOGRAPHY

- [1] Kah Jun Hong, Wee Ser, Zhiping Lin, and David Chee-Guan Foo. "Acoustic detection of excessive lung water using sub-band features." In: *Circuits and Systems Conference (DCAS), 2015 IEEE Dallas*. IEEE. 2015, pp. 1–4 (cit. on pp. [vii](#), [2](#)).
- [2] Kah Jun Hong, Wee Ser, and David Chee-Guan Foo. "An analysis of sounds for lungs with excessive water." In: *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*. IEEE. 2016, pp. 3662–3665 (cit. on pp. [vii](#), [2](#)).
- [3] K. J. Hong, S. Essid, W. Ser, and D.C.-G. Foo. "A robust audio classification system for detecting pulmonary edema." In: *Biomedical Signal Processing and Control* 46 (Sept. 2018), pp. 94–103 (cit. on pp. [viii](#), [2](#)).
- [4] Parveen J. Kumar and Michael L. Clark. *Kumar & Clark's clinical medicine*. Saunders Elsevier, 2009 (cit. on pp. [1](#), [5](#)).
- [5] Steven McGee MD. *Evidence-Based Physical Diagnosis*. Saunders, 2012 (cit. on p. [1](#)).
- [6] Zoltán Jambrik, Luna Gargani, Agnes Adamicza, József Kaszaki, Albert Varga, Tamás Forster, Mihály Boros, and Eugenio Picano. "B-lines quantify the lung water content: a lung ultrasound versus lung gravimetry study in acute lung injury." In: *Ultrasound in medicine & biology* 36 (12 Dec. 2010), pp. 2004–2010 (cit. on p. [5](#)).
- [7] Zoltan Jambrik, Simonetta Monti, Vincenzo Coppola, Eustachio Agricola, Gaetano Mottola, Massimo Miniati, and Eugenio Picano. "Usefulness of ultrasound lung comets as a nonradiologic sign of extravascular lung water." In: *The American journal of cardiology* 93 (10 May 2004), pp. 1265–1270 (cit. on p. [5](#)).
- [8] Eugenio Picano, Francesca Frassi, Eustachio Agricola, Suzana Gligorova, Luna Gargani, and Gaetano Mottola. "Ultrasound lung comets: a clinically useful sign of extravascular lung water." In: *Journal of the American Society of Echocardiography : official publication of the American Society of Echocardiography* 19 (3 Mar. 2006), pp. 356–363 (cit. on p. [5](#)).
- [9] Peter J Fagenholz, Jonathan A Gutman, Alice F Murray, Vicki E Noble, Stephen H Thomas, and N Stuart Harris. "Chest ultrasonography for the diagnosis and monitoring of high-altitude pulmonary edema." In: *Chest* 131 (4 Apr. 2007), pp. 1013–1018 (cit. on p. [5](#)).

- [10] G. R. Wodicka, K. N. Stevens, H. L. Golub, E. G. Cravalho, and D. C. Shannon. "A model of acoustic transmission in the respiratory system." In: *IEEE Transactions on Biomedical Engineering* 36.9 (Sept. 1989), pp. 925–934 (cit. on p. 5).
- [11] K. Mulligan, A. Adler, and R. Goubran. "Detecting regional lung properties using audio transfer functions of the respiratory system." In: *Proc. Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society*. Sept. 2009, pp. 5697–5700 (cit. on p. 6).
- [12] H. M. Zaeim, C. Scheffer, M. Blanckenberg, and K. Dellimore. "Evaluation of the use of frequency response in the diagnosis of pleural effusion on a phantom model of the human lungs." In: *Proc. 36th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society*. Aug. 2014, pp. 3418–3421 (cit. on p. 6).
- [13] Renard Xaviero Adhi Pramono, Stuart Bowyer, and Esther Rodriguez-Villegas. "Automatic adventitious respiratory sound analysis: A systematic review." In: *PloS one* 12.5 (2017), e0177926 (cit. on p. 7).
- [14] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. "Music emotion recognition: A state of the art review." In: *Proc. ISMIR*. Citeseer. 2010, pp. 255–266 (cit. on pp. 7, 37).
- [15] Zhouyu Fu, Guojun Lu, Kai Ming Ting, and Dengsheng Zhang. "A survey of audio-based music classification and annotation." In: *IEEE transactions on multimedia* 13.2 (2011), pp. 303–319 (cit. on pp. 7, 37).
- [16] Nandini Sengupta, Md Sahidullah, and Goutam Saha. "Lung sound classification using cepstral-based statistical features." In: *Computers in Biology and Medicine* 75 (2016), pp. 118–129 (cit. on pp. 7, 12).
- [17] Shohei Matsutake, Masaru Yamashita, and Shoichi Matsunaga. "Abnormal-respiration detection by considering correlation of observation of adventitious sounds." In: *Signal Processing Conference (EUSIPCO), 2015 23rd European*. IEEE. 2015, pp. 634–638 (cit. on pp. 7, 12).
- [18] Shunya Umeki, Masaru Yamashita, and Shoichi 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 (cit. on pp. 7, 12).

- [19] P Mayorga, C Druzgalski, RL Morelos, OH Gonzalez, and J Vidales. "Acoustics based assessment of respiratory diseases using GMM classification." In: *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*. IEEE. 2010, pp. 6312–6316 (cit. on pp. 7, 12).
- [20] Shoichi Matsunaga, Katsuya Yamauchi, Masaru Yamashita, and Sueharu Miyahara. "Classification between normal and abnormal respiratory sounds based on maximum likelihood approach." In: *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE. 2009, pp. 517–520 (cit. on pp. 7, 12).
- [21] Raphael Naves, Bruno HG Barbosa, and Danton D Ferreira. "Classification of lung sounds using higher-order statistics: A divide-and-conquer approach." In: *Computer methods and programs in biomedicine* 129 (2016), pp. 12–20 (cit. on p. 7).
- [22] Sezer Ulukaya, Gorkem Serbes, Ipek Sen, and Yasemin P Kahya. "A lung sound classification system based on the rational dilation wavelet transform." In: *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the IEEE*. 2016, pp. 3745–3748 (cit. on p. 7).
- [23] Kexin Zhang, Xuefeng Wang, Fangfang Han, and Hong Zhao. "The detection of crackles based on mathematical morphology in spectrogram analysis." In: *Technology and Health Care* 23.s2 (2015), S489–S494 (cit. on p. 7).
- [24] Ashok Mondal, Parthasarathi Bhattacharya, and Goutam Saha. "Detection of lungs status using morphological complexities of respiratory sounds." In: *The Scientific World Journal* 2014 (2014) (cit. on p. 7).
- [25] Marco A Tocchetto, Alexandre S Bazanella, L Guimaraes, JL Fragoso, and A Parraga. "An embedded classifier of lung sounds based on the wavelet packet transform and ann." In: *IFAC Proceedings Volumes* 47.3 (2014), pp. 2975–2980 (cit. on p. 7).
- [26] M Yeginer and YP Kahya. "Modeling of pulmonary crackles using wavelet networks." In: *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the IEEE*. 2006, pp. 7560–7563 (cit. on p. 7).
- [27] HG Martinez-Hernandez, CT Aljama-Corrales, R Gonzalez-Camarena, VS Charleston-Villalobos, and G Chi-Lem. "Computerized classification of normal and abnormal lung sounds by multivariate linear autoregressive model." In: *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the IEEE*. 2006, pp. 5999–6002 (cit. on pp. 7, 37).

- [28] Renard Xaviero Adhi Pramono, Syed Anas Imtiaz, and Esther Rodriguez-Villegas. "A Cough-Based Algorithm for Automatic Diagnosis of Pertussis." In: *PloS one* 11.9 (2016), e0162128 (cit. on p. 7).
- [29] Feng Yang, Wee Ser, Jufeng Yu, David Chee-Guan Foo, Daniel Poh Shuan Yeo, Pow-Li Chia, and Jennifer Wong. "Lung water detection using acoustic techniques." In: *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*. IEEE. 2012, pp. 4258–4261 (cit. on pp. 8, 37).
- [30] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." In: *Nature* 521.7553 (2015), pp. 436–444 (cit. on p. 8).
- [31] Li Deng, Dong Yu, et al. "Deep learning: methods and applications." In: *Foundations and Trends® in Signal Processing* 7.3–4 (2014), pp. 197–387 (cit. on p. 8).
- [32] Awni Hannun et al. "Deep Speech: Scaling up end-to-end speech recognition." In: (Dec. 17, 2014) (cit. on p. 8).
- [33] Nicholas D. Lane, Petko Georgiev, and Lorena Qendro. "Deep-Ear: Robust Smartphone Audio Sensing in Unconstrained Acoustic Environments Using Deep Learning." In: *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. UbiComp '15. Osaka, Japan: ACM, 2015, pp. 283–294 (cit. on p. 8).
- [34] Ding Liu, Paris Smaragdis, and Minje Kim. "Experiments on deep learning for speech denoising." In: *Fifteenth Annual Conference of the International Speech Communication Association*. 2014 (cit. on p. 8).
- [35] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. "WaveNet: A Generative Model for Raw Audio." In: (Sept. 12, 2016) (cit. on p. 8).
- [36] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." In: *nature* 521.7553 (2015), p. 436 (cit. on p. 8).
- [37] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. "Why Does Unsupervised Pre-training Help Deep Learning?" In: *J. Mach. Learn. Res.* 11 (Mar. 2010), pp. 625–660 (cit. on p. 8).
- [38] Yoshua Bengio. "Deep learning of representations for unsupervised and transfer learning." In: *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*. 2012, pp. 17–36 (cit. on p. 8).

- [39] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. “Decaf: A deep convolutional activation feature for generic visual recognition.” In: *International conference on machine learning*. 2014, pp. 647–655 (cit. on p. 8).
- [40] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. “Domain adaptation for large-scale sentiment classification: A deep learning approach.” In: *Proceedings of the 28th international conference on machine learning (ICML-11)*. 2011, pp. 513–520 (cit. on p. 8).
- [41] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “Bert: Pre-training of deep bidirectional transformers for language understanding.” In: *arXiv preprint arXiv:1810.04805* (2018) (cit. on p. 8).
- [42] Naoya Takahashi, Michael Gygli, Beat Pfister, and Luc Van Gool. “Deep Convolutional Neural Networks and Data Augmentation for Acoustic Event Detection.” In: *CoRR* (2016). arXiv: [1604.07160](#) (cit. on p. 8).
- [43] Emre Cakir, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen. “Polyphonic sound event detection using multi label deep neural networks.” In: *2015 International Joint Conference on Neural Networks (IJCNN)* (2015), pp. 1–7 (cit. on p. 8).
- [44] Galina Lavrentyeva, Sergey Novoselov, Egor Malykh, Alexander Kozlov, Oleg Kudashev, and Vadim Shchemelinin. “Audio Replay Attack Detection with Deep Learning Frameworks.” In: *INTERSPEECH*. 2017 (cit. on p. 8).
- [45] Philippe Hamel and Douglas Eck. “Learning features from music audio with deep belief networks.” In: *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*. Utrecht, The Netherlands, 2010, pp. 339–344 (cit. on p. 8).
- [46] Jing Huang and Brian Kingsbury. “Audio-visual deep learning for noise robust speech recognition.” In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (2013), pp. 7596–7599 (cit. on p. 8).
- [47] Elias Sprengel, Martin Jaggi, Yannic Kilcher, and Thomas Hofmann. “Audio Based Bird Species Identification using Deep Learning Techniques.” In: *CLEF*. 2016 (cit. on p. 8).
- [48] Honglak Lee, Peter Pham, Yan Largman, and Andrew Y Ng. “Unsupervised feature learning for audio classification using convolutional deep belief networks.” In: *Advances in neural information processing systems*. 2009, pp. 1096–1104 (cit. on p. 8).

- [49] Victor Bisot, Romain Serizel, Slim Essid, and Gaël Richard. “Feature Learning with Matrix Factorization Applied to Acoustic Scene Classification.” In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.6 (2017), pp. 1216–1229 (cit. on pp. 9, 57).
- [50] Daniel D Lee and H Sebastian Seung. “Learning the parts of objects by non-negative matrix factorization.” In: *Nature* 401.6755 (1999), p. 788 (cit. on pp. 9, 39, 40).
- [51] D Seung and L Lee. “Algorithms for non-negative matrix factorization.” In: *Advances in neural information processing systems* 13 (2001), pp. 556–562 (cit. on p. 9).
- [52] Nasser Mohammadiha, Paris Smaragdis, and Arne Leijon. “Supervised and unsupervised speech enhancement using non-negative matrix factorization.” In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.10 (2013), pp. 2140–2151 (cit. on p. 9).
- [53] ChingShun Lin and Erwin Hasting. “Blind source separation of heart and lung sounds based on nonnegative matrix factorization.” In: *Intelligent Signal Processing and Communications Systems (ISPACS), 2013 International Symposium on*. IEEE. 2013, pp. 731–736 (cit. on p. 9).
- [54] J. Makhoul. “Linear prediction: A tutorial review.” In: *Proceedings of the IEEE* 63.4 (Apr. 1975), pp. 561–580 (cit. on p. 11).
- [55] Jiang Hai and Er Meng Joo. “Improved linear predictive coding method for speech recognition.” In: *Proc. Joint and the Fourth Pacific Rim Conf Fourth Int Conf. Information, Communications and Signal Processing Multimedia*. Vol. 3. Dec. 2003, 1614–1618 vol.3 (cit. on p. 11).
- [56] G. White and R. Neely. “Speech recognition experiments with linear predication, bandpass filtering, and dynamic programming.” In: *and Signal Processing IEEE Transactions on Acoustics, Speech* 24.2 (Apr. 1976), pp. 183–188 (cit. on p. 11).
- [57] R. Christiansen and C. Rushforth. “Detecting and locating key words in continuous speech using linear predictive coding.” In: *and Signal Processing IEEE Transactions on Acoustics, Speech* 25.5 (Oct. 1977), pp. 361–367 (cit. on p. 11).
- [58] Tom Bäckström and Carlo Magi. “Properties of line spectrum pair polynomials—A review.” In: *Signal Processing* 86.11 (Nov. 2006), pp. 3286–3298 (cit. on p. 11).
- [59] K. K. Paliwal. “On the use of line spectral frequency parameters for speech recognition.” In: *Digital Signal Processing*. 1992, pp. 80–87 (cit. on p. 11).

- [60] E. Bozkurt, E. Erzin, C. E. Erdem, and A. T. Erdem. "Use of Line Spectral Frequencies for Emotion Recognition from Speech." In: *Proc. 20th Int. Conf. Pattern Recognition*. Aug. 2010, pp. 3708–3711 (cit. on p. 11).
- [61] S. Davis and P. Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences." In: *and Signal Processing IEEE Transactions on Acoustics, Speech* 28.4 (Aug. 1980), pp. 357–366 (cit. on p. 12).
- [62] Brian C. J. Moore, Brian R. Glasberg, and Thomas Baer. "A Model for the Prediction of Thresholds, Loudness, and Partial Loudness." In: *J. Audio Eng. Soc* 45.4 (1997), pp. 224–240 (cit. on p. 12).
- [63] Guodong Guo and Stan Z Li. "Content-based audio classification and retrieval by support vector machines." In: *IEEE transactions on Neural Networks* 14.1 (2003), pp. 209–215 (cit. on p. 12).
- [64] Martin McKinney and Jeroen Breebaart. "Features for audio and music classification." In: (2003) (cit. on p. 12).
- [65] Jeroen Breebaart and Martin F McKinney. "Features for audio classification." In: *Algorithms in Ambient Intelligence*. Springer, 2004, pp. 113–129 (cit. on p. 12).
- [66] Tong Zhang and C-C Jay Kuo. "Audio content analysis for on-line audiovisual data segmentation and classification." In: *IEEE Transactions on speech and audio processing* 9.4 (2001), pp. 441–457 (cit. on p. 12).
- [67] Juan José Burred and Alexander Lerch. "Hierarchical automatic audio signal classification." In: *Journal of the Audio Engineering Society* 52.7/8 (2004), pp. 724–739 (cit. on p. 12).
- [68] C. Joder, S. Essid, and G. Richard. "Temporal Integration for Audio Classification With Application to Musical Instrument Classification." In: *and Language Processing IEEE Transactions on Audio, Speech* 17.1 (Jan. 2009), pp. 174–186 (cit. on pp. 12, 13, 15).
- [69] L. Gerosa, G. Valenzise, M. Tagliasacchi, F. Antonacci, and A. Sarti. "Scream and gunshot detection in noisy environments." In: *Proc. 15th European Signal Processing Conf.* Sept. 2007, pp. 1216–1220 (cit. on pp. 12, 13).
- [70] A. Gray and J. Markel. "A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis." In: *and Signal Processing IEEE Transactions on Acoustics, Speech* 22.3 (June 1974), pp. 207–217 (cit. on p. 12).

- [71] J. Herre, E. Allamanche, and O. Hellmuth. "Robust matching of audio signals using spectral flatness features." In: *Proc. IEEE Workshop the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*. Oct. 2001, pp. 127–130 (cit. on p. 12).
- [72] S. Lim, J. Lee, S. Jang, S. Lee, and M. Y. Kim. "Music-genre classification system based on spectro-temporal features and feature selection." In: *IEEE Transactions on Consumer Electronics* 58.4 (Nov. 2012), pp. 1262–1268 (cit. on pp. 12, 15).
- [73] Dalwon Jang, Minh Jin, and C. D. Yoo. "Music genre classification using novel features and a weighted voting method." In: *Proc. IEEE Int. Conf. Multimedia and Expo*. June 2008, pp. 1377–1380 (cit. on p. 12).
- [74] Dalwon Jang and C. D. Yoo. "Music information retrieval using novel features and a weighted voting method." In: *Proc. IEEE Int. Symp. Industrial Electronics*. July 2009, pp. 1341–1346 (cit. on p. 12).
- [75] E. Scheirer and M. Slaney. "Construction and evaluation of a robust multifeature speech/music discriminator." In: *Proc. and Signal Processing 1997 IEEE Int. Conf. Acoustics, Speech*. Vol. 2. Apr. 1997, 1331–1334 vol.2 (cit. on p. 13).
- [76] S. O. Sadjadi and J. H. L. Hansen. "Unsupervised Speech Activity Detection Using Voicing Measures and Perceptual Spectral Flux." In: *IEEE Signal Processing Letters* 20.3 (Mar. 2013), pp. 197–200 (cit. on p. 13).
- [77] Theodoros Giannakopoulos, Dimitrios I. Kosmopoulos, Andreas Aristidou, and Sergios Theodoridis. "Violence Content Classification Using Audio Features." In: *Advances in Artificial Intelligence, 4th Hellenic Conference on AI, SETN 2006, Heraklion, Crete, Greece, May 18-20, 2006, Proceedings*. 2006, pp. 502–507 (cit. on p. 13).
- [78] O. Gillet and G. Richard. "Automatic transcription of drum loops." In: *Proc. and Signal Processing 2004 IEEE Int. Conf. Acoustics, Speech*. Vol. 4. May 2004, p. iv (cit. on p. 13).
- [79] Fredrik Vogel, Sverre Holm, and Ole Christian Lingjærde. *SPECTRAL MOMENTS AND TIME DOMAIN REPRESENTATION OF PHOTOACOUSTIC SIGNALS USED FOR DETECTION OF CRUDE OIL IN PRODUCED WATER* (cit. on p. 13).
- [80] G. Yu and J. Slotine. "Audio classification from time-frequency texture." In: *Proc. Speech and Signal Processing 2009 IEEE Int. Conf. Acoustics*. Apr. 2009, pp. 1677–1680 (cit. on p. 13).
- [81] J. Saunders. "Real-time discrimination of broadcast speech/-music." In: *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*. Vol. 2. May 1996, 993–996 vol. 2 (cit. on p. 13).

- [82] B. Kedem. "Spectral analysis and discrimination by zero-crossings." In: *Proceedings of the IEEE* 74.11 (Nov. 1986), pp. 1477–1493 (cit. on p. 13).
- [83] P. Kathirvel, M. Sabarimalai Manikandan, S. Senthilkumar, and K. P. Soman. "Noise robust zerocrossing rate computation for audio signal classification." In: *Proc. 3rd Int. Conf. Trendz in Information Sciences Computing (TISC2011)*. Dec. 2011, pp. 65–69 (cit. on p. 13).
- [84] Isabelle Guyon and André Elisseeff. "An introduction to variable and feature selection." In: *Journal of machine learning research* 3.Mar (2003), pp. 1157–1182 (cit. on pp. 14, 26).
- [85] Anil Jain and Douglas Zongker. "Feature selection: Evaluation, application, and small sample performance." In: *IEEE transactions on pattern analysis and machine intelligence* 19.2 (1997), pp. 153–158 (cit. on p. 14).
- [86] Jason Weston, Sayan Mukherjee, Olivier Chapelle, Massimiliano Pontil, Tomaso Poggio, and Vladimir Vapnik. "Feature selection for SVMs." In: *Advances in neural information processing systems*. 2001, pp. 668–674 (cit. on p. 14).
- [87] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. "Gene selection for cancer classification using support vector machines." In: *Machine learning* 46.1 (2002), pp. 389–422 (cit. on p. 14).
- [88] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001 (cit. on pp. 14, 15, 47).
- [89] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002 (cit. on p. 14).
- [90] J. A. Hartigan and M. A. Wong. "Algorithm AS 136: A K-Means Clustering Algorithm." In: *Applied Statistics* 28.1 (1979), p. 100 (cit. on p. 15).
- [91] David Arthur and Sergei Vassilvitskii. "K-means++: The Advantages of Careful Seeding." In: *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '07. New Orleans, Louisiana: Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035 (cit. on p. 15).
- [92] P Dhanalakshmi, S Palanivel, and Vennila Ramalingam. "Classification of audio signals using SVM and RBFNN." In: *Expert systems with applications* 36.3 (2009), pp. 6069–6075 (cit. on p. 15).
- [93] Lei Chen, Sule Gunduz, and M Tamer Ozsü. "Mixed type audio classification with support vector machine." In: *Multimedia and Expo, 2006 IEEE International Conference on*. IEEE. 2006, pp. 781–784 (cit. on p. 15).

- [94] Chien-Chang Lin, Shi-Huang Chen, Trieu-Kien Truong, and Yukon Chang. "Audio classification and categorization based on wavelets and support vector machine." In: *IEEE Transactions on Speech and Audio Processing* 13.5 (2005), pp. 644–651 (cit. on p. 15).
- [95] Lie Lu, Stan Z Li, and Hong-Jiang Zhang. "Content-based audio segmentation using support vector machines." In: *Proc. ICME*. Vol. 1. 2001, pp. 749–752 (cit. on p. 15).
- [96] Lie Lu, Hong-Jiang Zhang, and Stan Z Li. "Content-based audio classification and segmentation by using support vector machines." In: *Multimedia systems* 8.6 (2003), pp. 482–492 (cit. on p. 15).
- [97] Jia-Ching Wang, Jhing-Fa Wang, Cai-Bei Lin, Kun-Ting Jian, and W Kuok. "Content-based audio classification using support vector machines and independent component analysis." In: *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*. Vol. 4. IEEE. 2006, pp. 157–160 (cit. on p. 15).
- [98] Eugene S. Edgington. "Approximate Randomization Tests." In: *The Journal of Psychology* 72.2 (July 1969), pp. 143–149 (cit. on p. 19).
- [99] Benoit Mathieu, Slim Essid, Thomas Fillon, Jacques Prado, and Gaël Richard. "YAAFE, an Easy to Use and Efficient Audio Feature Extraction Software." In: *ISMIR*. 2010, pp. 441–446 (cit. on p. 25).
- [100] Zhang Xuegong. "Introduction to statistical learning theory and support vector machines." In: *Acta Automatica Sinica* 26.1 (2000), pp. 32–42 (cit. on p. 26).
- [101] Gabriel Rilling, Patrick Flandrin, and Paulo Gonçalves. "On empirical mode decomposition and its algorithms." In: *Proceedings of IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing NSIP-03*. Grado, Italy, June 2003 (cit. on pp. 26, 28).
- [102] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu. "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis." In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 454.1971 (Mar. 1998), pp. 903–995 (cit. on p. 28).
- [103] P. Flandrin, G. Rilling, and P. Goncalves. "Empirical Mode Decomposition as a Filter Bank." In: *IEEE Signal Processing Letters* 11.2 (Feb. 2004), pp. 112–114 (cit. on p. 28).
- [104] Semra İçer and Şerife Gengeç. "Classification and analysis of non-stationary characteristics of crackle and rhonchus lung adventitious sounds." In: *Digital Signal Processing* 28 (2014), pp. 18–27 (cit. on p. 32).

- [105] Kevin W Wilson, Bhiksha Raj, Paris Smaragdis, and Ajay Divakaran. "Speech denoising using nonnegative matrix factorization with priors." In: *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE. 2008, pp. 4029–4032 (cit. on p. 41).
- [106] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu. "Non-negative matrix factorization with the Itakura-Saito divergence: With application to music analysis." In: *Neural computation* 21.3 (2009), pp. 793–830 (cit. on pp. 44, 69).
- [107] Artur J Ferreira and Mário AT Figueiredo. "Boosting algorithms: A review of methods, theory, and applications." In: *Ensemble machine learning*. Springer, 2012, pp. 35–85 (cit. on p. 47).
- [108] Chun-Nam John Yu and Thorsten Joachims. "Learning structural svms with latent variables." In: *Proceedings of the 26th annual international conference on machine learning*. ACM. 2009, pp. 1169–1176 (cit. on p. 51).
- [109] Vicky Kumar Verma and N. Khanna. "Indian language identification using k-means clustering and support vector machine (SVM)." In: *Proc. Students Conf. Engineering and Systems (SCES)*. Apr. 2013, pp. 1–5 (cit. on p. 57).
- [110] G. Vyas and M. K. Dutta. "Automatic mood detection of indian music using mfccs and k-means algorithm." In: *Proc. Seventh Int. Conf. Contemporary Computing (IC3)*. Aug. 2014, pp. 117–122 (cit. on p. 57).
- [111] E. Tsunoo, G. Tzanetakis, N. Ono, and S. Sagayama. "Audio genre classification using percussive pattern clustering combined with timbral features." In: *Proc. IEEE Int. Conf. Multimedia and Expo*. June 2009, pp. 382–385 (cit. on p. 57).
- [112] Sebastian Krey, Uwe Ligges, and Friedrich Leisch. "Music and timbre segmentation by recursive constrained K-means clustering." In: *Computational Statistics* 29.1-2 (2014), pp. 37–50 (cit. on p. 57).
- [113] A. Mesaros, T. Heittola, and T. Virtanen. "TUT database for acoustic scene classification and sound event detection." In: *Proc. 24th European Signal Processing Conf. (EUSIPCO)*. 2016 (cit. on p. 61).
- [114] Victor Bisot, Romain Serizel, Slim ESSID, and Gael Richard. "Supervised nonnegative matrix factorization for acoustic scene classification." In: *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE)* (2016) (cit. on pp. 62, 69).

- [115] Marcelo A Colominas, Gastón Schlotthauer, and María E Torres. “Improved complete ensemble EMD: A suitable tool for biomedical signal processing.” In: *Biomedical Signal Processing and Control* 14 (2014), pp. 19–29 (cit. on p. 68).
- [116] Paris Smaragdis, Cedric Fevotte, Gautham J Mysore, Nasser Mohammadiha, and Matthew Hoffman. “Static and dynamic source separation using nonnegative factorizations: A unified view.” In: *IEEE Signal Processing Magazine* 31.3 (2014), pp. 66–75 (cit. on pp. 69, 70).
- [117] Long Leo Zhu, Yuanhao Chen, Alan Yuille, and William Freeman. “Latent hierarchical structural learning for object detection.” In: (2010) (cit. on p. 70).
- [118] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. “Object detection with discriminatively trained part-based models.” In: *IEEE transactions on pattern analysis and machine intelligence* 32.9 (2010), pp. 1627–1645 (cit. on p. 70).
- [119] N. Mohammadiha and A. Leijon. “Nonnegative HMM for Babble Noise Derived From Speech HMM: Application to Speech Enhancement.” In: *and Language Processing IEEE Transactions on Audio, Speech* 21.5 (May 2013), pp. 998–1011 (cit. on p. 70).
- [120] Gautham J. Mysore, Paris Smaragdis, and Bhiksha Raj. “Non-negative Hidden Markov Modeling of Audio with Application to Source Separation.” In: *Latent Variable Analysis and Signal Separation*. Springer Berlin Heidelberg, 2010, pp. 140–148 (cit. on p. 70).
- [121] G. J. Mysore and P. Smaragdis. “A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics.” In: *Proc. Speech and Signal Processing (ICASSP) 2011 IEEE Int. Conf. Acoustics*. May 2011, pp. 17–20 (cit. on p. 70).
- [122] G. Peeters. *A large set of audio features for sound description (similarity and classification) in the CUIDADO project*. Tech. Rep. IRCAM, 2004 (cit. on pp. 74, 75).