A Data Mining Approach for Acoustic Diagnosis of Cardiopulmonary Disease

by

Bryan C. Flietstra

B.S. Operations Research
United States Air Force Academy, 2006

SUBMITTED TO THE SLOAN SCHOOL OF MANAGEMENT IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN OPERATIONS RESEARCH
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2008

Signature of Author:
Sloan School of Management
Interdepartmental Program in Operations Research
May 15th, 2008

Approved by:
Dr. Natasha Markuzon
The Charles Stark Draper Laboratory, Inc.
Technical Supervisor

Certified by:
Professor Roy Welsch
Professor of Management Science, Statistics, and Engineering Systems
Thesis Advisor

Accepted by:
Professor Cynthia Barnhart
Professor, Civil and Environmental Engineering
Co-Director, Operations Research Center

[This Page Intentionally Left Blank]

A Data Mining Approach for Acoustic Diagnosis of Cardiopulmonary Disease

by

Bryan C. Flietstra

Submitted to the Sloan School of Management on

May 15th, 2008 in partial fulfillment of the requirements for the

Degree of Master of Science in Operations Research

Abstract

Variations in training and individual doctor's listening skills make diagnosing a patient via stethoscope based auscultation problematic. Doctors have now turned to more advanced devices such as x-rays and computed tomography (CT) scans to make diagnoses. However, recent advances in lung sound analysis techniques allow for the auscultation to be performed with an array of microphones, which send the lung sounds to a computer for processing. The computer automatically identifies adventitious sounds using time expanded waveform analysis and allows for a more precise auscultation.

We investigate three data mining techniques in order to diagnose a patient based solely on the sounds heard within the chest by a "smart" stethoscope. We achieve excellent recognition performance by using *k* nearest neighbors, neural networks, and support vector machines to make classifications in pair-wise comparisons. We also extend the research to a multi-class scenario and are able to separate patients with interstitial pulmonary fibrosis with 80% accuracy. Adding clinical data also improves recognition performance. Our results show that performing computerized lung auscultation offers a low-cost, non-invasive diagnostic procedure that gives doctors better clinical utility especially in situations when x-rays and CT scans are not available.

Technical Supervisor: Natasha Markuzon
The Charles Stark Draper Laboratory, Inc.

Thesis Advisor: Professor Roy Welsch
Professor of Management Science, Statistics, and Engineering Systems
Massachusetts Institute of Technology

[This Page Intentionally Left Blank]

Acknowledgements

There have been many people who have provided me with help, encouragement, or advice throughout the writing of this thesis. I would like to thank the many individuals for their support.

First, I thank God for blessing me with the opportunity to continue my studies here. He has blessed me with countless skills and abilities and for that I am very grateful. Isaiah 40:31.

Second, I would like to thank Dr. Natasha Markuzon of Draper Laboratory. I am very grateful for all that you have done for me as far as this project. Your advice and insights have been immeasurable. Болшое спасибо!

I would also like to thank Dr. Roy Welsch of MIT. Thank you for making time for me out of your busy schedule. Without your advice and ideas, I would not have been able to complete this thesis.

Furthermore, I would like to thank Dr. Andrey Vyshedskiy and Dr. Ray Murphy from Stethographics. From the first meetings over sake and sushi to the completion of this thesis and all the frustrations along the way, you have provided me with a very interesting and fulfilling research topic. I am truly thankful for being involved with this cutting edge and very practical research. Dr. Murphy - I hope you keep pursuing this dream and make it a reality.

Thank you to my study group at the ORC: Clay, Mo, Charles, and Chris. Thank you for all the help on problem sets and encouragement when I'd get frustrated. You've made this a great two years and without your help, I wouldn't have learned nearly as much.

To all my friends: Brian, Joe, Mr. Kayser, Anthony, Eric, Doug and all the other Air Force lieutenants in the area. Thank you for helping me keep my sanity and providing me with plenty of opportunities to do something other than homework or research. I have plenty of great memories from spending time with you all in Boston.

Finally, I'd like to thank my family for their constant support. Thank you for all your prayers and advice. You'd always be there for me and encourage me when things would go poorly. Thank you also for the occasional care package. Your chocolate chip cookies are amazing!

This thesis was prepared at The Charles Stark Draper Laboratory, Inc., under Internal Company Research Project 21796-001, Data Mining for Draper.

Publication of this thesis does not constitute approval by Draper or the sponsoring agency of the findings or conclusions contained herein. It is published for the exchange and stimulation of ideas.

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or The U.S. Government.

Bryan C. Flietstra, 2nd Lt., USAF　　May 15, 2008

[This Page Intentionally Left Blank]

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Advances in research methods and new technologies make medical practice a very dynamic field. Almost daily, various medical researchers publish results of a current medical study claiming a new drug lowers cholesterol, a new treatment option for cancer, etc. Sometimes a new technology is presented for use as a diagnostic tool. In this thesis, we present a "smart" stethoscope that will improve the diagnosis of cardiopulmonary disorders. The "smart" stethoscope will contain an embedded chip and rely on microphones and computerized algorithms to make an instant diagnosis. This technology will impact the way medicine is practiced, especially in remote areas where expensive and bulky devices such as chest x-rays and computed tomography (CT) scans are not available.

A similar impact occurred in 1816 when Laennec introduced the stethoscope. Instead of diagnosing patients largely based upon external symptoms, for the first time, a doctor was able to perform lung auscultation and effectively listen to internal chest sounds.

Over time, the design of the stethoscope has been improved upon and nearly everyone in the medical practice uses one. It has been at the forefront for performing auscultation for generations and has practically become a symbol of the medical profession. Figure 1.1 illustrates the enhancements made to the stethoscope and their inventors [1].

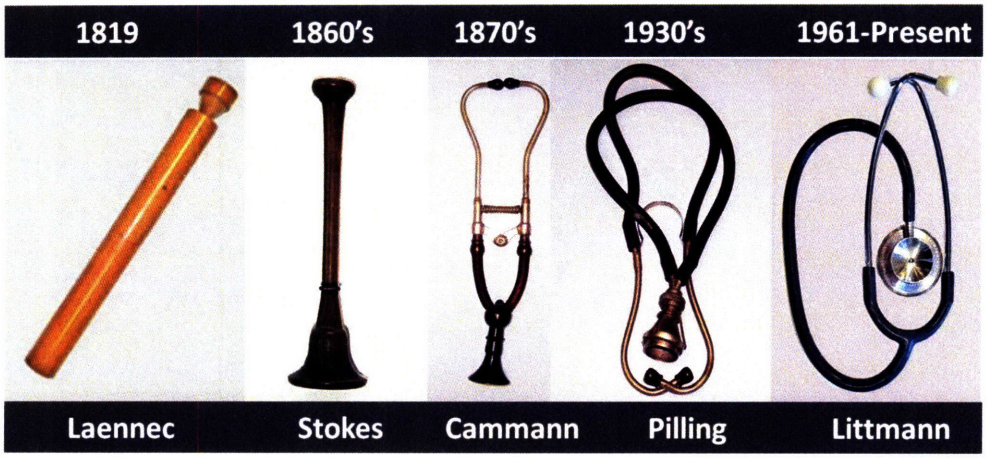


Figure 1.1 Evolution of the Stethoscope

Still, in spite of the improvement of the stethoscope, other medical diagnostic tools have also been introduced and seem to have partially phased out the stethoscope as the primary diagnostic tool. One of the biggest advances was the invention of the chest X-ray. Proponents of the X-ray cited its reliability as a primary benefit over auscultation using a stethoscope. Lung auscultation via a stethoscope principally relies on the doctor's ear, skill, and training. The differences in these three factors can cause great variability and diminish the clinical utility of the stethoscope. In spite of this, the stethoscope remains widely used as an initial diagnostic tool. A doctor will listen for sounds at several locations and then make a recommendation for a more thorough, objective test. These tests include chest X-rays, CT scans, magnetic resonance imaging (MRI), blood tests, spirometry, pulmonary arteriography, lung biopsy, and radioisotope scanning techniques [2].

Still, these tests can be very expensive to perform. New medical technology has been the primary cause for the rising health care costs and insurance premiums. There are two ways to combat these rising costs. First, doctors should be judicious as far as which tests to prescribe to which patients. Smartly applying various diagnostic tests to patients with certain symptoms can reduce the total costs [3]. For instance, not every patient requires a CT scan, so only perform the test on those where the most benefit can be gained by ordering the test. The second way to combat rising health care costs is to develop more cost effective treatments. A cheap test that can easily be read by a technician is of utmost importance in reducing health care costs. Many of the aforementioned tests cost health care providers nearly a thousand dollars per trial [3]. Not only are some of the currently practiced tests monetarily expensive, but they are also expensive from a waiting time standpoint. Oftentimes patients will have to wait in a queue for access to the specialized equipment. Even after the testing, more waiting can occur. It takes a fair amount of time for doctors and technicians to "read" the results or to wait for a lab to process the sample.

In this thesis, we will present and discuss an emerging technology that will help doctors make better decisions and also to speed up a patient's diagnostic time. We investigate a data mining approach to accurately diagnose patients based on the sounds contained in the chest while breathing. In a sense, we return to the diagnostic properties of Laennec's acoustic stethoscope. This time, instead of a doctor performing the analysis, a "smart" stethoscope will be used. In the "smart" stethoscope, microphones will perform the auscultation and then a computer will be used to analyze the sounds and ultimately make a recommendation for a diagnosis. This diagnostic tool will be interpreted by a doctor to prescribe further tests or begin treatment. Using a computerized approach will eliminate the variability in the doctor's skill and ear and ultimately improve the reliability of a diagnosis.

We expect the smart stethoscope to find application areas in many settings: in physician's offices, hospitals, nursing homes - essentially everywhere the stethoscope is used to listen to hearts and lungs. In addition, new areas of exploitation include settings where doctoral expertise or stationary medical equipment is not always available: diagnostics on tanker ships, oil rigs, embassies, soldiers operating in remote areas, and home monitoring by a visiting nurse. The diagnostic information provided by the "smart" stethoscope can be used on the spot or sent to a doctor for further analysis.

1.1 Thesis Overview

The goal of this research is to develop a decision analysis tool for doctors to use when diagnosing chest and lung disorders. It is based on automated auscultation and can be expanded to include clinical data such as temperature, blood pressure, etc. We aim to show that computerized auscultation is a viable tool that will provide cost effective and non-invasive diagnoses. Here, we provide a chapter overview of the remainder of the thesis.

Chapter 2 – Introduction to Time Expanded Waveform Analysis and Adventitious Lung Sounds

Chapter 2 provides an introduction to the study of lung sounds and common medical practice. We discuss the various types of sounds that can be heard when performing auscultation. They are crackles, wheezes, rhonchi, and squawks. These sounds are the main source s of information in distinguishing between diagnoses in our analysis. Next, we describe the five diseases we distinguish between: pneumonia (PN), congestive heart failure (CHF), interstitial (idiopathic) pulmonary fibrosis (IPF), asthma, and chronic obstructive pulmonary disease (COPD). Also included in the study are asymptomatic patients. We describe the diseases with respect to the adventitious lung sounds and also provide insight as to how doctors make a diagnosis for each. Additionally, we introduce the multi-channel lung sound analyzer used to record the sounds. Finally, we perform a literature review of other computerized auscultation studies.

Chapter 3 – Foundations for Data Mining Analysis

In Chapter 3, we provide an overview of the machine learning methods used in this research. In particular, we focus on supervised learning methods such as neural networks, *k* nearest neighbors, and support vector machines (SVM). Support vector machines are primarily used for binary classification purposes and we describe some other commonly accepted approaches to expand the binary problem to a multiple class scenario. A multi-class scenario is applicable to making a diagnosis from a wide spectrum of diseases.

Chapter 4 – Classifying Lung Sounds

In Chapter 4, we describe how the data we collect using the multi-channel lung sound analyzer is incorporated into the machine learning framework. Most of the features are collected from the auscultation by the “smart” stethoscope. In addition to these, we supplement this data with features that describe the distribution of the sounds around the chest. We also add some clinical features for our analysis. To perform the analysis, we first look at classifying individual crackles (a sound that will be fully described in Chapter 2). We also introduce a voting schema that will be used to increase diagnostic performance and make diagnoses on the patient level. Next we expand the machine learning process to include all adventitious lung sounds by classifying individual breaths. We provide the framework used to conduct the analysis, including determining training and testing sample sizes and various model validation approaches.

Chapter 5 – Results and Discussion

Chapter 5 presents results of lung sound classification. It describes the metrics we use to evaluate trial results. It includes the classification performance of all pair-wise comparisons, multi-class classifications, and gauges the performance of adding clinical information. In addition, several results are highlighted and discussed for their immediate impact on the medical field.

Chapter 6 – Summary, Conclusions, and Future Work

We discuss the overall effectiveness of our present model and propose ideas for future research. Also, we include a long-term vision of potential applications of this technology including remote telemedicine, and in-home patient monitoring.

1.2 Thesis Contributions

This research makes the following contributions:

- Shows that multi-channel lung auscultation is a viable method for medical research.
- Shows that interstitial pulmonary fibrosis crackles are distinguishable from crackles of other diseases using acoustic analysis.
- Demonstrates that most pairs of diseases can be separated based on sounds, including asthma and chronic obstructive pulmonary disease. Pneumonia and congestive heart failure patients can be separated by incorporating acoustic and clinical data.
- Introduces a hybridized approach to data mining that combines data from multiple sources to make a diagnosis.
- Shows that interstitial pulmonary fibrosis and asymptomatic patients can be correctly classified when several diseases are possibilities.

[This Page Intentionally Left Blank]

Chapter 2

Introduction to Time Expanded Waveform Analysis and Adventitious Lung Sounds

In this chapter, we seek to explore the physiology of the sounds heard throughout the chest during the breathing cycle and also provide a description of the diseases that are used in our study. Adventitious lung sounds have been described and useful in diagnostic procedures since the invention of Laennec's stethoscope. Here we describe the sounds known as crackles, wheezes, rhonchi, and squawks in accordance with accepted medical standards. We begin the chapter with an introduction of the computerized lung sound analyzer developed and used by Stethographics to record the patients who participated in this study. Its development is paramount to this study. The last section in the chapter describes some current studies relevant to computerized lung sound auscultation.

2.1 Computerized Auscultation

As advances in medical technology offered new methods to diagnose patients with lung diseases, the use of a stethoscope for auscultation waned in popularity and methods such as the chest X-ray became favored. A preeminent medical researcher in the field of lung sounds even claimed that auscultation had been reduced to a "perfunctory ritual" [4]. A primary cause of auscultation falling out of favor with pulmonologists is the high variability of doctor's listening abilities [5]. With no concrete standards, each physician could essentially hear the sounds differently and as a result possibly misdiagnose a patient. In order to combat this high variability, pioneering researchers began investigating the role of computer based technology in order to objectively measure and visualize the sounds inherent to cardiopulmonary diseases. In a groundbreaking study, Murphy et al. introduced a methodology known as Time Expanded Waveform Analysis (TEWA) [6]. At the time of the journal article, normal lung sounds could not be distinguished from the adventitious or abnormal lung sounds using conventional recorder speeds. Instead, they visualized the waveforms at a much higher frame rate, essentially zooming in on the waveform; thus the name, TEWA. For the first time, adventitious sounds could be visualized. TEWA creates reproducible visual displays that allow for a more objective approach to differentiating features of lung sounds and which also enhances the diagnostic utility of the sounds [6].

One of the first successful applications of TEWA was centered on the detection of an adventitious lung sound known as a crackle in workers with exposure to asbestos. TEWA was able to help doctors define the crackles as well as monitor the patients [7]. TEWA was also useful in setting standard definitions for various lung sounds [8]. One of the next applications of the technology was for the development of an automatic crackle counter. The results of a study comparing methods to detect crackles validated the computerized methodology since the results were highly correlated with doctor's counts [9]. These discoveries led to the development of the multi-channel lung sound analyzer. A full description is given in [2].

The multi-channel lung sound analyzer used in this thesis was developed by Stethographics (STG) and the model is known as STG-1602. The STG-1602 consists of a total of sixteen miniature microphones which are inserted into the chest pieces of stethoscopes. Fourteen of these microphone based stethoscopes are embedded into a soft foam pad and the two additional stethoscopes are placed on the trachea and heart. The foam pad is positioned on a gurney or examination bed with a cover placed over it for sanitary purposes. The patient lies on the pad and several full breath cycles are recorded. An illustration of the pad and a picture of the STG-1602 in use are shown in Figure 2.1.

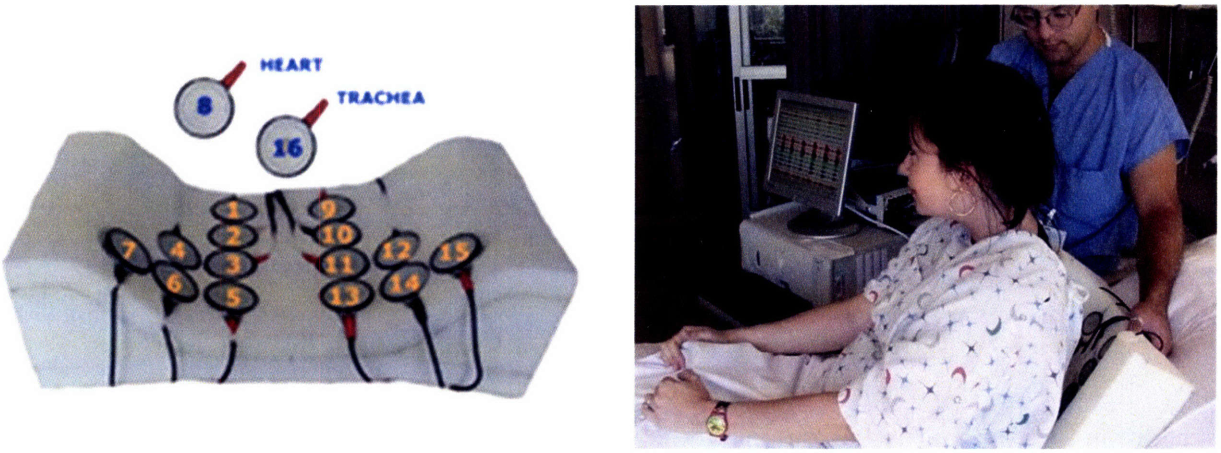


Figure 2.1 Illustration of STG-1602 and a Picture of a Recording

The lung sounds are fed through a signal processing box, an analog to digital converter, and finally into a computer running software specifically designed for this purpose. The software helps aid the diagnosis process in two ways. First, the lung sounds are displayed directly on a computer screen. Visual displays can help doctors notice the adventitious lung sounds in the breathing process. The visual display depicts both the inspiratory and expiratory waveforms for all 16 channels. Furthermore, the site of origin for the sounds is determined and the individual events can be viewed in three dimensions. The site of the individual sounds is found through the arrival times of the sounds at different microphones [10]. Figure 2.2 shows examples of both types of visualizations. Notice the large amount of abnormal activity in Channels 13 – 15 and in the lower left portion of the 3D view. These are adventitious lung sounds, which will be studied in more detail in Section 2.2.

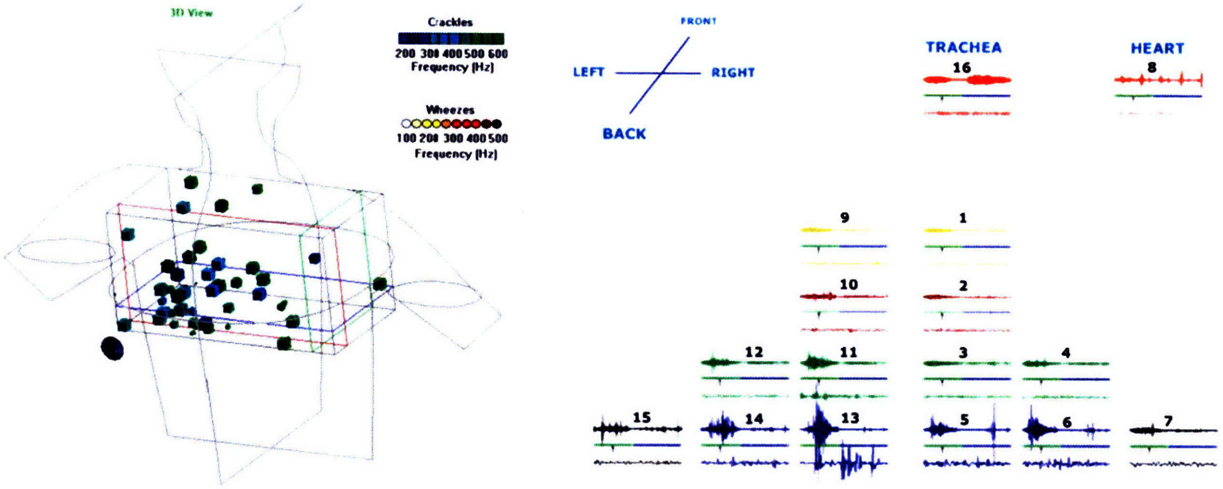


Figure 2.2 3D Visualization and Waveform Visualization for all Channels

In addition to providing visualizations of the waveforms and localizing the origins of the sounds, the software package of the STG-1602 automatically identifies the types of adventitious sounds studied in Section 2.2.

2.2 Adventitious Lung Sounds

Adventitious lung sounds are abnormal sounds that are heard in addition to the typical sounds associated with the breathing process. Their acoustic characteristics appear to be superimposed on the normal background sounds heard within the chest. These sounds can occur during both the inspiratory and expiratory phases of breathing. Furthermore, the sounds can be discontinuous (crackles), or continuous (wheezes, rhonchi). The occurrence of various adventitious lung sounds throughout the breathing cycle typically indicates that a patient has a cardiopulmonary disease. In this section, we fully explain both discontinuous and continuous breath sounds. A good introduction to these sounds is found on the instructional CD [11]. It even contains sample audio files for the sounds mentioned.

2.2.1 Discontinuous Lung Sounds - Crackles

Discontinuous lung sounds are characterized by their short duration and are often very sporadic in nature. The predominant type of sound in this category is known as a crackle or rale. An auditory crackle can be compared to the occasional popping sound made by a campfire.

Although no one can be sure, medical experts believe the crackles are the result of the sudden openings of airways. They may also occur as a result of fluid that is built up in the airways. Furthermore, crackles can be further subdivided and they can be characterized as either being "fine" or "coarse." Again, this distinction is made on the basis of the acoustic characteristics of the event. A fine crackle typically has a high pitch, low amplitude, and duration of less than 10 milliseconds. An analogy to this type of adventitious sound is that of bacon sizzling and popping when it is fried. On the other hand, coarse crackles can have low pitches, higher amplitudes, and normally last longer than 10 milliseconds. These coarse crackles can be compared to the sounds of water being poured out of bottle as described by Laennec. Still, in spite of the acoustic differences between fine and coarse crackles, medical researchers feel that they are generated from the some underlying physiologic causes. An example waveform of a breath containing several crackles is shown in Figure 2.1 [11]. The waveform on the top shows two full breaths. The waveform on the bottom is the time expanded waveform analysis and essentially provides a close up view of the sound.

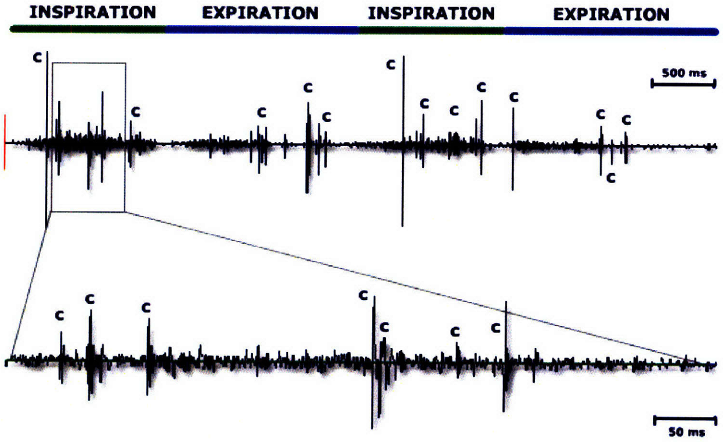


Figure 2.3 Breath Waveform with Crackles Denoted by "C"

When detecting crackles via the multi-channel lung sound analyzer, it is important to note that the sound of a crackle can be heard throughout the chest, which means it gets picked up by multiple channels. Detecting these crackles throughout the chest led to the concept of a *crackle family*. A crackle family is the set of waveforms that correspond to a single event within the chest. As a result, special care needs to be taken in order to consider only the dominant crackle. The dominant crackle is determined by the channel where the crackle has the highest recorded amplitude and this crackle is known as the *mother crackle*. All other recorded crackles that were generated from the same event are known as *daughter crackles* [12].

2.2.2 Continuous Lung Sounds

Continuous long sounds last much longer than the sporadic or explosive crackles mentioned previously. These sounds may last for almost the entirety of the patient's breath. Of these sounds, we look at wheezes, rhonchi, and squawks.

2.2.2.1 Wheezes

A wheeze is one type of continuous adventitious breath sound and sometimes has a musical type tone to it. A typical wheeze lasts for more than 200 milliseconds. Associated with a wheeze are high frequency sinusoidal waveforms and whistling sounds. Wheezes are believed to be caused by narrowing of the airways. A wheeze waveform is shown in Figure 2.2 below [11]. The most commonly associated disease with wheezes is asthma.

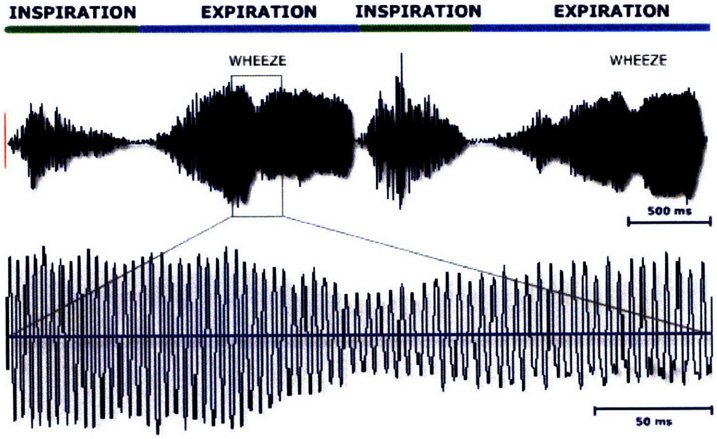


Figure 2.4 Time Expanded Wheeze Waveform

2.2.2.2 Rhonchi

A rhonchus is very similar to a wheeze and is characterized by its very low pitch. It also has a much lower frequency. Rhonchi are frequently caused by airway secretions but sometimes can be caused by a narrowing of the airways. A sample rhonchus is shown in Figure 2.3. Compared to Figure 2.4, it is easy to see the differences in the frequency [11].

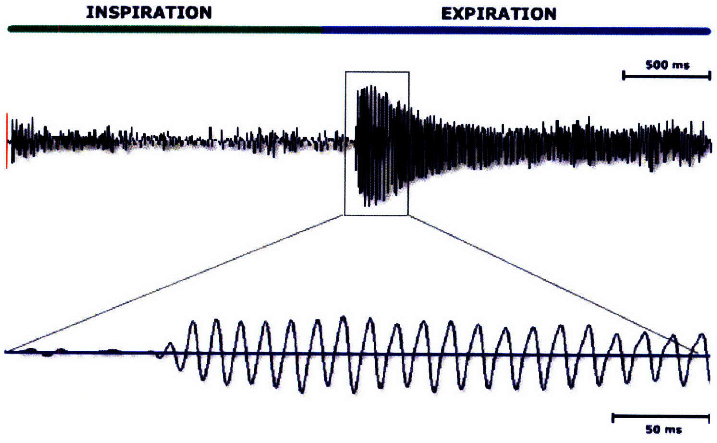


Figure 2.5 Time Expanded Rhonchus Waveform

2.2.2.3 Squawks

A squawk when compared to a wheeze or rhonchus is much shorter in duration, but not to the point of being characterized as a discontinuous adventitious lung sound. They sound like a quick squeak and have a brief sinusoidal waveform. The waveform of a squawk is depicted in Figure 2.4 [11].

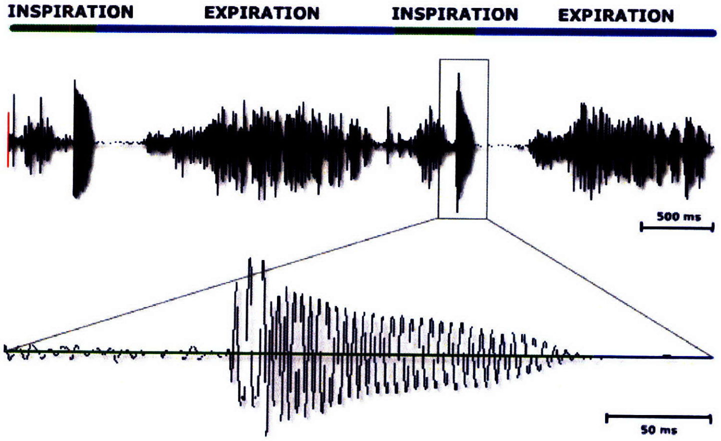


Figure 2.6 Time Expanded Squawk Waveform

2.3 Cardiopulmonary Disease Overview

In this section, we look at each of the five types of diseases that we aim to distinguish. They are pneumonia (PN), congestive heart failure (CHF), asthma, chronic obstructive pulmonary disease (COPD), and idiopathic pulmonary fibrosis (IPF). We also include asymptomatic patients, which are also referred to as normal patients. They are patients without a known cardiopulmonary disorder. In each subsection, we seek to achieve the following goals:

- Provide a brief description of the disease.
- Identify current diagnostic procedures.
- Present a brief overview of potential adventitious lung sounds present in a patient with the specified disease.
- Provide an illustration of the waveforms associated with the specified disease.

All waveform illustrations are taken from [11].

2.3.1 Pneumonia

Pneumonia is an infection of the lung, most commonly caused by bacteria, but also by viruses, fungi, and parasites. The infection causes portions of the lung to fill with fluid. It is the sixth leading cause of death in the United States, and the leading cause of death from infectious disease [13]. Oftentimes, pneumonia develops when a person already has a weakened immune system. Most patients with pneumonia display some sort of respiratory symptoms including a cough and sputum production. Other symptoms that are typically present include fever and increased respiratory rate [13].

Adventitious lung sounds are usually present in a patient with pneumonia. Crackles that occur at the base of the lung are the most common, but other sounds such as wheezes and squawks can be present as well. The crackles tend to be consolidated within the region of the lung containing the infection. Furthermore, the lung sounds may be quieter than normal [11]. In Figure 2.8, the patient has a significant amount of inspiratory crackles which are consolidated at the left base. In the illustration, the numbers correspond to the channel where the waveform was recorded. Also, the small green and blue bar below the waveform depict the inspiration and expiration periods of the breathing phases, respectively. The time expanded waveform is directly below this bar and represents 100 milliseconds.

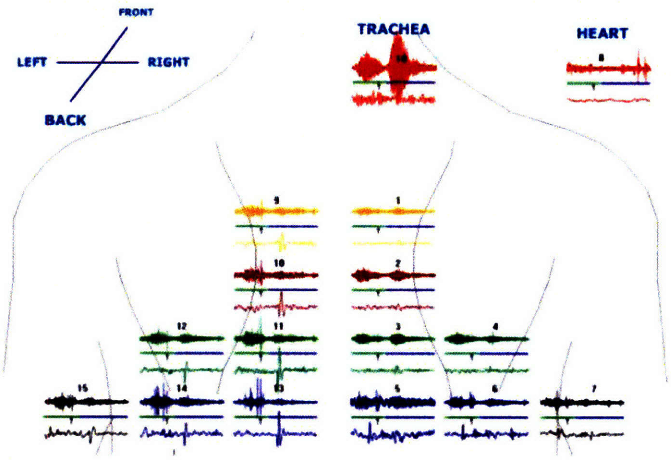


Figure 2.7 Waveform Patterns of a Pneumonia Patient

Typical diagnostic procedures for pneumonia include ordering a chest x-ray, computed tomography (CT) scans, and possibly lab work including blood tests and sputum analysis. A chest x-ray is the most commonly applied diagnostic technique. In a patient with pneumonia, the chest x-ray will possibly show white areas known as infiltrates that indicate an infection [14]. Still, chest x-rays do not remove the problem of observer variability so more thorough tests may be needed. A CT scan is often referred to as the gold standard since it is more sensitive to infiltrates than a simple chest x-ray. Still, a CT scan is only performed if a chest x-ray does not produce results. Although not a standalone diagnostic technique, the analysis of a patient's sputum can help identify the presence of the type of bacteria causing the infection. As a result, an appropriate antibiotic can be prescribed [13].

2.3.2 Congestive Heart Failure

Congestive heart failure is a serious condition in which the heart cannot pump enough blood to the body. It is a chronic, long-term condition. It can develop over time as a result of factors such as high blood pressure, obesity, coronary artery disease, or it can develop suddenly as a result of a heart attack. Many symptoms of heart failure result from the congestion that develops as fluid backs up into the lungs and leaks into the tissues. Some symptoms of CHF include shortness of breath, cough, fatigue, and swelling of the ankles and feet [15].

Although this is primarily a heart condition, adventitious lung sounds still occur because of the buildup of fluid within the lungs. The primary lung sounds heard in CHF are crackles that occur in the bases of the lungs. Unlike localized crackles in pneumonia, the crackles that occur in CHF tend to be more symmetric and occur in both lungs simultaneously. Crackles that occur higher in the chest may indicate increasing severity of the illness. Some wheezes and rhonchi can also occur particularly in the late expiratory phase of breathing [11]. In Figure 2.9, crackles can be seen occurring in the bottom of both lungs.

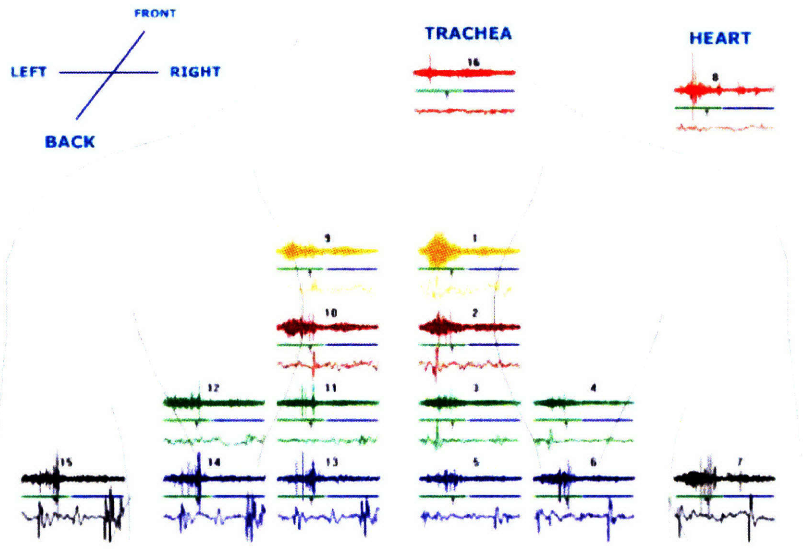


Figure 2.8 Waveform Patterns of a Congestive Heart Failure Patient

A typical starting point for diagnosing CHF is taking a patient history and performing a physical examination. A physical may reveal swelling of the legs and ankles which is a good indication of CHF. If a physical does not produce a useful diagnosis, an echocardiogram may be performed. It is an effective but expensive diagnostic procedure. It is an ultrasound that can reveal the size and the performance of the various chambers of the heart. Doctors can use the results from the echocardiogram to measure the amount of blood pumped to the body in each heartbeat [15]. It also can reveal other cardiac abnormalities which can be pertinent in a diagnosis. Also, chest x-rays have some utility in determining the size and function of the heart.

2.3.3 Chronic Obstructive Pulmonary Disease

Chronic obstructive pulmonary disease is defined as a "disease state characterized by airflow limitation that is not fully reversible" [16]. COPD encompasses a class of diseases that are closely related to one another. It is the fourth leading cause of death in the United States. The two dominant diseases that make up the class are chronic bronchitis and emphysema. Chronic bronchitis is defined as chronic, sputum producing cough that lasts for more than three months of the year for two consecutive years. Emphysema destroys the alveoli, the place within the lungs where the exchange of oxygen and carbon dioxide occurs [17]. Both chronic bronchitis and emphysema can be caused by smoking. Because of their common cause, the diseases often occur together and the diagnosis and treatment options are very similar. Other symptoms of COPD include decreased lung function, shortness of breath, wheezing, and experiencing difficulty exhaling [17].

Lung sounds present in patients with COPD normally include wheezing as the expiratory phase comes to an end. Also common are rhonchi, but they generally clear after coughing. A patient also may have a few crackles along the base of the lungs. Furthermore, the basic breath sounds are decreased in intensity [11]. The patient depicted in Figure 2.10 has low intensity lung sounds as well as several crackles focused at the bases of the lungs. Unfortunately for the purposes of illustration, this patient does not have any wheezing present.

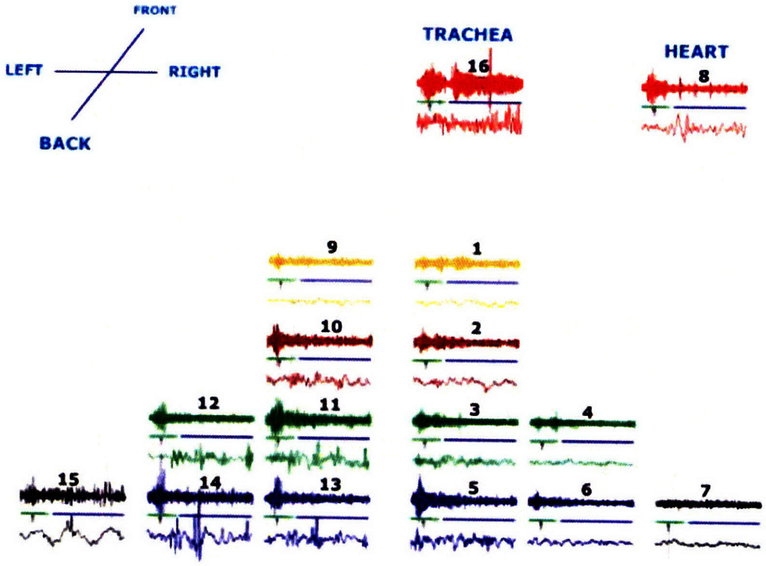


Figure 2.9 Waveform Patterns of a COPD Patient

Diagnosing COPD can be a difficult task since it is often mistaken as asthma since the symptoms are very similar. A key step to diagnosing COPD is obtaining a patient's medical and personal history. COPD should be suspected in any patient over age 50 with a history of smoking [17]. A more precise diagnosis can be made using pulmonary function tests known as spirometry. These tests measure the airflow obstruction in the lungs when the patient breathes out. It measures the maximum volume and the force of the air as it is exhaled from the lungs.

Lower flow rates are observed in a patient with COPD. A chest x-ray is typically obtained in addition to spirometry in order to distinguish COPD from CHF [16].

2.3.4 Asthma

Asthma is one of the most prevalent chronic diseases in the United States, affecting 57 per 1,000 persons. The National Institute of Health defines asthma as "a chronic inflammatory disorder of the airways that causes signs and symptoms of airflow limitation, wheezing, breathlessness, and cough" [18]. Notice that these are practically the same symptoms as those pointed out for COPD in Section 2.3.4. One of the main distinguishing characteristics between the two diseases is that of the patient history. Oftentimes, asthma can be diagnosed at a young age whereas COPD tends to develop as a result of frequent smoking over the course of one's life.

The dominant adventitious lung sound in asthma is wheezing. In spite of this, wheezing may only be present during an asthma attack. If a patient's asthma isn't active, the patient may appear asymptomatic. For this reason, a lot of importance is placed on factors other than just the wheezing. It is important to note that not all wheezing is asthma related. Figure 2.11 shows a patient with wheezing and decreased lung sounds. Wheezing is present in both inspiration and expiration.

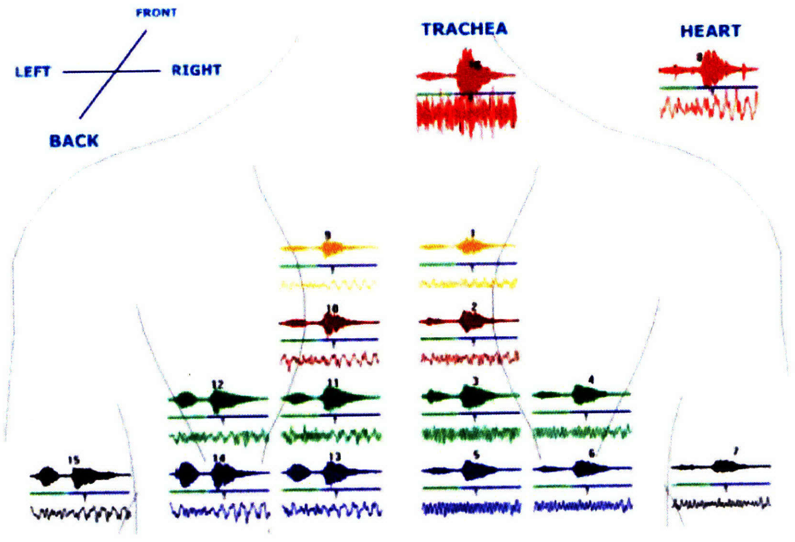


Figure 2.10 Waveform Pattern in an Asthma Patient

Diagnosing asthma is quite complicated because the definition of it is so broad. In order to make a diagnosis of asthma, three criteria are important:

1. Symptoms consistent with airflow limitation
2. Airflow limitation is partially reversible with an inhaler
3. Other diseases are excluded as possibilities.

The answers to these three criteria can be obtained through patient history, physical examination, and spirometry [18].

2.3.5 Idiopathic Pulmonary Fibrosis

Idiopathic pulmonary fibrosis (IPF) is a disease characterized by the scarring and thickening of the lungs. It is called idiopathic since there is no known cause for the disease. The disease most often occurs in the elderly and does not have a favorable prognosis. No cure exists and treatment does not help too often. It is a fairly rare disease that may cause problems when attempting to diagnose a patient since doctors do not see the disease that frequently. Some common symptoms include chest pain, shortness of breath, and a dry cough [19].

Fine crackles at the bases of the lungs are the most common adventitious lung sound. They tend to occur towards the end of the inspiratory portion of the breath cycle. As the patient's condition worsens, more crackles are noticed throughout the entire chest. When the crackles begin to be heard in the expiratory phase, it is another sign of a worsening condition. Squawks are also heard occasionally in patients with IPF [11]. Since the lung sounds closely resemble that of CHF, it is commonly misdiagnosed as such. A patient with many inspiratory crackles is shown in Figure 2.11.

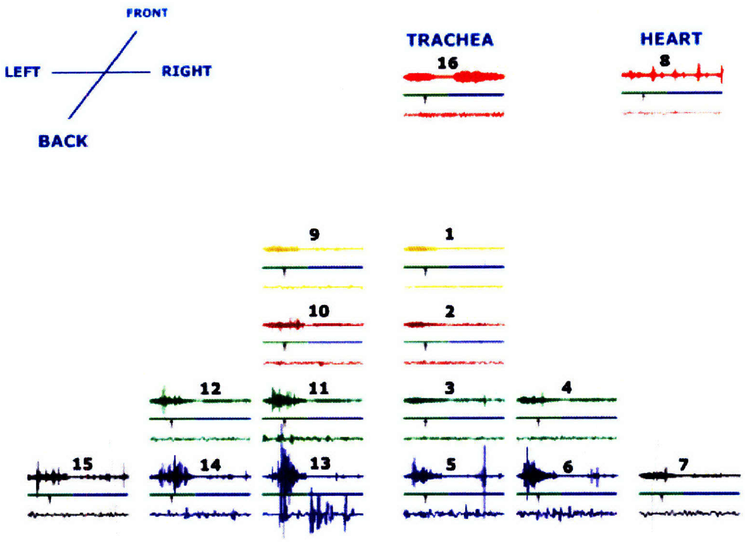


Figure 2.11 Waveform Pattern of an IPF Patient

Diagnosing IPF is frequently troublesome because of variability in chest x-rays. No true telltale signs exist when examining the x-rays. A better diagnostic technique is the CT scan. It can depict the extent of the fibrosis. Patchy infiltrates are often present at the base of the lungs. In Section 2.3.2, we pointed out that pneumonia has similar infiltrates – another possibility for misdiagnosis. In order to rule out other potential diagnoses, a transbronchial lung biopsy is sometimes performed [20].

2.3.6 Normal

Obviously, this is the one category that does not pertain to a specific disease. They are asymptomatic patients in the fact that they don't present any typical signs of an existing cardiopulmonary disease. The patients in this set have come to the hospital for a routine check-up or annual physical and have agreed to have their breath sounds recorded. From these patients, a subset was taken in order to statistically match the ages and demographics of the diseased population. Initially, one would think that a normal patient would not have any adventitious lung sounds. This is not the case for many normal patients. In fact, many of them possess crackles and wheezing to some extent. Figure 2.12 shows the waveforms of a typical normal patient. No adventitious lung sounds are present.

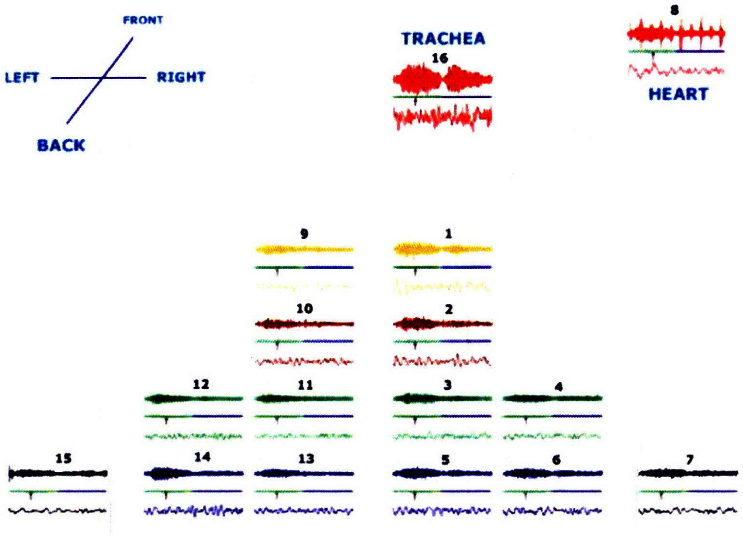


Figure 2.12 Waveform Patterns of a Normal Patient

2.4 Previous Studies Involving Computerized Auscultation

Computerized auscultation has resulted in several successful studies. As mentioned previously in Section 2.1, one of the first successful applications of TEWA is that Murphy et al were able to detect and characterize crackles in workers with exposure to asbestos. TEWA was used to train medical technicians for the surveillance of the workers [7]. Early on, computerized auscultation was also used to verify the number of crackles heard within a patient. Being able to perform accurate crackle counts justified the use of computers when listening to the chest [9].

Building on the early results of the detection of asbestosis via computerized auscultation, al Jarad and Strickland et al. compared the performance of TEWA to that of chest radiography and CT scans. They discovered that TEWA performed better than chest radiography in detecting the early phases of asbestosis and performed just as well as CT scans [21].

In 1994, Bettencourt et al, performed a study most similar in scope to ours. They utilized TEWA and tried to predict four diseases: PN, CHF, COPD, and IPF. They used multiple logistic regression to make diagnoses and were able to make diagnoses with 68% accuracy. The study differs from ours since they used a very different set of features and also their patient sample sizes were much smaller. Still, this is a very relevant example of the capability of the computerized auscultation [22]. In another study, Kawamura et al used time expanded waveform analysis to study 18 patients with IPF and 23 patients with crackles who did not have this disease. They too were able to separate IPF crackles from other diseases with some success [23].

Gavriely and Nissan evaluated the addition of computerized lung sound analysis to a questionnaire and spirometry measurements in a respiratory health screening program of 493 subjects. Although they did not perform TEWA, they detected adventitious lung sounds that were outside the normal range. The investigators found that the sensitivity for detection of respiratory disease rose from 71% to 87% by adding the lung sound information to the traditional tests [24].

Building on these benchmark studies led to the multi-channel STG-1602 in use today. In its first application it was used to study the properties of the sound transmission within the lung and its relationship to lung volume [25]. More recently, Murphy et al successfully determined that the lung sounds in patients with pneumonia can be separated from the lung sounds in asymptomatic patients. The study further verified the applicability of analysis by computerized auscultation [5]. Using the multi-channel analyzer, Vyshedskiy et al sought to describe how sounds travel throughout the chest. They documented that crackles in patients with CHF and PN were transmitted over a larger area than those of patients with IPF. They also found that the crackles of IPF also have a higher frequency [12].

Most studies in the field do not focus on determining the differences within diseases, but instead revolve around finding new methods to define the acoustic waveform more accurately. Their rationale is that the better the sounds are defined, the better the underlying physiology can be understood and a doctor will be able make a better diagnosis. One study that tries to better describe the sounds is performed by Kandaswamy et al. They use a more advanced technique known as wavelets to process the lung sounds since they are non-stationary. Once the wavelets are computed, they use neural networks (explained in Section 3.4.2) to determine if the sound is a crackle, squawk, wheeze, rhonchus, or normal [26]. A study by Taplidou and Hadjileontiadis aimed to construct an automatic technique for wheeze detection and monitoring using spectral analysis. Their efficient method defines wheezes very well even in the presence of background noise [27].

Furthermore, a recent paper by Güler, Polat, and Ergün use neural networks and genetic algorithms [28] to better distinguish adventitious lung sounds from the background noise. In their study, they claim that a time-frequency based modeling approach , such as TEWA, does not effectively reduce the background sounds [29]. Contrary to their criticism, we show that TEWA is still an effective means to summarize lung sounds.

2.5 Summary

This chapter has provided a brief background to the burgeoning field of computerized auscultation. We presented the methodology known as time expanded waveform analysis that was used to define more clearly the adventitious lung sounds such as crackles, wheezes, rhonchi, and squawks. We discussed five diseases that we investigate in this thesis and also an asymptomatic (normal) class. Finally, we reviewed previous studies involving computerized auscultation. Since it is a very new field, few studies exist that are similar in scope to this study. In the next chapter, we depart from the medical field and introduce data mining and machine learning techniques which are featured in our analysis.

[This Page Intentionally Left Blank]

Chapter 3

Foundations for Data Mining Analysis

As described in Chapter 1, the primary of goal of this research is to be able to provide a decision analysis tool for doctors to use when diagnosing chest and lung disorders. For the purpose of developing the automated diagnostic device, we will incorporate data mining and machine learning models. These models need to satisfy the following requirements:

- Fast recognition in the test phase
- Be able to operate in near real-time
- Be able to perform classifications when there are multiple classes
- Must be data-driven since all our knowledge stems from a large dataset

We will show that models introduced in this chapter satisfy the above mentioned requirements and therefore will be useful in developing the smart stethoscope.

This chapter introduces the scientific disciplines involved in the thesis and provides a detailed explanation of each. Furthermore, this chapter provides a description of the three main classification techniques employed in the thesis: neural networks, *k*-nearest neighbors, and support vector machines.

3.1 Scientific Disciplines

This thesis focuses on two disciplines that are very closely related. The first is the discipline of data mining and the second is machine learning. This thesis falls within both of these scientific disciplines. Loosely defined, data mining is the extraction of useful knowledge from vast databases. Nearly analogous to data mining is machine learning. Machine learning focuses on using computer based algorithms to learn and model real world behavior. Most of the time, characterizing the research as one discipline or the other is often a matter of semantics. In sections 3.1.1 and 3.1.2, we further define both machine learning and data mining and also explain how computerized cardiopulmonary diagnoses fit into both disciplines.

3.1.1 Data Mining

Computers and increased digital storage capacity have allowed electronic databases to grow to unprecedented sizes. Oftentimes, relationships and fundamental information about the data cannot be easily inferred because of the large size of the database. Data mining is an emerging scientific discipline that is focused on discovering these relationships. One textbook defines the field as “the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner” [30]. The author stresses that the data normally involved within a data mining analysis usually is not collected for this purpose. Instead, it can be collected beforehand for a separate purpose entirely and it isn’t until later when data mining techniques can be applied to reap some new, insightful, knowledge. In this sense, data mining is often an analysis on indirect data.

In our case, data was initially collected in order to provide teaching aids to those in the medical profession. Audio recordings of patients with certain cardiopulmonary disorders were analyzed in depth and compiled in a software package which provided both visual and auditory aids to the medical professional. Recently, because of the large amount of data, we are now applying data mining techniques to broaden the amount of knowledge we can discover from this dataset. The datasets include thousands of individual sounds which make the problem difficult to be analyzed using a rudimentary analysis.

3.1.2 Machine Learning

Machine learning is a field of research that falls within the realm of artificial intelligence. Loosely defined, artificial intelligence is a branch of computer science that is concerned with the automation of intelligent behavior [31]. Generally speaking, machine learning attempts to develop algorithms to allow computers to understand and model real world processes. Hence, the intuitive name "machine learning." Machine learning techniques have been widely applied to adaptive control theory, brain modeling, evolutionary learning, and statistics. In this thesis, we focus on the statistics discipline. Statistical learning occurs in two forms: supervised learning, and unsupervised learning. In supervised learning, a model is built with the purpose to be able to map input objects to a desired output value. If the model is built to predict a continuous output it is known as "regression", if it is used to predict a class label, it is called "classification." In unsupervised learning, there is no output value. Instead, the purpose is to be able to describe the relationships among the inputs or how the data is organized. Classification problems are often described as "Black Box" problems. Within the "Black Box" lies the complex modeling technique that ultimately provides a classification. We will look at several techniques later on in this chapter. Figure 3.1 provides an illustration of the "Black Box."

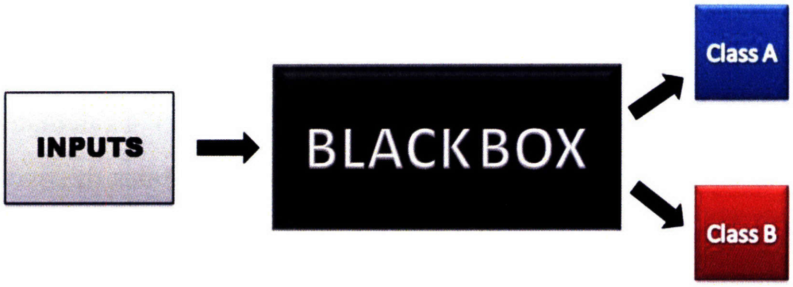


Figure 3.1 "Black Box" Illustration of Classification

We use supervised learning because the data are labeled in advance. Specifically, this problem represents a classification. Each patient has a specific disease which we will try to classify based on available lung sounds and clinical data. Problem formulation will be discussed in more detail in Chapter 4.

3.2 Classification Process Overview

In order to accurately define the classification problem, we first need to introduce some notation and terminology. We let $\boldsymbol{x_i} \in \mathbb{R}^{\boldsymbol{d}}$ represent the *d*-dimensional *input feature vector* that represents the characteristics of the i^{th} sample we are attempting to classify. Other texts may refer to the feature vector as an *input vector* or *attribute vector*. The individual components of $\boldsymbol{x_i}$ are likewise referred to as *features*, *attributes*, or *input variables*. These features are usually defined by the user and are normally numeric or categorical in type. Each feature vector has an associated *class* or *label* $y \in \{-1, +1\}$.

Combining an individual feature vector, $\boldsymbol{x}_i$, with a label, y_i, results in a *sample* $(\boldsymbol{x}_i, y_i)$. A *training set, S,* consists of a series of *N* samples which can be written as follows:

$$S = \{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_N, y_N)\} \quad \forall\, i = 1..N \tag{3.1}$$

The goal of the classification algorithm, is to use the training set to create a function which can map an unknown vector, $\boldsymbol{x}$, onto a class, y. This process is known as *training* or *learning.* The function should be able to classify samples within the training set accurately (*empirical performance*) and also be able to perform well on a *test set* of data unseen in the modeling steps (*generalization performance*).

The end result of the classification is a *classification function*, $h(\boldsymbol{x}): \mathbb{R}^{\mathrm{d}} \longrightarrow \mathbb{R}$ that computes a single value from all of the input variables. Once $h(\boldsymbol{x})$ is calculated, most algorithms have a threshold that is used to determine a cutoff value that separates the two classes. The most common threshold is zero for ease of interpretability of the *decision function*, $f(\boldsymbol{x})$. Setting the threshold at zero results in the following decision function:

$$f(\boldsymbol{x}) = sign(h(\boldsymbol{x})) \tag{3.2}$$

This decision function returns a +1 or -1 for the binary classification process.

Evaluation of empirical and generalization performance is done in terms of *error*, or *risk*, used interchangeably. Before we define error, we first explain the concept of *loss.* A *loss function,* Loss(f($\mathbf{x}_i$), y_i), is a measure of how far the prediction varied from the actual class. There are many types of loss functions for both unsupervised and supervised learning techniques. For supervised learning with a binary decision function, an applicable loss function is the following:

$$Loss(f(\boldsymbol{x}_i), y_i) = \begin{cases} \mathbf{1} & \text{if } f(\boldsymbol{x}_i) = y_i \\ \mathbf{0} & \text{otherwise} \end{cases} \quad (3.4)$$

This function is known as the 0/1 loss function since it returns a 0 when the classification is correct and a 1 otherwise.

Now that we have defined the loss function, we can return to the concept of error. Let the *test error,* or *generalization error,* be defined by the symbol ε. It is the expected prediction error over an independent test sample

$$\varepsilon = \mathbb{E}[Loss(f(\boldsymbol{x}_i), y_i)] \quad (3.5)$$

The test error is the best estimate of how a classifier performs when it is subjected to unknown feature vectors that have been generated from the source distribution. In order to calculate ε exactly, one would need to know the underlying distribution function of the data in order to calculate the expectation. In general, this distribution is never known, which is why the performance must be estimated by a test sample. Similarly, the *training error,* or *empirical error* can be defined by the symbol $\overline{\varepsilon_N}$.

$$\overline{\varepsilon_N} = \frac{1}{N}\sum_{i=1}^{N} Loss(f(\boldsymbol{x}_i), y_i) \quad (3.6)$$

Training error is the average loss over the training sample [32]. The training error is an estimate of the test error; however, at times it isn't a good one. Training error is directly related to model complexity. Often times, it is possible to achieve almost zero training error by building a very complex model, but the model generalizes very poorly. As a model becomes more complex, both training and test error decrease. However, at some point the model becomes too complex and focuses too much on explaining the intricacies in the training data. As a result, the test error begins to increase. This tradeoff is shown in Figure 3.2 [32].

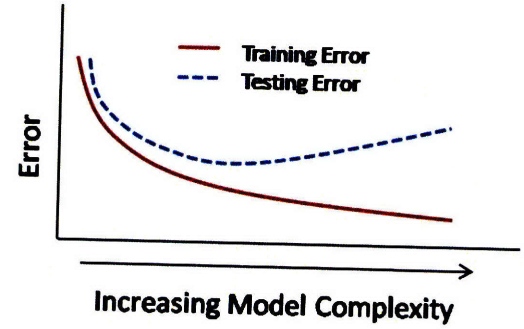


Figure 3.2 Tradeoff of Error and Model Complexity

When a model is fit too much based upon the training data, it is known as over-fitting. Figure 3.3 provides an illustration of two models, one that grossly over-fits the data, and another which will be able to generalize well.

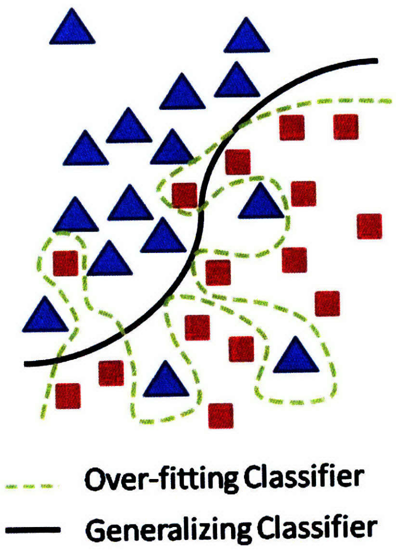


Figure 3.3 Example of Over-fitting

Obviously, the goal of any classifier is to maximize generalization performance. In order to be effective in the real world, a classifier must be able to perform well when asked to classify unseen data. However, within the learning process of a classifier, different types of errors are minimized depending on the type of classifier chosen. Some, such as *k*-nearest neighbors and neural networks, minimize the training error, or the *empirical risk*. These methods are founded upon the hope that data in the test set are generated from the same distribution as those that the model has been trained upon. As a result, good performance on the training set will likely translate into similarly good performance on a test set. These algorithms are said to follow the *empirical risk minimization* (ERM) principle. Generalization error can be minimized by using cross validation to select the best parameters for the model.

On the other hand, a method developed by Vapnik and Chervonenkis known as support vector machines (SVM) seeks to minimize both the empirical error and generalization error simultaneously. A function that takes into account both empirical and generalization error is said to follow the *structural risk minimization* (SRM) principle [33]. A technique within the learning process seeks to control the generalization error whereas in ERM methods, the generalization error can only be examined after the model has been fully learned. By taking into account both types of error within the learning stage, there is potential for greater generalization performance than with ERM methods. The next section goes into more depth explaining the concepts of risk.

In summary, building a classifier takes several steps that can be seen in the process flow shown below in Figure 3.4 [34].

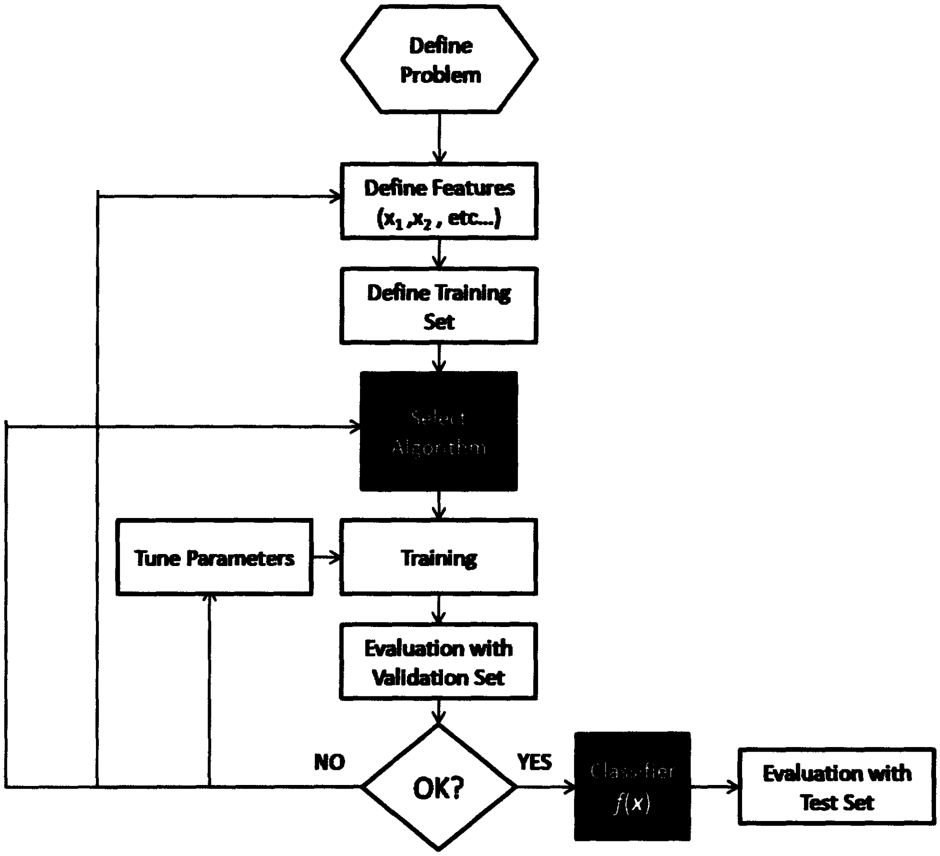


Figure 3.4 Process Flow Chart for Building a Classifier

As one can see, building a classifier is an iterative process. If one algorithm doesn't perform as well as hoped, the parameters can be tuned, and the model can be retrained. Also, the process can go back even further to redevelop the features and introduce new ones, or even choose a new classification algorithm altogether.

In the following sections, we further explain the concepts of empirical risk minimization and of structural risk minimization. In addition, we discuss the three learning methods briefly mentioned in the preceding paragraphs. They include *k* nearest neighbor, neural networks, and support vector machines. Finally, this section has been a very brief overview of the process of learning a classifier from data. Much more information can be found in [32, 33, 35].

3.3 Learning Theory and Risk Revisited

As mentioned previously, the concept of risk minimization is central to any learning algorithm. In this section, we provide a summary of the learning theory and develop a framework for both empirical risk minimization and structural risk minimization. To start, we expand on our notation of a decision function $f(\mathbf{x})$. Let $f(x,w)$ be a specific decision function defined by fixed parameters $\boldsymbol{w}$. This vector, $\boldsymbol{w}$, can be viewed as a set of weight parameters for each corresponding feature. In the training phase, depending on the training set, the classifier will ultimately choose a weight vector for use in classifying. With this new notation, we define the test error in terms of a *actual risk* , $R(w)$.

$$R(\boldsymbol{w}) = \int |y - f(\boldsymbol{x}, \boldsymbol{w})| dP(\boldsymbol{x}, y) \tag{3.7}$$

If $R(w) = 0$, then the classifier will never make an error and generalizes perfectly for any unknown feature vector. In Equation 3.7, $P(\boldsymbol{x}, y)$ represents the joint cumulative distribution function of the features and class. Since it is unknown, the distribution is empirically learned from the training set.

We also update the Equation 3.6 in terms of empirical risk to remain consistent with the literature [36].

$$R_{emp}(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N} Loss(f(\boldsymbol{x}_i, \boldsymbol{w}), y_i) \tag{3.8}$$

Unlike the actual risk, the empirical risk does not depend on the unknown probability distribution, only the training set and chosen decision function. Also of note, the empirical risk can also be minimized with respect to the weight vector, $\boldsymbol{w}$. These two components provide the foundation for empirical risk minimization.

Empirical risk minimization is the process of determining a decision rule by finding a weight vector $\boldsymbol{w}_{emp}$ from all potential vectors $\boldsymbol{w} \in \mathcal{W}$ that minimizes the risk. More specifically,

$$R_{emp}(\boldsymbol{w}_{emp}) = \inf_{\boldsymbol{w} \in \mathcal{W}} R_{emp}(\boldsymbol{w}) \tag{3.9}$$

It also can be show that

$$\inf_{\boldsymbol{w} \in \mathcal{W}} R_{emp}(\boldsymbol{w}) \overset{P}{\rightarrow} \inf_{\boldsymbol{w} \in \mathcal{W}} R(\boldsymbol{w}) \quad \text{as} \quad N \rightarrow \infty \tag{3.10}$$

This says that as the training set grows larger and larger, the minimum empirical risk converges in probability to the minimum actual risk. Derivations and proofs of empirical risk minimization can be found in [37].

In one of the most recent developments in the statistical learning field, Vapnik introduced a bound on the actual risk [33]. With the 0/1 loss function and a parameter η such that 0 ≤ η ≤ 1, the following bound holds with probability 1-η [36]:

$$R(\boldsymbol{w}) \leq R_{emp}(\boldsymbol{w}) + \sqrt{\left(\frac{h\left(\log\left(\frac{2N}{h}\right)+1\right)-\log\left(\frac{\eta}{4}\right)}{N}\right)} \tag{3.11}$$

In Equation 3.11, h is a non-negative integer known as the Vapnik Chervonenkis (VC) dimension, N is the number of samples in the training set, and the risks are the same as defined in Equations 3.7 and 3.8. Loosely defined, the VC dimension is a measure of the complexity of the family of classifiers $f \in \mathcal{F}$. For simplicity's sake, a full derivation of the VC dimension is not mentioned here. It is sufficient to assume that the more complex a classifier becomes, the higher it's VC dimension.

From Equation 3.11, one can see that the actual risk can be limited by minimizing training error, having a large training set, and also controlling the size of the VC dimension. Limiting the VC dimension is the fundamental concept of structural risk minimization. In the derivation of support vector machines we show how this concept is applied. A much more thorough reference on SRM and an explanation of the VC dimension is found in [36].

3.4 Explanation of Specific Learning Algorithms

After providing an ample background on the processes of data mining, machine learning, and classification, we finally are able to delve into the black box presented earlier. These subsections will highlight the three learning algorithms explored in this thesis. We present each method with discussion on its advantages and disadvantages.

3.4.1 K Nearest Neighbors

K nearest neighbors (kNN) is one of the oldest learning algorithms and is still useful in many cases. It was pioneered by Fix and Hodges in 1951 [38]. It belongs to a class of algorithms known as lazy learning algorithms. In lazy learning, the classifiers are based solely on the training set, and no additional model needs to be fit. Given an unknown sample, the kNN algorithm finds the k samples in the data set closest in distance to the unknown sample and then classifies using a majority vote [32]. The parameter *k* is specified by the user. Typical distance metrics include the Euclidean metric,

$$D(x_i, x_0) = \|x_i - x_0\| \qquad (3.12)$$

or the Manhattan metric,

$$D(x_i, x_0) = \sum_{j=1}^{m} |x_i^j - x_0^j| \qquad (3.13)$$

where *m* is the number of features.

Figure 3.5 below illustrates the kNN concept. The green circle in the middle is the sample to be classified. If *k* is chosen to be three, the circle will be classified as a member of the blue triangle class. However, if *k* is chosen to be seven, the circle will be classified as a member of the red square class.

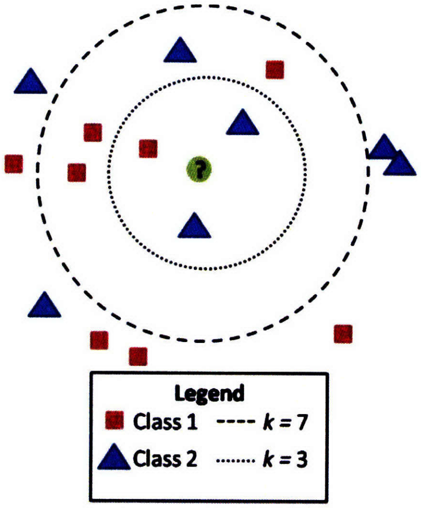


Figure 3.5 *k* Nearest Neighbor Classification Example

For being a relatively simple classifier, it has the ability to perform very well on certain datasets. One pioneering study involved the recognition of handwritten numerals and kNN performed the best out of several learning approaches [39].

In spite of its good performance on some datasets, kNN is not without its flaws. One primary concern is its large memory storage requirements. The model consists of every element of the training set so trying to implement it as a classifier may be very slow. A second concern is that the choice of the neighborhood size, *k*, greatly affects the performance of the algorithm. If the data is noisy, i.e. the points are relatively intermixed; a small *k* could potentially result in classification errors. Similarly, if the region that defines a certain class is very small in number compared to another class, often times this region will be completely over looked due to the prevalence of the other class in the training data set. These problems can be partially solved by varying the parameter *k* [34]. Also, the kNN approach is very sensitive to perturbations in the data and also irrelevant features since all features bear the same weight. Another problem with this classification method is that the decision boundary is hard to conceptualize since it only depends on the training set in a very high dimension feature space. Still, *k* nearest neighbors provides a simple, easily understandable classifier that has the potential to perform well.

3.4.2 Neural Networks

Unlike *k* nearest neighbors, neural networks are often viewed as the hardest classification algorithm to grasp. They are an attempt to create a classifier built upon the architecture of the human brain. In the basic single hidden layer neural network, there are three total layers. The first is the input layer which inputs all feature vectors within the training set. The second layer is the output layer which contains the results of the classification. The third layer is referred to as a "hidden" layer. All three layers consist of a set of neurons, thus the name neural networks. This architecture can be seen in Figure 3.6.

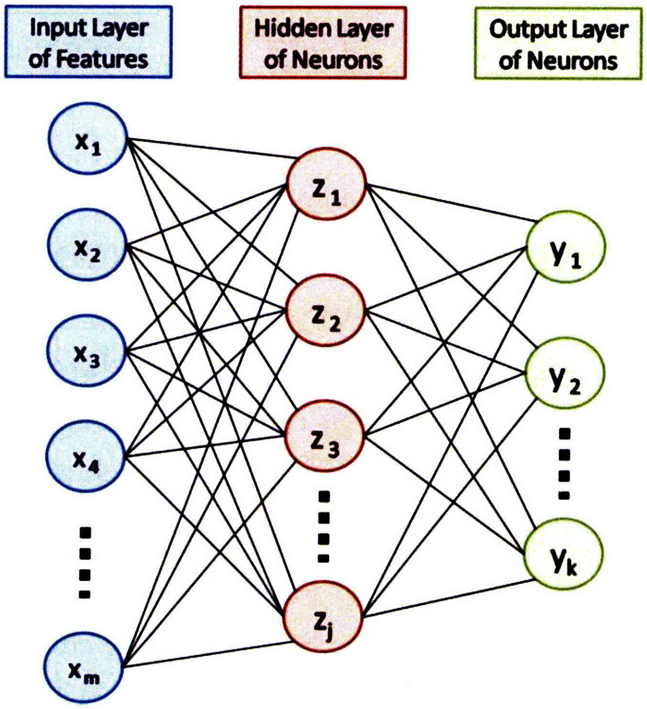


Figure 3.6 Neural Network Architecture with a Single Hidden Layer

The premise of a neural network is to create linear combinations of the feature vectors in the hidden layer, and then model the output layer as a nonlinear function of the hidden layer [32]. Within the neurons of the hidden layer are various activation functions which allow the neural network to represent complicated non-linear relationships. Each connection in the network has a certain weight associated with it. When the neural network is trained, the weights are constantly modified and adjusted in order to minimize the training error. More complex neural networks can be made by adding neurons to the hidden layer or even adding more hidden layers.

The most common method to train a neural network is through a process known as back propagation. Kotsiantis provides six key steps to training a neural network through back propagation [34].

1. Present a training sample to the neural network
2. Compare the network's output to the desired output from that sample. Calculate the error in each output neuron.
3. For each neuron, calculate what the output should have been, and a scaling factor, how much lower or higher the output must be adjusted to match the desired output. This is the local error.
4. Adjust the weights of each neuron to lower the local error.
5. Assign "blame" for the local error to neurons at the previous level, giving greater responsibility to neurons connected by stronger weights.
6. Repeat all previous steps until a stopping criterion is reached.

Valid stopping criteria include stopping after the training sample has been presented to the network a certain number of times (known as epochs), stopping when the error reaches a threshold, or stop if there is no improvement in error over several epochs.

Because of neural networks very high complexity, extremely low training errors are normal, and achieving zero training error is not unheard of. Still this method also has some drawbacks. One that has been hinted at already is that with very high complexity also comes the possibility of over-fitting the training data. The ability to generalize well is of utmost importance. In order to combat over-fitting, premature stopping criteria are sometimes invoked. Another drawback is that there is no sure way to set up a network with the appropriate number of layers and neurons. A network with too few neurons will not be able to capture the relationships in the data, and a network with too many neurons will tend to over-fit. Finding an appropriate size can be done through cross-validation or trial and error. An additional drawback of building a large neural network is that the length of training time grows with the size of the problem. The back propagation algorithm is not very fast and for large datasets with lots of features, training the algorithm becomes a very time intensive procedure. Finally, one last drawback is that the solution space is non-convex with many local optima. Oftentimes, the algorithm gets stuck in one of these local optima leading to poor generalization performance [32].

Nevertheless, neural networks are one of most commonly studied learning methods with literally hundreds of articles that document its successful application in practice. Good textbooks for a more in depth analysis are [37, 40]

3.4.3 Support Vector Machines

As mentioned previously, support vector machines (SVM) are one of the newest methods in the supervised learning field. They are developed by Vapnik in his book [33]. Unlike *k* nearest neighbors and neural networks, support vector machines attempt to minimize generalization error within the framework of the algorithm. They do this by controlling both the training error, and the VC dimension as shown in Equation 3.11. Generally speaking, a support vector machine seeks create a hyperplane that separates the two data classes. Not only does the hyperplane separate the data, but also it is oriented in such a fashion that creates the maximum "margin" on both sides of it ensuring the largest possible separation between the two classes. This concept will become clearer in the following paragraphs.

For ease of explanation, we assume that the training data can be separated by a linear hyperplane as shown in Figure 3.7 [41]. All data in the negative class lie on one side, whereas all data in the positive class lie on the opposite side of the hyperplane. The central question to support vector machine is: which hyperplane separates the data the best?

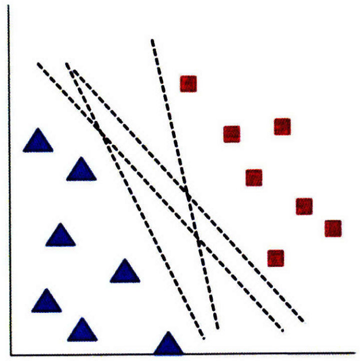


Figure 3.7 Linearly Separable Data in 2D with Several Hyperplanes

Algorithms that follow the empirical risk minimization principle would not distinguish from the sets of hyperplanes above. However, support vector machines are able to find an optimal hyperplane. To determine the best separating hyperplane, we first introduce some more mathematic notation. All points that lie on the hyperplane satisfy the equation $\boldsymbol{w^T} \cdot \boldsymbol{x} + b = 0$, where $\boldsymbol{w}$ is normal to the hyperplane, $\frac{|b|}{||\boldsymbol{w}||}$ is the perpendicular distance from the hyperplane to the origin, and $||\boldsymbol{w}||$ is the Euclidean norm of $\boldsymbol{w}$. Finally, let d_+ and d_- represent the distances from the hyperplane to the nearest sample on the positive and negative side, respectively [36]. The sum of the distances is the margin. These terms of interest are depicted in Figure 3.8.

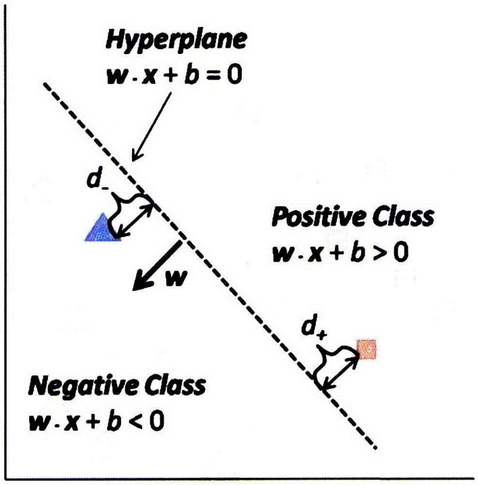


Figure 3.8 Geometric Definitions for a Separating Hyperplane

The hyperplanes depicted above can be defined by the following equations:

$$\boldsymbol{w^T} \cdot \boldsymbol{x_i} + b \geq 1 \quad \text{if } y_i = +1 \tag{3.14}$$

$$\boldsymbol{w^T} \cdot \boldsymbol{x_i} + b \leq -1 \quad \text{if } y_i = -1 \tag{3.15}$$

Combining Equations 3.14 and 3.15 results in the following expression,

$$y_i(w^T \cdot x_i + b) \geq 1 \quad \forall i = 1, 2, \dots, N \tag{3.16}$$

This constraint enforces the fact that all samples must be classified on the correct side of the hyperplane. Since the data is linearly separable, there exist points such that Equations 3.14 and 3.15 are tight. Essentially, these tight constraints form two additional hyperplanes that are parallel to the optimal separating hyperplane. The distance between the two hyperplanes is known as the margin. With some arithmetic it can be shown that the distance from one of the hyperplanes to the separating hyperplane is equal to $\frac{1}{\|w\|}$. This derivation can be found in an appendix.

In order to train a support vector machine, it requires solving an optimization problem. The goal of the SVM is to find the separating hyperplane with the largest margin subject to the constraint of classifying all the points correctly. Remember, this data is assumed to be linearly separable. The optimization problem is as follows:

$$\text{maximize}_{w,b} \ \frac{1}{\|w\|}$$

$$\text{subject to} \quad y_i(w^T \cdot x_i + b) \geq 1 \quad \forall i = 1, 2, \dots, N$$

Notice that minimizing $\frac{1}{2} \ \|w\|^2$ will produce the same result as the previous formulation. This results in the similar problem below.

$$\text{minimize}_{w,b} \ \frac{1}{2} \ \|w\|^2$$

$$\text{subject to} \quad y_i(w^T \cdot x_i + b) \geq 1 \quad \forall i = 1, 2, \dots, N$$

This reformulation translates to solving a convex quadratic programming program. Minimizing a convex function is very beneficial since there are no local minima and therefore a global optimum can always be found. There are plenty of software programs that can solve a problem of this type. The samples for which the inequality constraint holds define the location for the optimum separating hyperplane. These samples are known as the support vectors and thus the name of the method. By constructing the optimal separating hyperplane in this manner, the VC dimension is constrained [33]. Limiting the VC dimension is a benefit of SVM that neural networks and *k* nearest neighbors do not possess. Regulating the VC dimension allows for better generalization performance theoretically. A final figure depicting an optimal separating hyperplane for our contrived example is shown below.

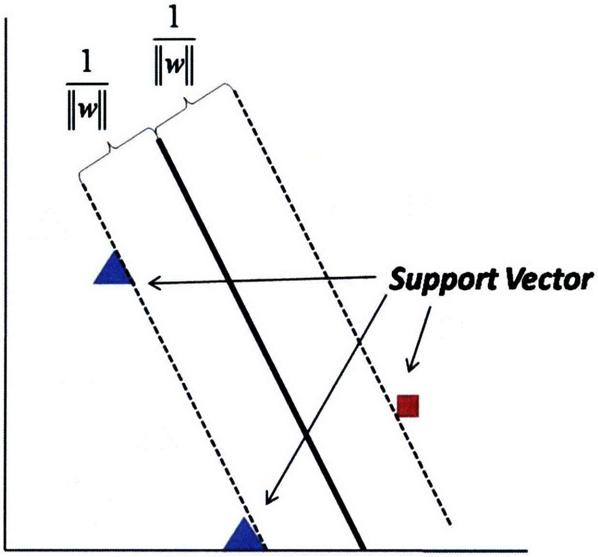


Figure 3.9 Optimal Separating Hyperplane and Support Vectors

In many cases the linearly separable assumption does not apply because the data set contains some overlap. A formulation known as the *soft margin* allows for some training samples to be misclassified. In order to do this, slack variables, ξ_i , are introduced to the constraints. The slack variable takes a positive value if the constraint cannot be satisfied for a given sample. For each misclassification, a penalty parameter, *C*, is added to the objective function. For large values of *C,* the margin is smaller in order to correctly classify more points. For small *C,* the support vector machine will place more importance on creating a large margin than classify all the training samples correctly. The new optimization problem is rephrased as follows:

$$\text{minimize}_{\mathbf{w},b} \ \frac{1}{2}\|\boldsymbol{w}\|^2 + C * \sum_{i=1}^{N} \xi_i$$

$$\text{subject to} \quad y_i(\boldsymbol{w}^T \cdot \boldsymbol{x}_i + b) \geq 1 - \xi_i \quad \forall i = 1, 2, \dots, N$$

$$\xi_i \geq 0 \qquad \forall i = 1, 2, \dots, N$$

One final property of support vector machines to bring up here is their ability to extend to non-linear problems as well. In order to do this, the features are mapped to a higher dimensional space known as the *transformed feature space.* A linear separation in the higher dimensional space, corresponds to a non-linear separation in the input space [34]. A kernel function, $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, calculates the inner product of two input vectors and transforms the data into a higher dimension. Popular choices for kernel functions include a polynomial kernel with degree *p*,

$$K(x_i, x_j) = (x_i \cdot x_j + 1)^p \tag{3.17}$$

and a radial basis function (RBF) kernel with width σ,

$$K(x_i, x_j) = e^{-\|x_i \cdot x_j\|^2 / 2\sigma^2} \tag{3.18}$$

Choosing to use a kernel is a decision made by the user before training the model. Using kernels lengthens the amount of training time significantly. [34].

Support vector machines are very popular in practice because of their ability to generalize well and their computational efficiency. Both of these properties stem from the fact that the SVM problem is formulated as a convex optimization problem. This formulation guarantees an optimal solution in a finite amount of time. A major drawback of SVM is that the general problem does not translate into a multi-class problem theoretically. Methods implementing multiple binary classifications are shown in Section 3.5.

Several good resources on the derivation of support vector machines are available in [33, 36].

3.4.4 Classification Methods Summary

All three classification methods mentioned here have certain advantages and disadvantages which have been mentioned throughout the section. In a recent journal article, Kotsiantis provides a rating for each method in tabular form [34]. That table is modified for our purposes and shown below in Table 3.1. One star represents poor performance, whereas four stars represent the best performance out of all classes of learning algorithms.

	k Nearest Neighbor	Neural Networks	Support Vector Machines
Ability to Generalize	★★	★★★★	★★★★
Speed of Learning	★★★★	★	★★
Speed of Classification	★	★★★★	★★★★
Tolerance to irrelevant features	★★	★★★	★★★
Model interpretability	★★★	★★	★★

Table 3.1 Comparisons of Learning Algorithms

3.5 Extensions to Multiple Class Problems

Most of the derivations of the three methods in Section 3.4 were given in the context of a binary classification problem. In this thesis, it is important to be able to distinguish different diseases based upon a pair-wise comparison. In practice, suppose a doctor has studied the symptoms of a patient and has narrowed down the possible diagnoses to two possible diseases. A pair-wise acoustic analysis between these two diseases would be beneficial to aid the doctor make the correct diagnosis. Although this is an important contribution of this thesis, we feel that a true multi-class problem is where the most potential gains lie. A potential long term vision of this project is that a patient with symptoms pointing towards some sort of cardiopulmonary disease comes into the hospital, lies down on the acoustic analyzer, and a computer makes an instantaneous diagnosis out of all possible diseases. Therefore it will be beneficial to frame the problem in this multi-class way.

The *k* nearest neighbor algorithm is relatively simple to expand to a multi-class context. Simply include samples from all classes in the training set, and the nearest neighbor algorithm will make the classifications accordingly.

Similarly, extending the multi-class framework to neural networks is a simple modification. Again, include samples of all classes into the training set, and modify the output layer of the network. The network should have one output neuron for each class to be modeled.

Extending the multi-class scenario to the support vector machine algorithm is trickier. In the formulation we presented of SVM, only binary comparisons are possible. In spite of this, ongoing research documents how to formulate SVM as a multi-class classifier. Some popular attempts involve forming multiple binary classifiers. The two classifiers we look at in this thesis are the One-Vs-One (OVO) classifier and the One-Vs-All (OVA) classifier.

First, One-Vs-One was pioneered by Kreβel in [42]. This methodology involves creating binary classifiers for all pairs of classes. If we let k be the number of classes, the method involves constructing $\binom{k}{2}$ separate SVMs, one for each pair. For a simple three class example, the comparisons would be Class 1 vs. Class 2, Class 1 vs. Class 3, and finally Class 2 vs. Class 3. In order to assign a unknown sample to a class, it is run through each trained classifier. Each classifier "votes" for a class and the class receiving the most votes for a sample wins. In case of a tie in the amount of votes, the tiebreaker will be the classifier involving the two tied classes and seeing the output of that comparison [42].

The second approach we consider is the One-Vs-All approach. Instead of computing $\binom{k}{2}$ classifiers, it only calculates k SVMs. It constructs a SVM for each class k as the positive class, and lumps all other classes into one combined negative class. For our simple three class example, the SVMs are Class 1 vs. Class 2&3, Class 2 vs. Class 1&3, and Class 3 vs. 1&2. For an unknown sample, its classification is based on which hyperplane it was the furthest from. The point must lie on the positive side of the hyperplane which implies it is on the singular side of the hyperplane instead of the side where the classes are grouped. The distance from the hyperplane indicates a degree of confidence of the prediction. A point that lies far from the separating hyperplane is a much more confident prediction than one that lies near the hyperplane. In equation form, the decision is:

$$\boldsymbol{f(x) = \arg\max_{i=1,2,\dots,k}(w_i{}^T \cdot x + b_i)} \qquad (3.19)$$

A recent paper by Rifkin heralds this approach and claims it is well founded in regularization theory and is a time tested and very appropriate formulation of the multiclass problem [43]. The decision boundaries for both OVO and OVA are shown in Figure 3.10.

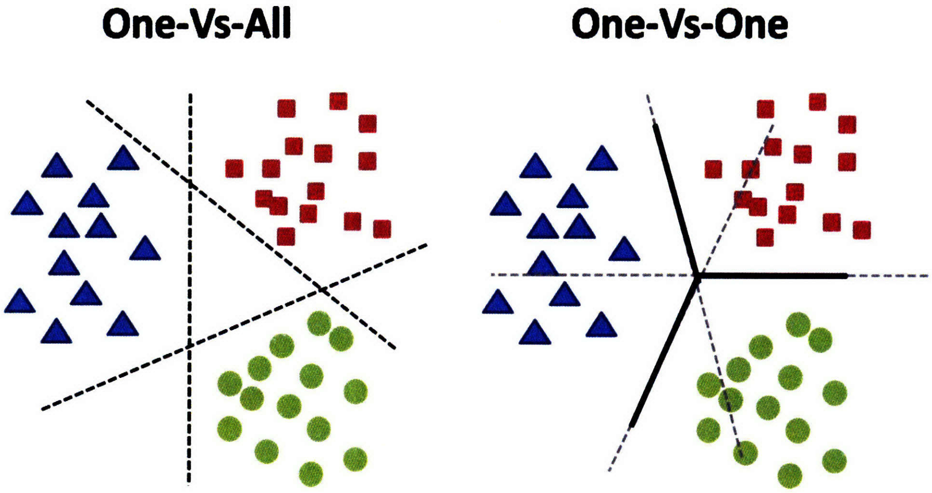


Figure 3.10 Decision Boundaries for Multiclass Support Vector Machines

3.6 Summary

In summary, this Chapter has provided a brief introduction of the data mining and machine learning concepts that this thesis revolves around. Hopefully, a novice in the field now has enough background in order to understand the central concepts to this thesis. Furthermore, we have provided partial derivations and explanations of three commonly used methods for classification problems: *k* nearest neighbors, neural networks, and support vector machines. Each method has certain advantages and disadvantages associated with them. We choose to explore all three methods for use as a baseline performance metric. In general, if a method works for data mining problem, other methods should work too. However their performance may vary to a certain degree. By using several methods, we effectively check our work in order to present data mining as a viable option to explore for acoustic cardiopulmonary diagnosis.

Chapter 4

Classifying Lung Sounds

In Chapter 4, we examine the precise methodology in which we apply the techniques of machine learning and data mining to extract reasonable information from the adventitious lung sounds data set. This chapter will examine how we implement the steps to build a classifier as we described in Figure 3.4. In particular, we discuss the development of the input features. We also discuss frameworks that we use to explore the problem of diagnosing lung disorders based upon the computerized auscultation.

First we classify the data using individual crackles only for both pair-wise and multi-class comparisons. We call this methodology Method 1. With this approach, we seek to answer the following question: does the underlying physiologic source of the crackle cause audible differences that cannot be picked up by the human ear? The second approach combines the individual crackles and other adventitious sounds into cumulative features that describe breaths instead of just a single sound. This is called Method 2. In this framework, the learning algorithm seeks to makes a classification of the individual breaths. This methodology gives us a much larger sample size than an analysis where each patient is summarized by a single feature vector. We also discuss other important factors such as determining training and testing sample sizes and various model validation approaches.

4.1 Method 1 - Classifying Individual Crackles

This method addresses only individual crackle features. It has the smallest set of features and does not include any features describing the other adventitious lung sounds. In this method, data obtained from the auscultation of squawks, wheezes, and rhonchi are left out. By leaving this data out, we acknowledge the fact that we may not have enough acoustic information to predict diseases where these sounds play a predominant role such as asthma and COPD. Not only is it the most basic as far as the algorithm development, it also is the most basic from a physician's standpoint too. It explores the fundamental differences between crackles. No clear common medical explanation of crackles is known, but it is hypothesized that crackles are generated by different processes and as a result have different acoustic characteristics. This method is predominantly used for diagnosing patients where crackles are the principal component. The following subsections detail the feature definitions, training set selection, validation, and testing procedures developed for Method 1.

4.1.1 Feature Definitions

For Method 1, we develop features that are only based upon individual crackle sounds. Each individual crackle is fully defined by a single set of features. The features have been defined and derived by Andrey Vyshedskiy et al. of Stethographics, Inc. They have been recorded by the STG-1602 multi-channel lung sound analyzer. The multi-channel approach lets us find the precise location on the chest where the waveforms occur. Figure 4.1 shows a close-up of a typical crackle waveform. Several characteristics of the waveforms are labeled. The crackle analysis starts by identification of the crackle's highest deflection or highest peak. The half period to the left of the highest peak is marked as T_1. The half period to the right of the highest peak is marked as T_2. Crackle frequency is calculated from four consecutive half periods, with T_1 as the first half period. Crackle amplitude is marked with A_1, A_2, etc. Crackle polarity is defined positive if the highest peak is upward and negative if the peak is downward.

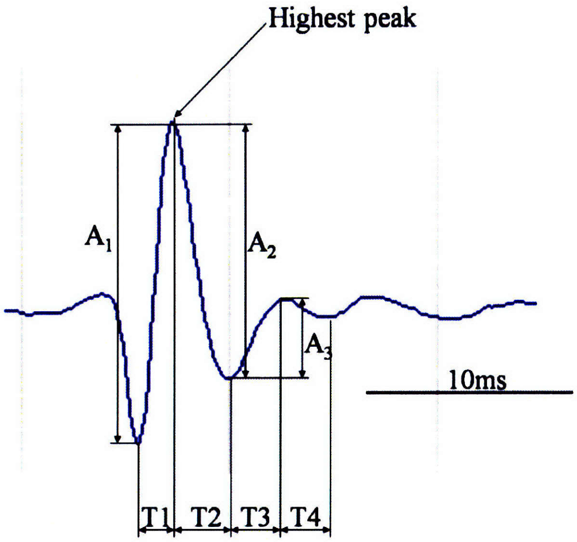


Figure 4.1 Example Crackle Waveform with Labels

These labels are referred to in Table 4.1 which lists all of the features used to design this experiment. All in all, 22 features are used to describe each crackle. They are listed with brief descriptions in Table 4.1.

Feature Name	Description
Zero Crossings (ZXS)	The number of times the crackle waveform crossed the baseline
T1	First half period
T2/T1	Ratio of the 2nd and 1st half periods
Half Period Variability (Tvar)	Standard Deviation $\{T_1,T_2,...,T_n\}$ /Mean $\{T_1, T_2,..., T_n\}$
Frequency (Freq)	Crackle frequency calculated from 4 half periods: T1, T2, T3, and T4
Timing	Discrete values of 1,2,3 represent the phases (early, mid, and late) of inspiration, whereas 4,5,6 represent expiration
Crackle Transmission Coefficient (CTC)	The degree of crackling sound transmission through the ipsilateral chest, as calculated from crackle family observation by multiple microphones
Amplitude	Amplitude of the highest peak (arbitrary units)
A2/A1	Ratio of the 2nd and 1st amplitudes
A3/A1	Ratio of the 3rd and 1st amplitudes
Amplitude Variability (Avar)	Standard Deviation $\{A_1,A_2,...,A_n\}$ /Mean $\{A_1, A_2,..., A_n\}$
PolarityUp	Direction of the highest peak (1 or 0)

Feature Name	Description
PeakSharpness	The measure of the sharpness of the highest peak
DelayDist_Intercept, DelayDist_Slope, & DelayDist_Correlation	For each crackle family, the delay between daughter crackles and mother crackles was analyzed as a function of linear distance between the corresponding microphones. A linear regression was performed to find the intercept, slope, and correlation.
AmplDist_Intercept,AmplDist_Slope, & AmplDist_Correlation	For each crackle family, the amplitude of the daughter crackles was analyzed as a function of linear distance between the daughter crackle microphone and the mother crackle microphone. A linear regression was performed to find the intercept, slope, and correlation.
DelayAmpl_Intercept, DelayAmpl_Slope, & DelayAmpl_Correlation	For each crackle family, the delay between daughter crackles and mother crackles was analyzed as a function of crackle amplitude. A linear regression was performed to find the intercept, slope, and correlation.

Table 4.1 Crackle Feature Definitions Used in Method 1

4.1.2 Data Pre-Processing and Standardization

To be able to process the data, we first needed to take certain steps in order to "clean" it. All crackles with missing attributes, or errors were deleted. Furthermore, algorithms perform more efficiently if the data is scaled before a model is trained. Scaling the data also has the added effect that no feature is represented by values of a significantly higher order of magnitude than the other features. Since the features can be represented on a similar scale, no preference is given to an individual feature. Most standardizations map the values to a range of [-1, 1] or [0, 1]. We chose to normalize the data between [-1, 1]. We used the following standardization for all features:

$$s = \frac{2*(x-\min_j(x))}{\max_j(x)-\min_j(x)} - 1 \tag{4.1}$$

where the $\min_j$ and $\max_j$ terms represent the minimum values of x across all samples for the feature to be standardized. x represents the particular value, and s is the standardized equivalent of this value.

The total size of the data sets for Method 1 is shown in Table 4.2.

Disease	Number of Patients	Number of Crackles
PN	123	5518
CHF	95	3204
IPF	39	4362
COPD	96	2463
Asthma	64	1118
Normals	187	1286

Table 4.2 Number of Crackles in Data Sets

4.1.3 Training Set Selection

In order to select a training set, we first make the assumption that all crackles are independent events. The basis for this assumption is that each crackle occurs as a result of an isolated physiological process. In Method 1, it does not matter which breath, or which patient a crackle occurred in for the purposes of making a test set.

To generate the training set for the pair-wise and multi-class comparisons, we ensured that an equal number of crackles for each disease were placed in the training set. We split the data in accordance with the following percentages: 70% Training, 15% Validation, and 15% Testing. To maintain the idea of training on an equal number of crackles for each disease, the percentages were taken from the disease with the fewest number of crackles. The idea behind training on a balanced training set is that no class receives preferential treatment by the learning algorithm due to its prevalence in the training set. In the case of an unbalanced training set, the algorithm may be able to achieve very low training error rates by classifying everything into the dominant class. However, this classifier will not be able to generalize well with the data withheld in the testing set.

One exception to this equal size rule is that the testing sets will not be of the same size. The test sets are composed of all samples not chosen to be a part of the training or validation sets. Since the sets are initially imbalanced, one test set may be of much larger size. However, this has no affect on the training of the algorithm. Figure 4.1 provides an illustration of the splitting process used in Method 1.

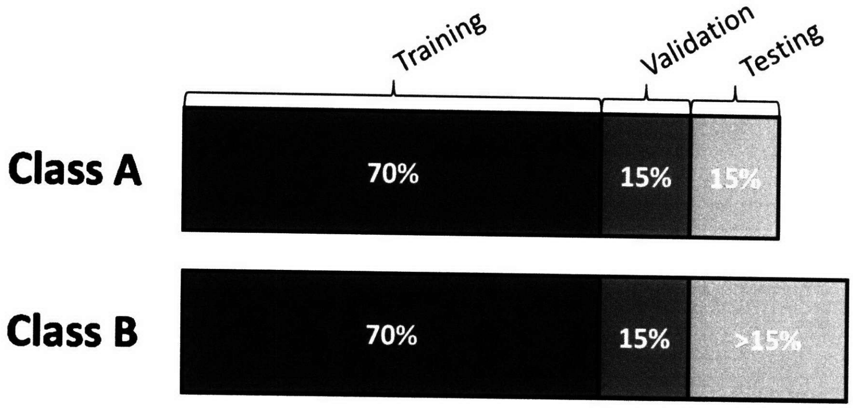


Figure 4.2 Splitting of Data to Form a Balanced Data Set

In some cases, we were not able to take the full 70% of the crackle data to train on. The large number of crackles for some diseases caused the computer to run into memory storage issues especially when conducting multiple trials. Therefore, instead of using a full 70%, a smaller percentage to split the data was chosen to speed up the training process yet still maintain a balanced training set.

4.2 Method 2 – Classifying Individual Breaths

In the second method of training the data set, we perform the analysis based on information that has been aggregated by breath. This methodology gives us the benefit of reducing the sample size to something manageable by the computer, but it doesn't reduce the sample size too much as would be the case if the patients were combined at the patient level. There are only 39 unique patients for the IPF data set, and this sample size would be very small for a machine learning analysis. Combining features into a breath level analysis maintains the ideals of being able to classify patients. Each breath is associated with the patient's disease and will be classified as such. As with Method 1, a voting schema will be applied in order to make a diagnosis on the patient level. Unlike Method 1, we use the full amount of acoustic data including wheezes, rhonchi, and squawks. Additionally, we derive several distribution features that define the location and distribution of crackles. All in all, this leads to many more features for the algorithms to process than were available in Method 1.

4.2.1 Feature Definitions

In Method 1, since there were only 22 features, we were able to describe them all easily within Table 4.1. Adding the features to characterize the squawks, wheezing, and rhonchi increases the number of acoustic features to 91. These will be fully listed in an appendix. One quick note about the additional sound features is that their features need to be defined at the patient level. As a result, special considerations need to be applied when portioning the data set. Otherwise, identical feature vectors could exist in both the training, validation, and testing phases which will produce overestimates of the accuracy.

The 22 features for each crackle are modified by taking the median value of all the crackles that occur within a breath. The median was used to calculate the central tendency of the crackles because it is a more robust estimate than the mean. Using a median would eliminate potential outliers from the data set.

On top of these 91 features are an additional 18 features that describe the distribution of the sounds across the chest. Figure 4.3 helps to describe the distribution. In the figure, four quadrants are labeled Top Left (TL), Top Right (TR), Bottom Left (BL), and Bottom Right (BR). Crackles that occur in the specific regions are counted and utilized as a new feature.

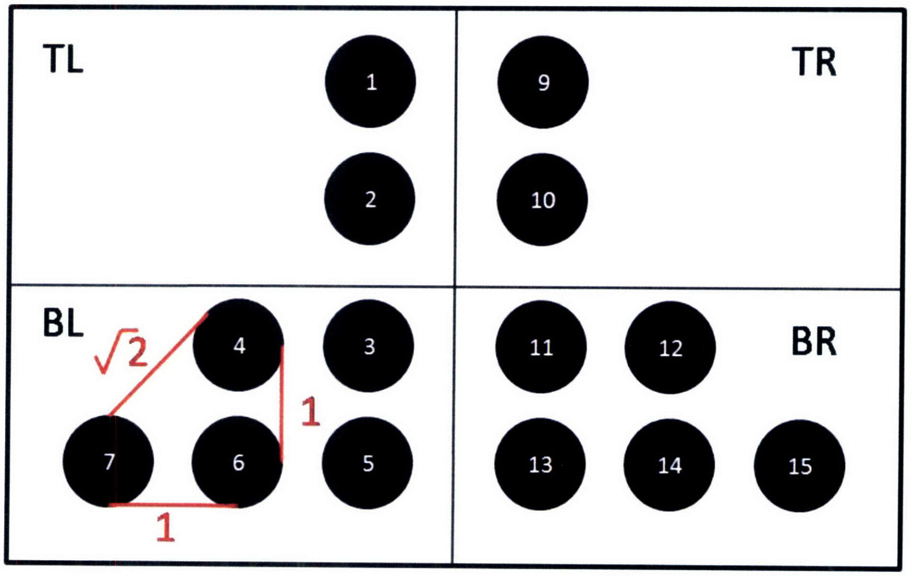
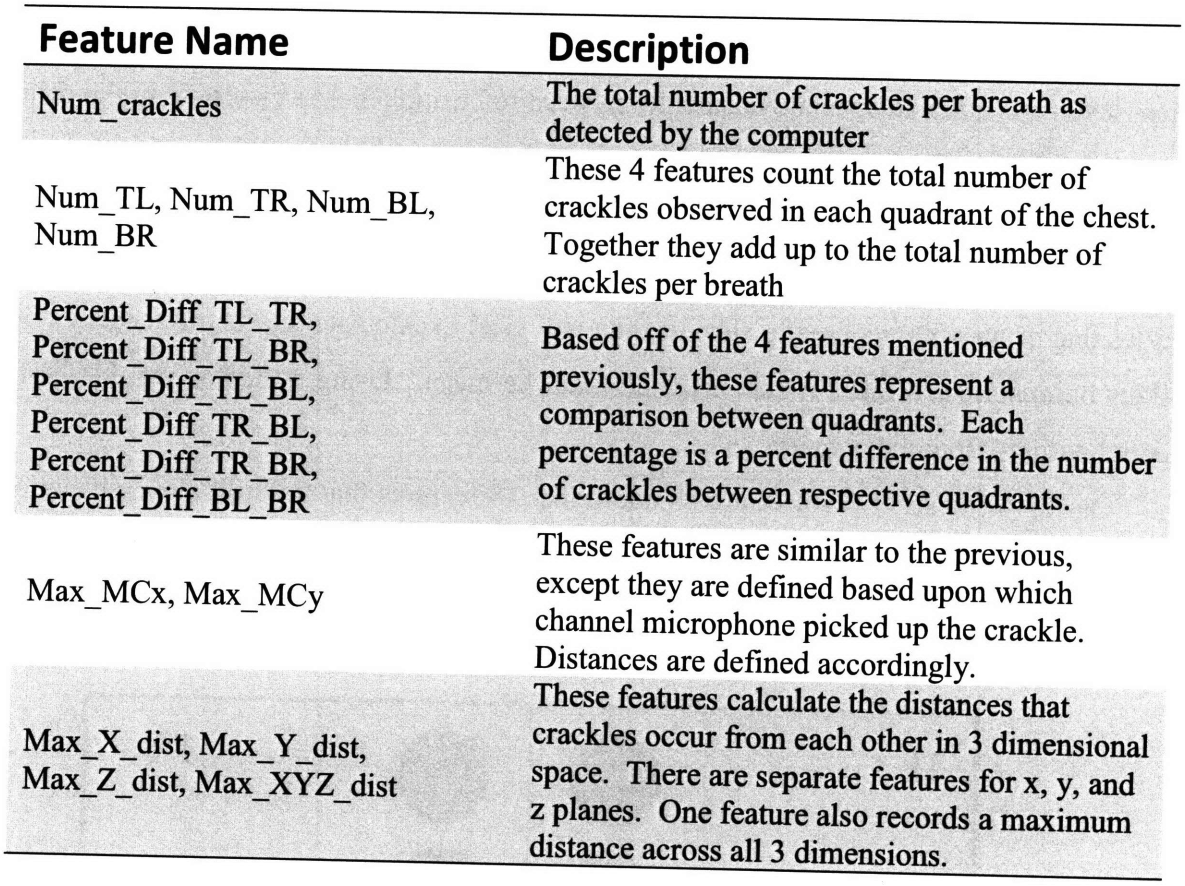


Figure 4.3 STG-1602 Drawing for Distribution Features

Another thing seen on the picture is the distances used to measure an artificial distance between the channels. The goal here is to be able to calculate the maximum distance between crackles as recorded by the mother channel. An imaginary grid is placed over the microphones and each microphone is said to be one unit apart horizontally and vertically. Therefore a diagonal distance has measurement $\sqrt{2}$ as can be seen in Figure 4.3. Specific definitions of the distribution features can be given in Table 4.3.



Feature Name	Description
Num_crackles	The total number of crackles per breath as detected by the computer
Num_TL, Num_TR, Num_BL, Num_BR	These 4 features count the total number of crackles observed in each quadrant of the chest. Together they add up to the total number of crackles per breath
Percent_Diff_TL_TR, Percent_Diff_TL_BR, Percent_Diff_TL_BL, Percent_Diff_TR_BL, Percent_Diff_TR_BR, Percent_Diff_BL_BR	Based off of the 4 features mentioned previously, these features represent a comparison between quadrants. Each percentage is a percent difference in the number of crackles between respective quadrants.
Max_MCx, Max_MCy	These features are similar to the previous, except they are defined based upon which channel microphone picked up the crackle. Distances are defined accordingly.
Max_X_dist, Max_Y_dist, Max_Z_dist, Max_XYZ_dist	These features calculate the distances that crackles occur from each other in 3 dimensional space. There are separate features for x, y, and z planes. One feature also records a maximum distance across all 3 dimensions.

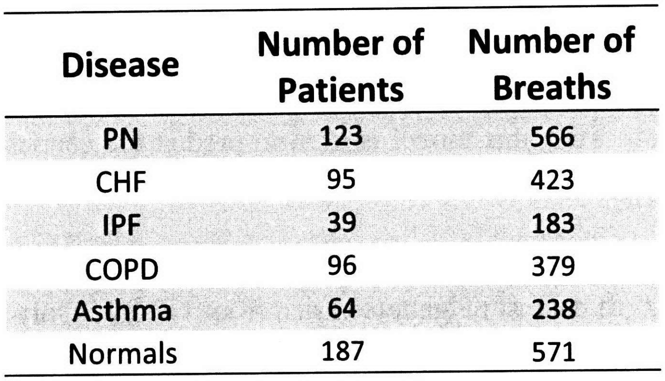
Table 4.3 Definitions of Distribution Features

The last row in the table describes four more features. These features were derived from the 3D Visualization shown in Figure 2.2 and try and capture the maximum spreads of the crackles in 3D space.

4.2.2 Data Pre-Processing and Standardization

Data was read from the same data files as with Method 1, so no further cleaning of the data was required. Some patients did not display any crackles, but they had other adventitious lung sounds. As a result, they could only be represented as one breath since that is how those sounds were modeled by Stethographics. Furthermore, all crackle features and distribution features will be "0" for the breaths without any crackles.

All data was standardized by using the same method presented in Section 4.1.2. The number of breaths for each disease is shown in Table 4.4.



Disease	Number of Patients	Number of Breaths
PN	123	566
CHF	95	423
IPF	39	183
COPD	96	379
Asthma	64	238
Normals	187	571

Table 4.4 Number of Breaths in Data Sets

4.2.3 Training Set Selection

Selecting a training set occurred in a slightly different manner for Method 2. Instead of simply selecting 70% of the breaths for training as in Method 1, we needed to take special care to ensure that all breaths belonging to a patient ended up in the same set whether that is training, validation, or testing. The reasoning behind this is that we can no longer make the independence assumption. Individual crackle events are independent, but the other lung sounds are calculated per patient. This means that identical feature vectors are repeated for each breath. These are obviously not independent. Therefore, the training sets must be designed in a way as to keep all breaths together that come from the same patient.

To partition these breath samples into the sets we need, we still prefer to keep the 70/15/15 ratio we had earlier. This time, to implement the splitting, we take 70% of the number of the patients, and then put all breaths associated with that patient into the training set. The same goes for both validation and testing sets. Ideally, the number of breaths in each set will be somewhat equal since the number of patients in each set is the same. Still, there are no guarantees that the data will be split equally with this approach.

4.3 Validation and Testing

Although it hasn't been formally defined yet within this thesis, validation is an important step towards developing a model that generalizes well. In Figure 3.4, validation falls under the action block of "Tune Parameters." Validation is the process of varying the model parameters in order to fine tune the models predictive ability. After a model is trained using the training data, the generalization performance is measured based off of the validation data set. At no point does the model ever "see" the data withheld for testing. After many iterations and variations, a final parameter is decided on. Then this model is used to predict the classes of the data completely withheld in the testing set.

Each of our three algorithms had several parameters that required tuning. For k nearest neighbors, the number k ,of nearest neighbors varied from 1 to 17. Only odd numbers for k were used to prevent ties. Neural networks required the modification of the number of nodes in the hidden layer. Finally, support vector machines required the validation of the cost parameter C, the type of kernel used, and also parameters associated with the kernels such as the parameter p for the polynomial kernel in Equation 3.18 and the parameter σ for the width of the RBF kernel. We implemented cross validation manually, that is changing the parameters by hand, and then comparing results. This process was not automated in our learning algorithms. The following table lists our validation choices for the aforementioned parameters.

Algorithm	Parameter	Possible Values
k Nearest Neighbors	k	1, 3, ..., 17
Neural Networks	# of Neurons	10,15,20, ..., 40
SVM	C	{0.1, 0.25, 0.5, 1, 1.5, 5, 10}
polynomial	p	1, 2, 3
RBF	σ	{0.1, 0.15, 0.2, 0.3}

Table 4.5 Parameter Values for Validation

After the best parameters for each method are selected, the model is put to the test by classifying the testing data. The tests are repeated 50 times so that an average performance metric can be estimated. The datasets are the exact same ones as used in the validation process. All results are discussed in Chapter 5.

4.4 Voting Schema

One further item is the application of a voting scheme. Although predicting which disease a crackle resembles is useful, it is much more beneficial to be able to make a recommendation to a doctor as to what disease a patient might have. To make a recommendation at the patient level we rely on a voting scheme. The voting scheme uses a "majority wins" rule to extend the classifications of crackles and breaths to that of patients. In Method 1, the classification of individual crackles, we seek to diagnose all patients that have at least one crackle in the testing set. Every crackle in the testing set casts a "vote" for the disease it most closely resembles. A patient is diagnosed by the disease that gets the most votes. In Method 2, a very similar approach is taken but instead of voting by each crackle, the votes are cast by breaths. A patient is assigned to a disease class based on the total number of votes cast for the specific disease by a breath. This novel approach allows us to extend the pair-wise crackle and breath tests to a patient level diagnosis.

4.5 Computer Implementation Notes

We first implemented all tests via an open source software known as Weka [44]. Although important benchmarks were achieved in this software, in order for us to achieve the repeatability we desired and also the large number of training runs, we implemented our methodology in MATLAB. The algorithms *k* nearest neighbors and neural networks were implemented through internal toolboxes within MATLAB. A more efficient implementation of the SVM algorithm is provided through the SVM^{light} program developed by Thorsten Joachims [45]. An interface to use the program in MATLAB is provided by Tom Briggs [46]. Weka also was used as a check for the models developed in MATLAB.

4.6 Summary

This chapter has shown how we connect the data mining and machine learning techniques presented in Chapter 3 to the computerized auscultation presented in Chapter 2. We discussed the derivation of several features, and also discussed our validation and testing approaches. We also presented two pair-wise methodologies that will be used to explore the process of diagnosing a patient based off of computerized auscultation. The first seeks to classify individual crackles, and the second classifies breaths. We also introduced a voting schema that will be used to make diagnoses at a patient level.

Chapter 5

Results and Discussion

In this chapter, we present the results and demonstrate the feasibility of computerized auscultation. We will show that we have achieved good recognition performance through pair wise comparisons between diseases. We also will show that interstitial pulmonary fibrosis patients and asymptomatic patients are well separated when performing these binary comparisons. Furthermore, we show that IPF and normals can be distinguished very easily from all other diseases in a multi-class classification. We will also discuss combining clinical data with the acoustic data as a way to improve performance.

We begin the chapter by defining the metrics we use to summarize our findings. We then present results for the classification of the individual crackles and also the classification of breaths. The results are presented for both pair-wise and multi-class classifications.

5.1 Classification Metrics

Before we present the results, we first must explain some terminology that we use to compare the different classifiers and ultimately gauge the overall performance of the computerized auscultation. The following performance metrics will be introduced here: *sensitivity, specificity,* and *accuracy.*

To introduce these terms, we provide a brief example. Suppose we want to set up a test that differentiates patients as either having pneumonia or being normal. PN will be the positive class, and normal will be the negative class. If a patient has PN and is predicted to have PN, it is a *True Positive* (TP). However, if the patient is misdiagnosed, it is a *False Negative* (FN). If a patient is normal and classified as such, it is called a *True Negative* (TN). Similarly, if the normal patient is predicted to have pneumonia, it is a *False Positive* (FP). Accuracy measures the total number of correct predictions out of the entire tested population. It is simply *1-error*, with error as defined in Chapter 3. However, this metric may not be the best measurement since it can be skewed by the amount of each class in the testing set. For instance, if there were 90 PN patients and 10 normal patients in a test set, a classifier could achieve a seemingly good accuracy of 90% by classifying all patients as PN. 100% of the PN patients were classified correctly, but 0% of the normal patients.

Accuracy is still a useful metric, but more detailed measurements are required to ensure the classifier is balanced; that is predicts both classes with similarly good performance. For this we turn to sensitivity and specificity. Sensitivity is the proportion of all positive patients that tested positive to the total number of positive patients in the study. It is expressed as:

$$sensitivity = \frac{TP}{\mathrm{TP+FN}} \tag{5.1}$$

Similarly, specificity measures the ratio of all negative patients classified as such to the total number of negative patients. More explicitly:

$$specificity = \frac{TN}{\mathrm{TN+FP}} \tag{5.2}$$

These two measures will be used to ensure we do not have any unbalanced classifiers.

5.2 Method 1 – Classifying Individual Crackles

The results of the pair-wise comparisons between diseases using the defined metrics are shown in Table 5.1. The results in this section are for Method 1. Method 1 focuses on classifying the individual crackles whereas all other sounds are left out. Comparing IPF vs. PN yields the same result as PN vs. IPF, so there are fifteen total comparisons to run. In the table, each of the numbers is the average classification performance of the respective classifiers on fifty randomly generated data sets. Fifty tests were considered to get a good estimate on the actual performance of the classifier. All three learning algorithms (support vector machines, neural networks, and *k* nearest neighbors are presented. To read the table, the positive class is listed in the first column and the ability of the algorithm to predict that disease is associated with sensitivity. For the negative class, it is listed second and associated with specificity. All results are color coded for viewing convenience. A red tint indicates good performance, yellow mediocre performance, and blue poor performance.

		SVM			Neural Networks			k Nearest Neighbor		
Positive Class	Negative Class	Sens	Spec	Acc	Sens	Spec	Acc	Sens	Spec	Acc
IPF	PN	0.798	0.715	0.743	0.767	0.729	0.743	0.775	0.733	0.747
	CHF	0.799	0.726	0.777	0.753	0.691	0.732	0.767	0.734	0.757
	Asthma	0.798	0.768	0.797	0.783	0.724	0.778	0.777	0.735	0.773
	COPD	0.802	0.713	0.789	0.735	0.738	0.736	0.759	0.694	0.746
	Normals	0.742	0.747	0.742	0.741	0.751	0.742	0.74	0.655	0.732
PN	CHF	0.484	0.684	0.525	0.563	0.614	0.576	0.583	0.613	0.59
	Asthma	0.662	0.554	0.658	0.625	0.587	0.623	0.619	0.626	0.619
	COPD	0.478	0.743	0.516	0.581	0.636	0.590	0.584	0.62	0.589
	Normals	0.705	0.727	0.707	0.707	0.715	0.708	0.656	0.638	0.655
CHF	Asthma	0.663	0.611	0.660	0.612	0.574	0.608	0.647	0.616	0.644
	COPD	0.548	0.593	0.562	0.534	0.610	0.559	0.6	0.6	0.6
	Normals	0.721	0.646	0.715	0.675	0.662	0.673	0.649	0.621	0.646
Asthma	COPD	0.592	0.675	0.667	0.621	0.613	0.619	0.587	0.637	0.63
	Normals	0.631	0.739	0.703	0.705	0.598	0.641	0.617	0.663	0.647
COPD	Normals	0.681	0.604	0.671	0.638	0.600	0.630	0.629	0.601	0.625

Table 5.1 Complete Results for Classifying Individual Crackles

Idiopathic pulmonary fibrosis crackles were separated the best out of all six classes. A crackle is correctly identified as IPF nearly 80% of the time. These numbers are seen in the first five rows. The ability to diagnose IPF acoustically is consistent with previously reported opinions [23]. Separating IPF crackles with high accuracy is very important within the medical community since it is a very rare disease and as a result it is commonly misdiagnosed as another pulmonary condition. As a result, this terminal condition could set in even faster without appropriate treatments.

Furthermore, asymptomatic patients (normals) were classified fairly well. They can be separated from IPF, PN, CHF, and asthma with over 70% accuracy using support vector machines. Normals separate from COPD with 67.1% accuracy. This is somewhat surprising because adventitious lung sounds aren't normally associated with "healthy" patients. Still, the adventitious lung sounds are present and the crackles provide enough subtle differences to make a classification. CHF, COPD, and asthma were much harder to classify. This is largely because these diseases are associated more so with wheezes and rhonchi than with crackles. The worst comparisons by far were PN vs CHF (52.5% accuracy for SVM) and PN vs COPD (51.6% accuracy for SVM).

We compared the three algorithms and found that they all perform somewhat comparably. Figure 5.1 compares the sensitivities and specificities of support vector machines, *k* nearest neighbors, and neural networks for the IPF vs. COPD comparison.

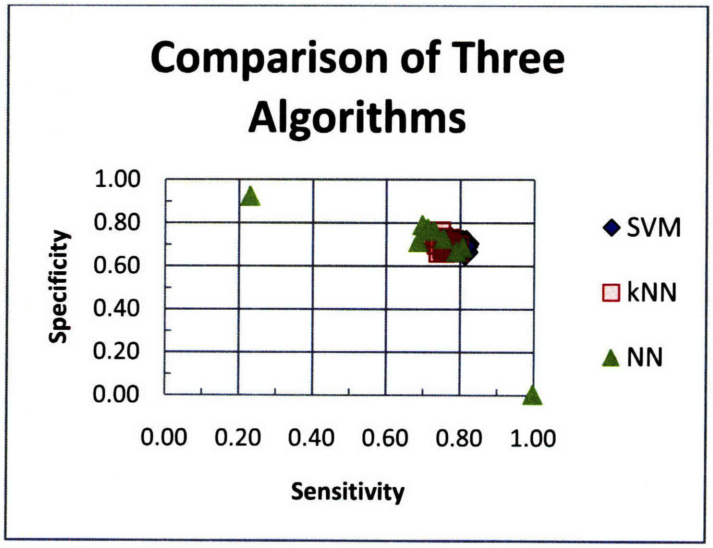


Figure 5.1 Sensitivity and Specificity of 10 Data Sets for IPF vs COPD Comparison

Each point in the figure represents one individual classification's sensitivity and specificity. There are ten points for each algorithm which represent ten different trained models. Each algorithm was trained and tested on the same data sets. There are several test runs where the neural network does not converge and results are very erratic. Other than that, the performance of support vector machines, neural networks, and *k* nearest neighbor are fairly close to each other.

To further explore the training error with neural networks we plot the training points associated with the two node output layer. Figure 5.2 shows two graphs generated from these output nodes. In the training of the neural network, we use a unary encoding scheme according to the literature [37]. With a two-neuron output layer, the network tries to model the positive class as [1 0] and the negative class as [0 1]. As the network trains, the classification points move towards the respective output pair. This formulation makes for a convenient 2-D plot. The graph on the left is an example of a neural network that converged. Each point represents a single crackle classification. They are very small since there are over 2,000 points in each class. The decision boundary is represented by the thick black line. Ideally, all blue points would lie to the right of this line and the red points would lie to the left. Instead of achieving a clean separation, the right network is erratic and the placement of the points makes no sense. As a result, all neural networks that appeared in this manner were disregarded in the calculation of the average prediction performance in Table 5.1 and in the rest of the chapter.

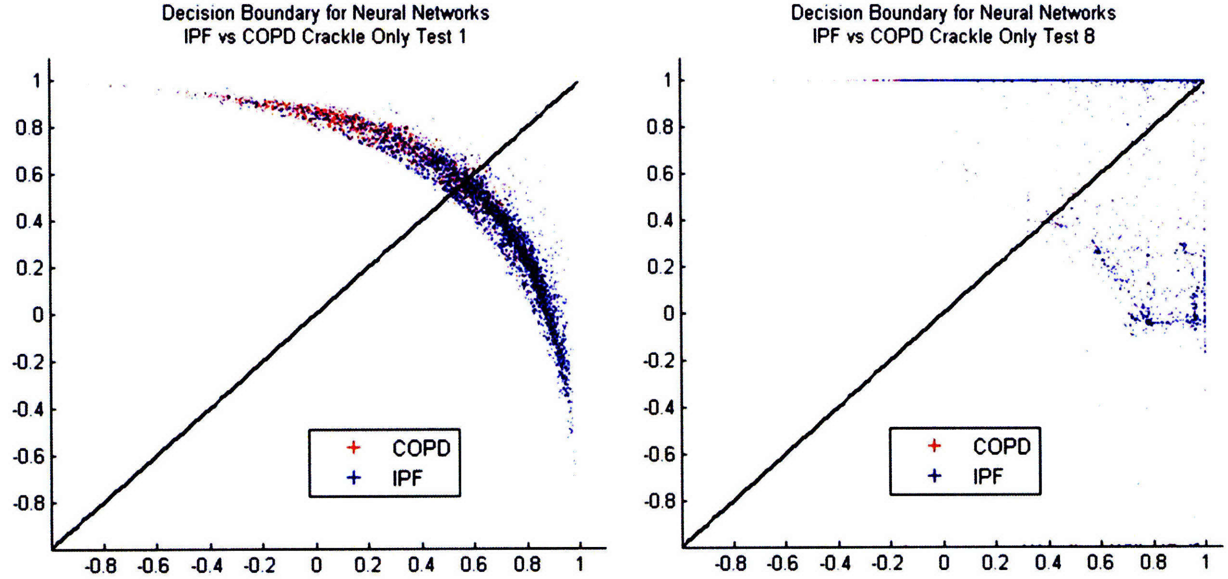
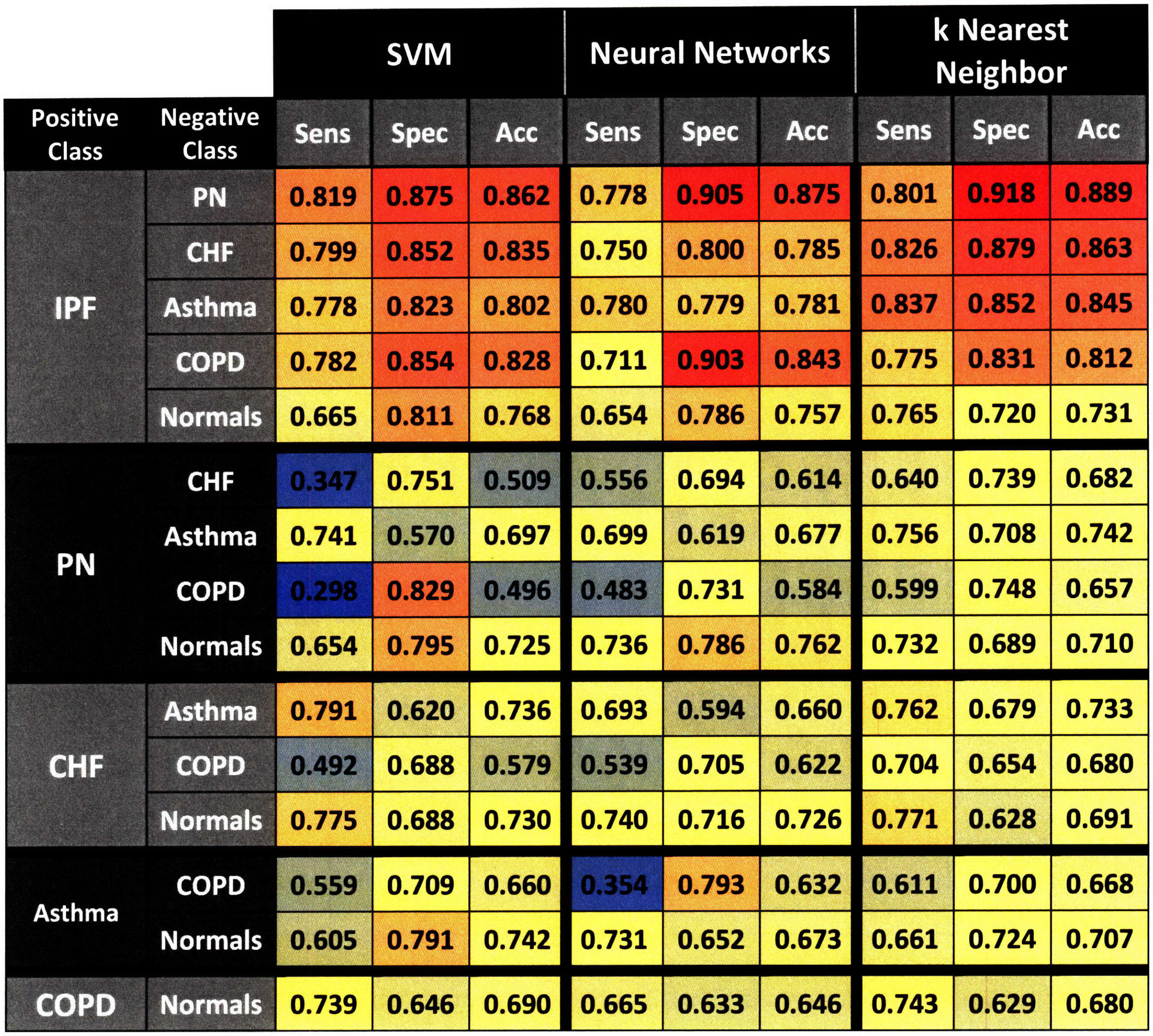


Figure 5.2 Two Neural Networks, One of Which Does Not Converge

5.3 Diagnosis of Patients by Voting

Here we discuss the voting mechanism originally introduced in Section 4.4 which will be used to make predictions at a patient level. We extend the classification of the individual crackles to a more useful patient diagnosis. With the decent performance of the individual crackle classifications, we show even better performance when predicting the diagnosis of a patient. Patients who have crackles exhibit some that possess characteristics of a certain lung disease whereas others may be indistinguishable. The expectation is that the majority of the crackles contain some distinguishable information that increases separation by using voting. The results confirm our assumptions and are shown in Table 5.2 below.



		SVM			Neural Networks			k Nearest Neighbor		
Positive Class	Negative Class	Sens	Spec	Acc	Sens	Spec	Acc	Sens	Spec	Acc
IPF	PN	0.819	0.875	0.862	0.778	0.905	0.875	0.801	0.918	0.889
	CHF	0.799	0.852	0.835	0.750	0.800	0.785	0.826	0.879	0.863
	Asthma	0.778	0.823	0.802	0.780	0.779	0.781	0.837	0.852	0.845
	COPD	0.782	0.854	0.828	0.711	0.903	0.843	0.775	0.831	0.812
	Normals	0.665	0.811	0.768	0.654	0.786	0.757	0.765	0.720	0.731
PN	CHF	0.347	0.751	0.509	0.556	0.694	0.614	0.640	0.739	0.682
	Asthma	0.741	0.570	0.697	0.699	0.619	0.677	0.756	0.708	0.742
	COPD	0.298	0.829	0.496	0.483	0.731	0.584	0.599	0.748	0.657
	Normals	0.654	0.795	0.725	0.736	0.786	0.762	0.732	0.689	0.710
CHF	Asthma	0.791	0.620	0.736	0.693	0.594	0.660	0.762	0.679	0.733
	COPD	0.492	0.688	0.579	0.539	0.705	0.622	0.704	0.654	0.680
	Normals	0.775	0.688	0.730	0.740	0.716	0.726	0.771	0.628	0.691
Asthma	COPD	0.559	0.709	0.660	0.354	0.793	0.632	0.611	0.700	0.668
	Normals	0.605	0.791	0.742	0.731	0.652	0.673	0.661	0.724	0.707
COPD	Normals	0.739	0.646	0.690	0.665	0.633	0.646	0.743	0.629	0.680

Table 5.2 Voting Results for Crackle Only

For classes such as IPF where the individual crackle classification was good, the voting results perform even better. If roughly 75% of all crackles can be confirmed or rejected as being similar to the crackle form of an IPF crackle, the voting schema incorporates the high success rate and increases the chances of successfully diagnosing a patient. For example, in the IPF vs PN comparison, individual crackles were distinguished by support vector machines with 74.3% accuracy. Applying the voting increased the accuracy to 86.2%. On the contrary, in cases where the individual crackles were not as easily classified, the errors seem to magnify. This occurs in the PN vs CHF and also the PN vs COPD comparisons. Classification accuracy of PN vs. CHF drops from 52.5% to 50.9% with voting and drops from 51.6% to 49.6% in the PN vs COPD comparison. Since the crackles themselves possessed very little recognition, the voting schema could not help the classification performance. Still, the voting approach has shown a significant difference in crackle prediction performance. Figure 5.3 shows the changes in classification performance after the voting scheme has been applied to the output of the SVM model.

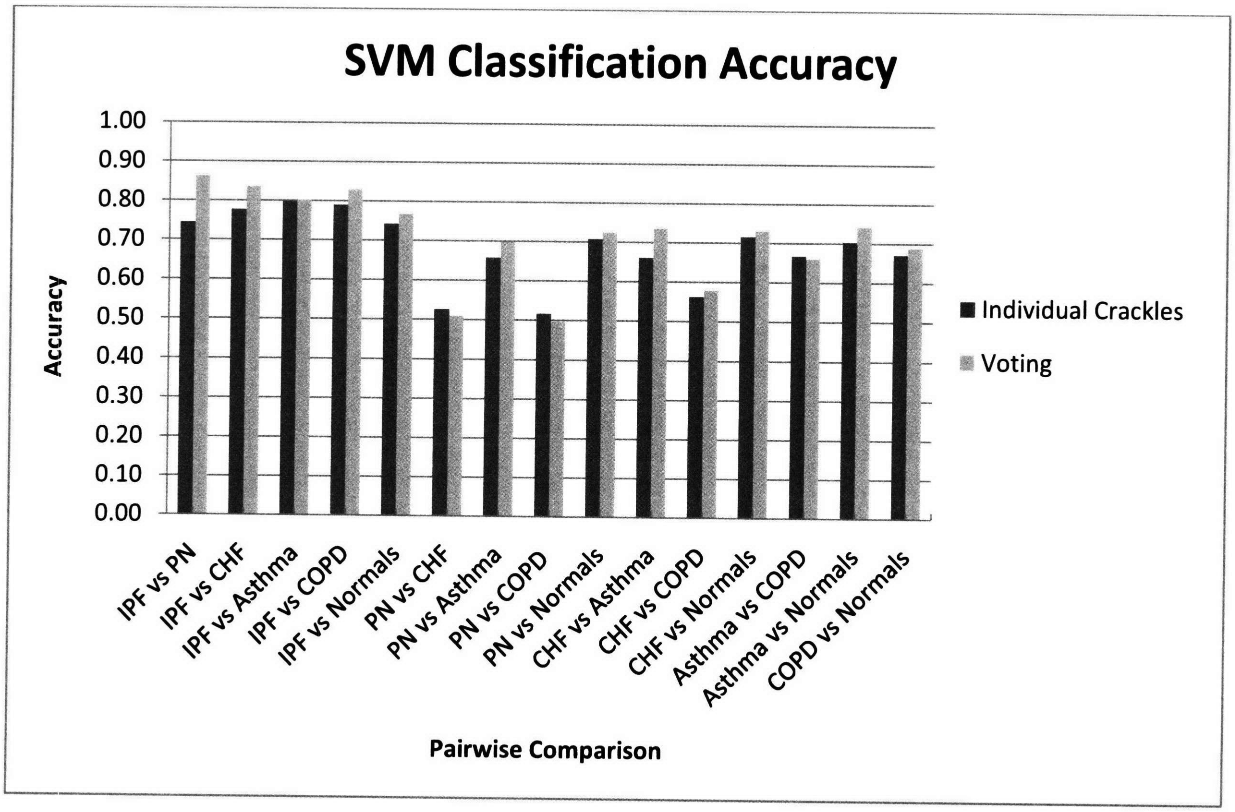


Figure 5.3 Changes in Classification Accuracy with Voting by Crackle

Although the accuracies shown in the figure are not necessarily high enough for complete diagnostic use, they can be used in conjunction with other methods for a doctor's final diagnosis. The highest classification accuracies at the patient level are in seen in all the IPF comparisons. It is accurately predicted at least 75% of the time with all algorithms.

One particular advantage of classification by using only crackles is that the listening device can likely be simplified. Instead of a multi-channel pad, a single stethoscope with a microphone embedded in it is able to capture the differences within the individual crackle sounds. This has the added benefit of being very familiar to a patient. A doctor would be able to apply the smart stethoscope in the same manner as a traditional stethoscope. It would not be much of a departure from normal medical practice so it could possibly ease a patient's transition towards computerized auscultation.

All in all, Method 1 provides good classification for IPF and fairly good classification for asymptomatic patients indicating that crackle features for these two diseases are distinct. On the contrary, other diseases are much harder to classify by using crackles. PN and CHF are two diseases that are known for having crackles, but still their results indicate they are almost indistinguishable using this analysis. Diseases like asthma and COPD still have crackles, but their dominating features are that of wheezes and rhonchi. Incorporating these features will aid in diagnosing theses diseases.

5.4 Method 2 – Classification by Using Breath Analysis

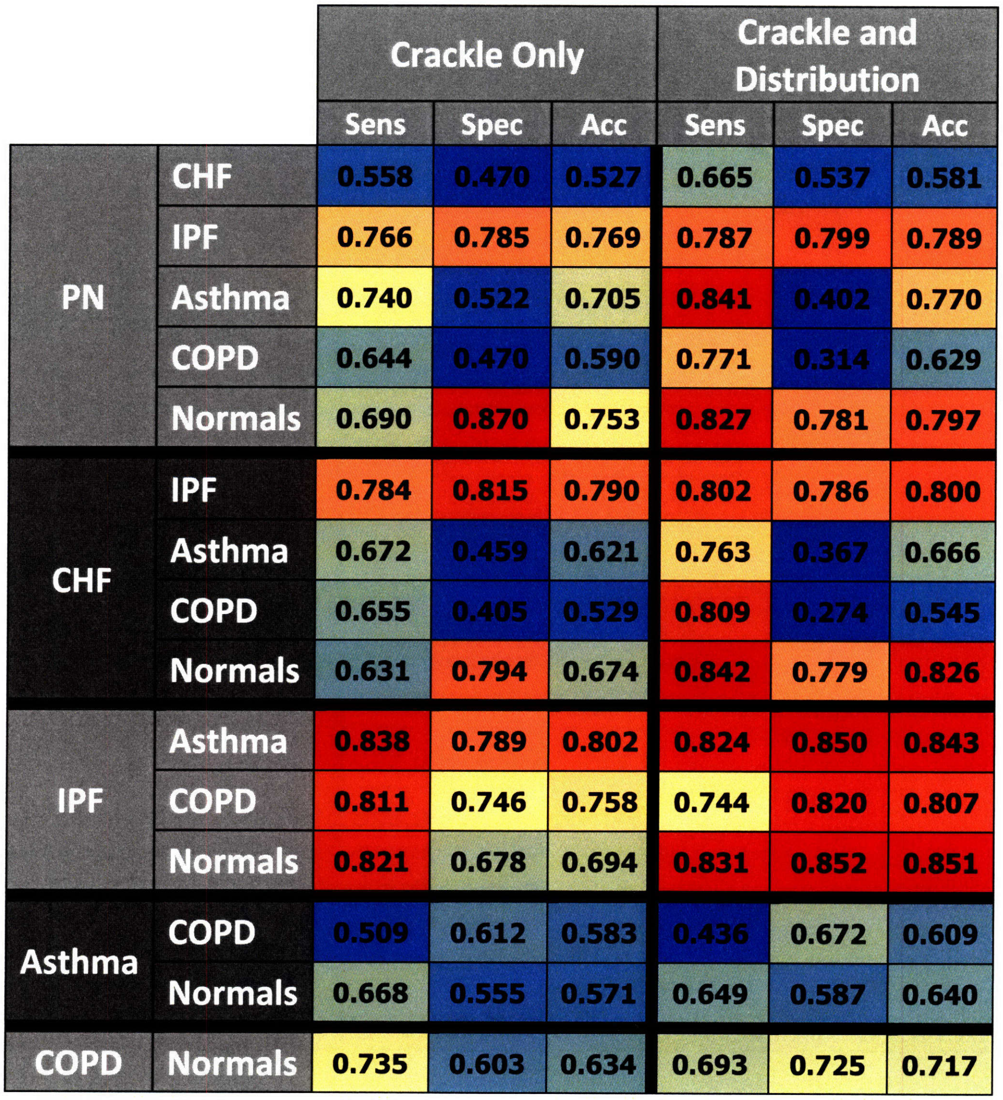
We now shift our focus to the second method of analysis, combining features per breath. We will make classifications based on full breaths instead of just individual crackles as before. This time, although we still continue to use neural networks and *k* nearest neighbors, we shift our focus to that of support vector machines. We do this because they seemed to have similar performance capabilities in Method 1, but the support vector machines took much less time to compute. When we add the full adventitious lung sounds data set, the number of features goes from 22 to 107. Training hundreds of neural networks with 107 features would take a very long time.

To study the effectiveness of classification of breaths, we perform multiple training runs on feature subsets of the full adventitious lung sound data. Doing this incrementally provides insights as to which data are the most important in making a classification. The subsets are listed in the Table 5.3.

Subset Name	Number of Features	Description
Crackle Only	23	This set contains all variables in Method 1, but combines them as a median according to their breath.
Crackle and Distribution	41	This set contains the Crackle Only set and also the distribution features mentioned in Section 4.2.1.
Additional sounds	66	This set contains all acoustic features not used in Method 1 and are typically computed as an average at the patient level. Features represent crackles, wheezes, rhonchi, and squawks.
Full Data	107	This is a combination of the all previously mentioned sets.

Table 5.3 List of Data Subsets Used to Test Method 2

The results are presented in Table 5.4 and Table 5.5 in a similar fashion as before. All fifteen pair-wise comparisons are shown in two consecutive tables. The first table contains the Crackle Only and Crackle and Distribution feature subsets. The second table has the Additional Sounds and the Full Data feature subsets. Tables are also provided in an appendix that shows a breakout of each class versus all five other classes which may be easier to read, but are too lengthy for inclusion here.

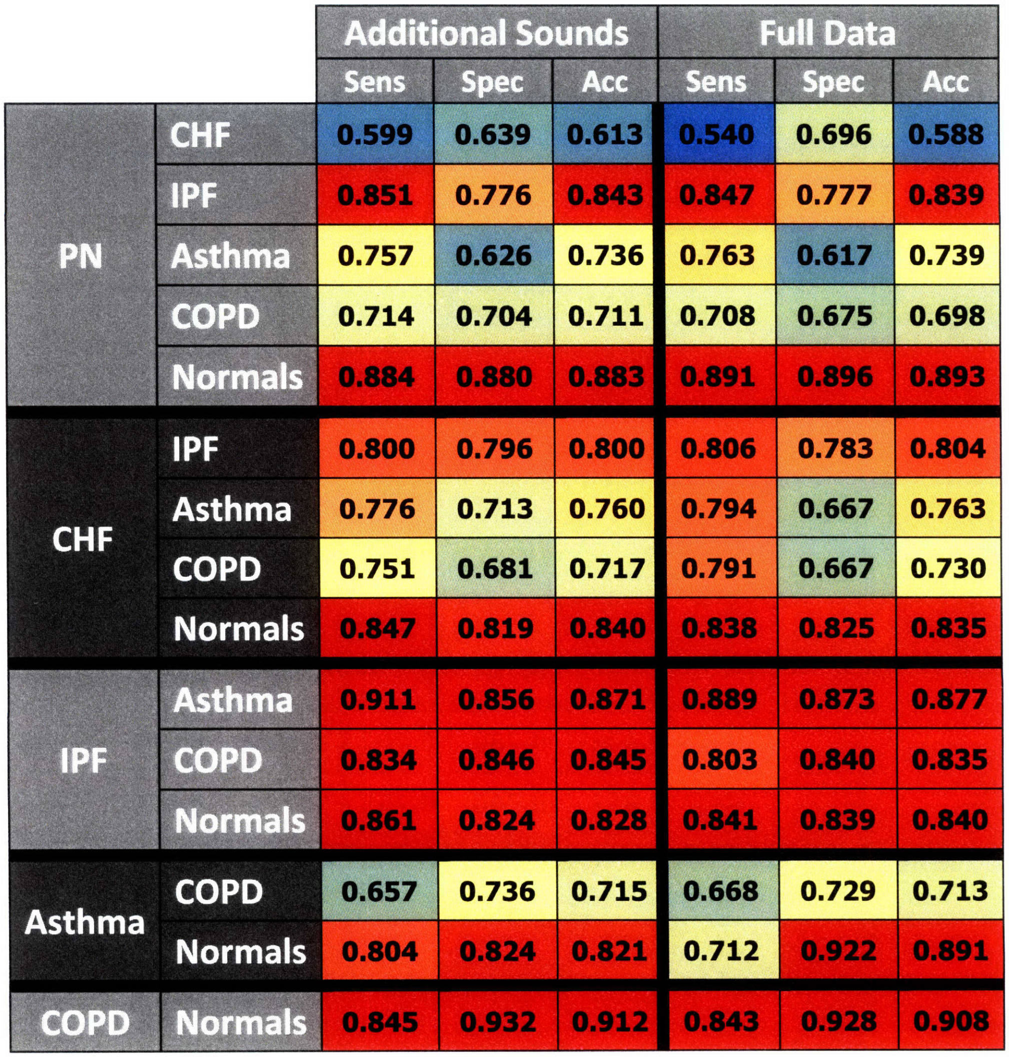


		Crackle Only			Crackle and Distribution		
		Sens	Spec	Acc	Sens	Spec	Acc
PN	CHF	0.558	0.470	0.527	0.665	0.537	0.581
	IPF	0.766	0.785	0.769	0.787	0.799	0.789
	Asthma	0.740	0.522	0.705	0.841	0.402	0.770
	COPD	0.644	0.470	0.590	0.771	0.314	0.629
	Normals	0.690	0.870	0.753	0.827	0.781	0.797
CHF	IPF	0.784	0.815	0.790	0.802	0.786	0.800
	Asthma	0.672	0.459	0.621	0.763	0.367	0.666
	COPD	0.655	0.405	0.529	0.809	0.274	0.545
	Normals	0.631	0.794	0.674	0.842	0.779	0.826
IPF	Asthma	0.838	0.789	0.802	0.824	0.850	0.843
	COPD	0.811	0.746	0.758	0.744	0.820	0.807
	Normals	0.821	0.678	0.694	0.831	0.852	0.851
Asthma	COPD	0.509	0.612	0.583	0.436	0.672	0.609
	Normals	0.668	0.555	0.571	0.649	0.587	0.640
COPD	Normals	0.735	0.603	0.634	0.693	0.725	0.717

Table 5.4 Crackle Only and Crackle and Distribution Datasets for SVM by Breath

The per breath analysis on the crackle only and the crackle and distribution data sets leads to very similar results to that of Method 1. IPF and normals were separated the best out of all six classes. IPF could be separated from CHF and asthma with 80% accuracy and separated from COPD and PN with 75% accuracy with the Crackle Only feature subset. Adding distribution features yields nearly a 5% improvement for the comparison of IPF with asthma and COPD. Furthermore, adding the distribution features greatly aid the classification of IPF vs normals. Classification accuracy jumps from 69.4% to 85.1% for this comparison.

As in Method 1, COPD and asthma are very difficult to classify by using only crackles and their distribution about the chest. Adding the other adventitious sounds improve the classification accuracy as shown in Table 5.5.



		Additional Sounds			Full Data		
		Sens	Spec	Acc	Sens	Spec	Acc
PN	CHF	0.599	0.639	0.613	0.540	0.696	0.588
	IPF	0.851	0.776	0.843	0.847	0.777	0.839
	Asthma	0.757	0.626	0.736	0.763	0.617	0.739
	COPD	0.714	0.704	0.711	0.708	0.675	0.698
	Normals	0.884	0.880	0.883	0.891	0.896	0.893
CHF	IPF	0.800	0.796	0.800	0.806	0.783	0.804
	Asthma	0.776	0.713	0.760	0.794	0.667	0.763
	COPD	0.751	0.681	0.717	0.791	0.667	0.730
	Normals	0.847	0.819	0.840	0.838	0.825	0.835
IPF	Asthma	0.911	0.856	0.871	0.889	0.873	0.877
	COPD	0.834	0.846	0.845	0.803	0.840	0.835
	Normals	0.861	0.824	0.828	0.841	0.839	0.840
Asthma	COPD	0.657	0.736	0.715	0.668	0.729	0.713
	Normals	0.804	0.824	0.821	0.712	0.922	0.891
COPD	Normals	0.845	0.932	0.912	0.843	0.928	0.908

Table 5.5 Other Sounds and Full Datasets for SVM by Breath

Adding the remaining sounds (wheezes, rhonchi, squawks) to supplement the crackle features clearly improves classification performance. Most diseases can be distinguished with at least 75% accuracy. The PN and CHF comparison is the most difficult with accuracy only around 60%. Furthermore, COPD remains the most difficult disease to predict with accuracies around 70% for all pair-wise comparisons with the exception of normals and IPF.

One important trait of the SVM classifiers which can be seen in Table 5.5 is that the sensitivity and specificity are close to each other which indicate that the model is well trained. Because of this, accuracy can be used as a balanced assessment tool regardless of the number of breaths of each disease in the test set. Figure 5.4 plots the accuracies for all four data sets: Crackle Only, Crackle and Distribution, Additional Sounds, and Full Data.

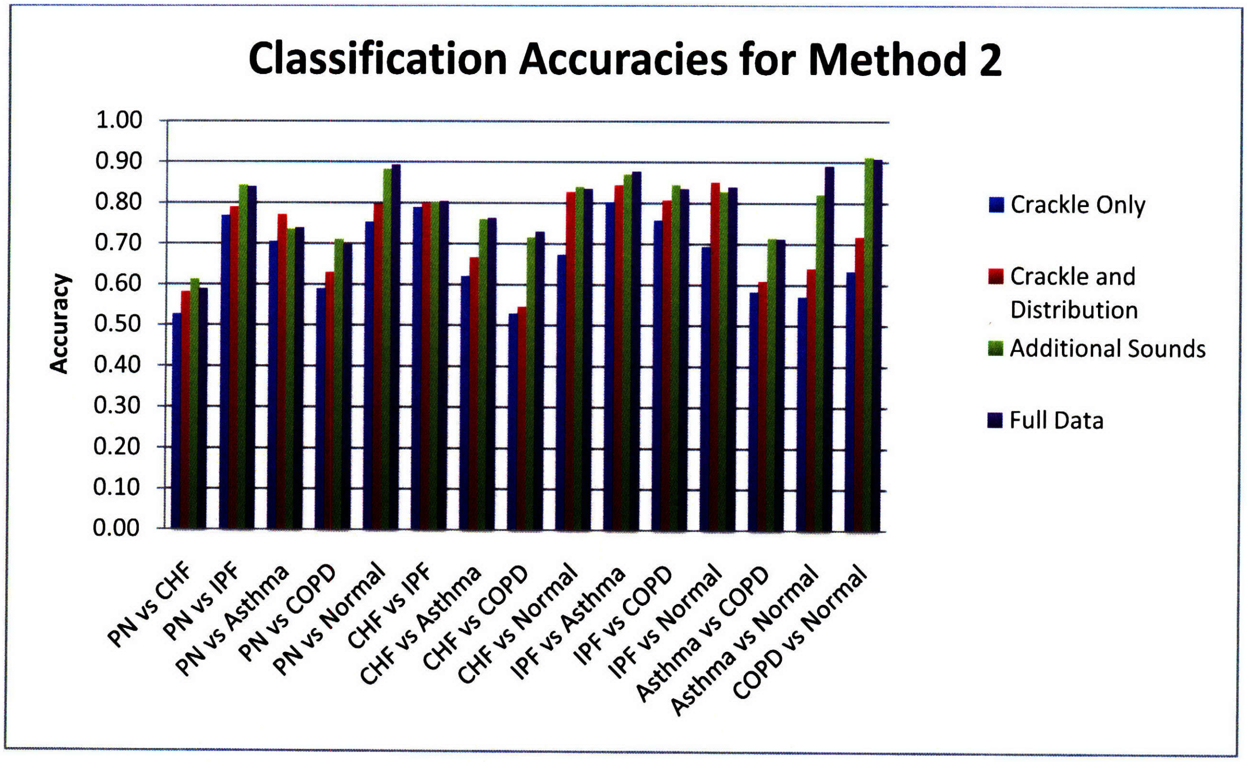
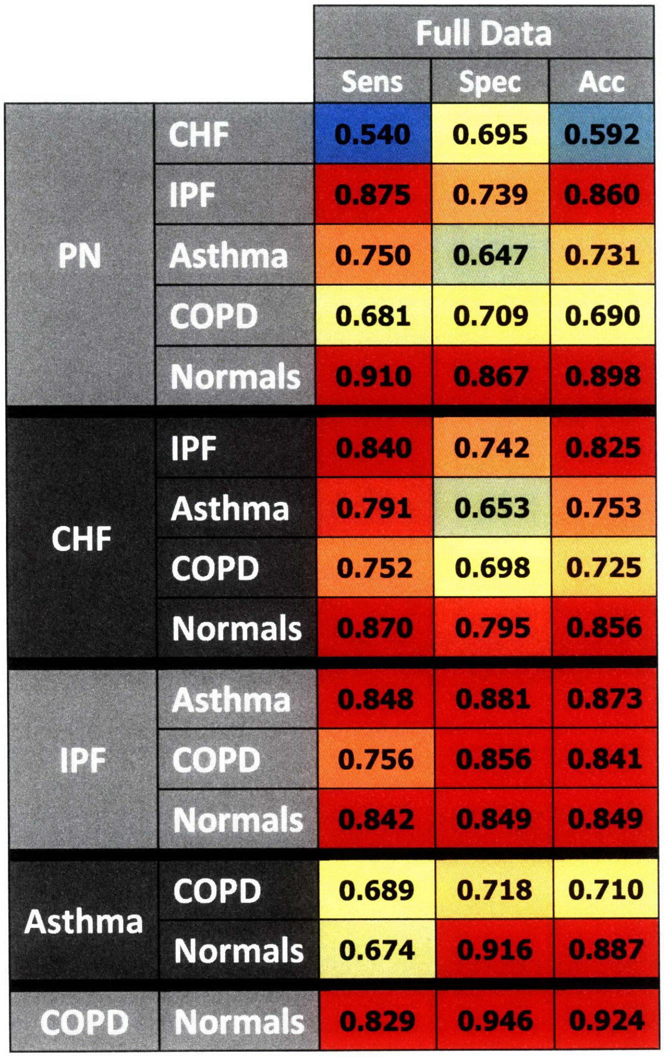


Figure 5.4 Pair-Wise Accuracies for Different Data Sets

Accuracies based on only the crackle features were by far the worst performing for all data sets. Adding distribution information features to the crackle data moderately helped the classification. The Additional Sounds data set and Full Data were by far the most important to make a diagnosis. With only a few exceptions, using all the sounds proved to be beneficial for the performance of the algorithm. The results show that all recorded adventitious lung sounds are important in making a lung disease diagnosis.

5.5 Method 2 – Voting Applied to Classification of Breaths

In the previous section, we show that classification based on the set of all adventitious lung sounds is much better than the data sets that rely on crackles only. We again implement the voting mechanism to further improve the classification accuracy of our models and to bring the diagnosis to the patient level. We present the voting results for the Full Data set here in Table 5.6 whereas the rest of the data sets are shown in the appendix.



		Full Data		
		Sens	Spec	Acc
PN	CHF	0.540	0.695	0.592
	IPF	0.875	0.739	0.860
	Asthma	0.750	0.647	0.731
	COPD	0.681	0.709	0.690
	Normals	0.910	0.867	0.898
CHF	IPF	0.840	0.742	0.825
	Asthma	0.791	0.653	0.753
	COPD	0.752	0.698	0.725
	Normals	0.870	0.795	0.856
IPF	Asthma	0.848	0.881	0.873
	COPD	0.756	0.856	0.841
	Normals	0.842	0.849	0.849
Asthma	COPD	0.689	0.718	0.710
	Normals	0.674	0.916	0.887
COPD	Normals	0.829	0.946	0.924

Table 5.6 Voting Results for Method 2 Performed on the Full Data Set

Voting is performed by summing up the total number of breaths in a patient that pertain to a certain disease. In the case of a tie (ie. three breaths PN and three breaths CHF), the patient is listed as an uncertain diagnosis, but less than 1% of all patients fell into this category. As with previous tests, IPF and asymptomatic patients are the ones that can be diagnosed best with this technology. IPF is routinely separated with over 85% accuracy. Performing even better is the classification of the normal patients. They are separated from COPD with accuracy of 92.4% and from PN with accuracy of 89.8%. Other diseases are much harder for the algorithms to predict because of their similarity in sound patterns. COPD and asthma is one such comparison. However, there still is some separation with classification accuracy of 71%. The PN and CHF pair-wise comparison remains the worst performing with an accuracy of only 59.2%.

Overall, voting did not have much of an impact when compared to the classification accuracies of the breaths. Both the individual breath classification and the voting results are shown in Figure 5.5.

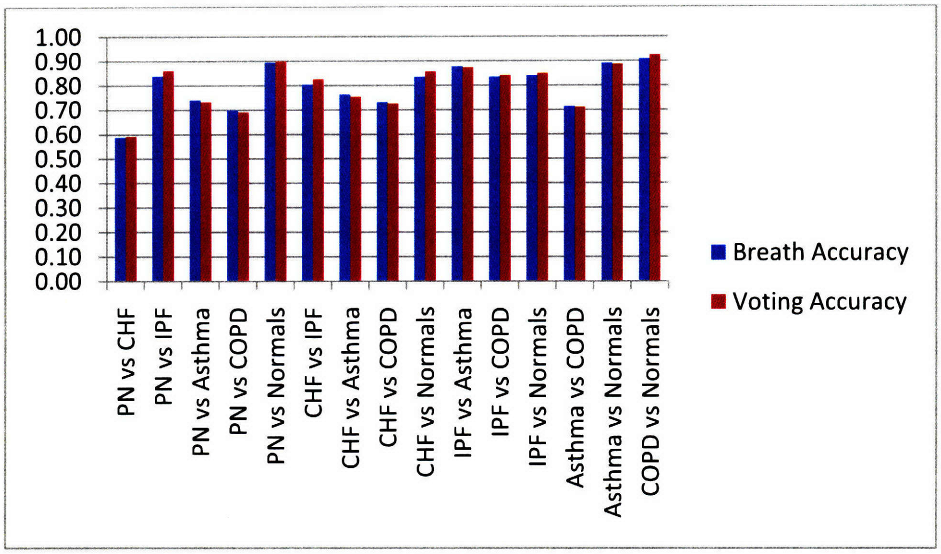


Figure 5.5 Breath Classification Accuracy and Voting Accuracies

For the most part, the classification was nearly the same. One possible explanation of the similarities in performance is that a lot of the votes are unanimous decisions. If enough of the patients have a unanimous vote, the differences in accuracy would be negligible. For example, if the test set only has two patients with six breaths each, all twelve breaths could be classified correctly. This classification performance is 100% accuracy in both the breath classification and in the voting. If this happens with many patients in the dataset, the results will be very similar as shown. Still, the voting methodology allows us to make diagnoses on a patient level so it remains a useful test.

5.6 Addition of Clinical Data

To further explore the available data, we add clinical features to the data set. These features are common measurements that a nurse collects as part of a patient's initial care or which can easily be obtained. We add the following features to the dataset:

- Age
- Gender
- Temperature
- Heart Rate
- Respiratory rate (RR)
- Systolic and diastolic blood pressure
- Oxygen saturation levels
- Presence or absence of cough (productive or not)
- and white blood cell count

We add these features since a doctor would have similar information when making a diagnosis. Adding them into the design of the smart stethoscope will improve the diagnostic ability of the machine since it incorporates the sounds and traditional medical examinations. We still follow the same testing process as introduced in Section 5.4 with patients being summarized by breath. However, we perform it on a completely different data set. For these tests, we eliminated all patients without the clinical information attached to their file. Doing so greatly reduced the number of patients available for testing. Less than half of all the patients had full clinical information. The exact number of patients is shown in Table 5.7. By using only a subset of the clinical features and ignoring the respiratory rate and coughing features, we are able to increase the amount of patients by a fair amount. These numbers are also shown in Table 5.7.

Disease	Total Patients	Full Clinical	No RR or Cough
PN	123	60	87
CHF	95	24	45
IPF	39	2	2
COPD	96	29	47
Asthma	64	16	19
Normals	187	0	0

Table 5.7 Number of Patients with Clinical Data

Figure 5.6 shows the results of training the support vector machine classifiers for pair-wise comparisons of the clinical data set. We do not include the normals or the IPF patients in our testing since there are hardly any patients. We again study the full sound data as tested previously which includes all crackles, wheezes, rhonchi, squawks, and distribution features of the crackles around the chest. In the figure these tests are denoted as "All Sounds." The "Clinical Added" data sets contain the same patients, but with the added benefit of the clinical information. Also, “Larger Set” denotes trials performed without the features of respiratory rate and the coughing features so more patients could be used in training the algorithm.

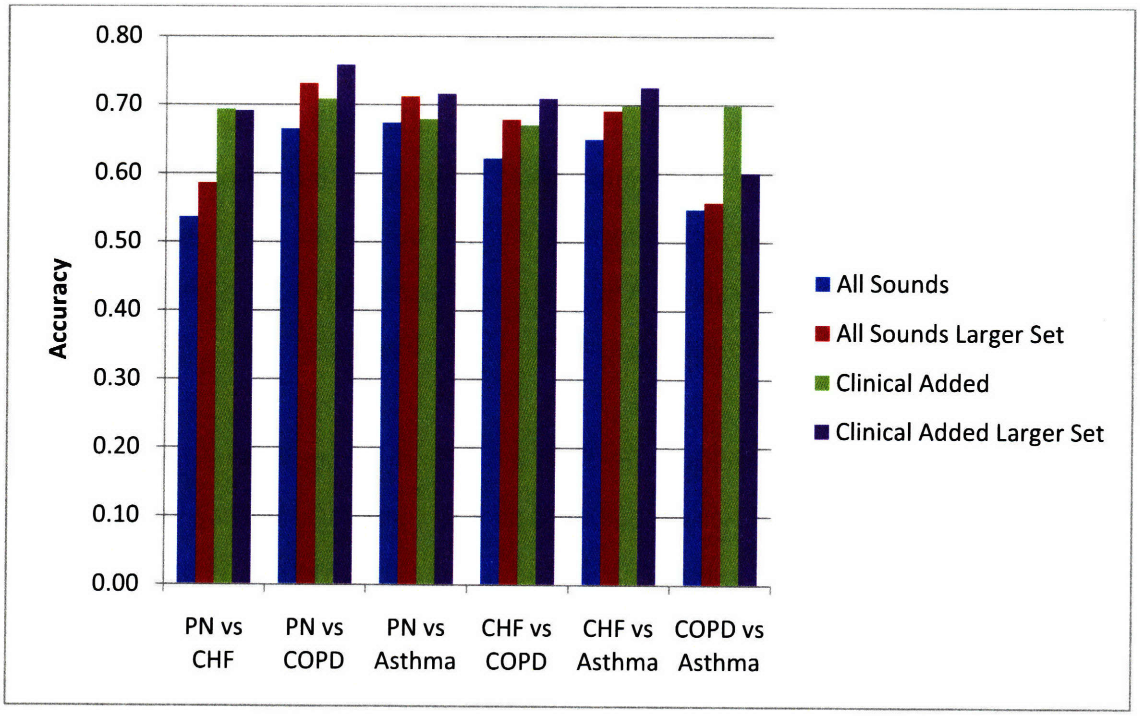


Figure 5.6 The Effect of Clinical Data on the Classification Performance

In the figure, the bars represent the average classification accuracy obtained by fifty classifications. The first two columns (blue and red) show the benefit of adding more patients to the training data set. By adding more patients, all pair-wise comparisons showed a significant improvement in performance. More importantly, incorporating the clinical information also adds significant improvements in performance of the classification. With the exception of the PN vs. asthma comparison, the remaining comparisons show a minimum improvement of 5% over the exact same data without the clinical information.

The computer that performs the auscultation could easily have a user interface where these parameters are input by a doctor or nurse so the algorithm could take advantage of them in predicting a patient’s disease. Doing so will improve the classification performance as we have shown here.

5.7 Multi-Class Classification

Extending the pair-wise comparisons to a multi-class comparison is an important step for medical practice. Instead of simply predicting one disease out of a choice of two, a multi-class diagnosis would help a doctor distinguish patients from all possible diseases. To perform the multi-class classification, we follow the multi-class formulations introduced in Section 3.5. We also maintained the same rules in selecting patients for the training set in order to keep the training set balanced as best as possible. A majority of the patients would end up in the test set since the small number of IPF patients would be a limiting factor. This may decrease performance marginally.

We used support vector machines (Section 3.4.3) and *k* nearest neighbors (Section 3.4.1) for our modeling since the neural networks would take too long to train on such large data sets. We apply the learning algorithms to both the individual crackles (Method 1) and also the patients by breath (Method 2). For Method 2, only the full data set was used since it generally outperformed the other datasets. All tests were performed by training the algorithms using the best parameters as found by validation in accordance with Section 4.3.

The results of the multi-class formulations are summarized in Figure 5.7. Only individual accuracies are displayed. In a multi-class scenario, sensitivity and specificity lose interpretability since there is no distinct positive or negative class. The left chart shows the multi-class performance of the classification of individual crackles. The right chart shows the classification of the breaths. The blue bars show the performance of the support vector machine one-vs-one method, the red bars are for the support vector machine one-vs-all accuracy, and finally the green bars represent the *k* nearest neighbor accuracy.

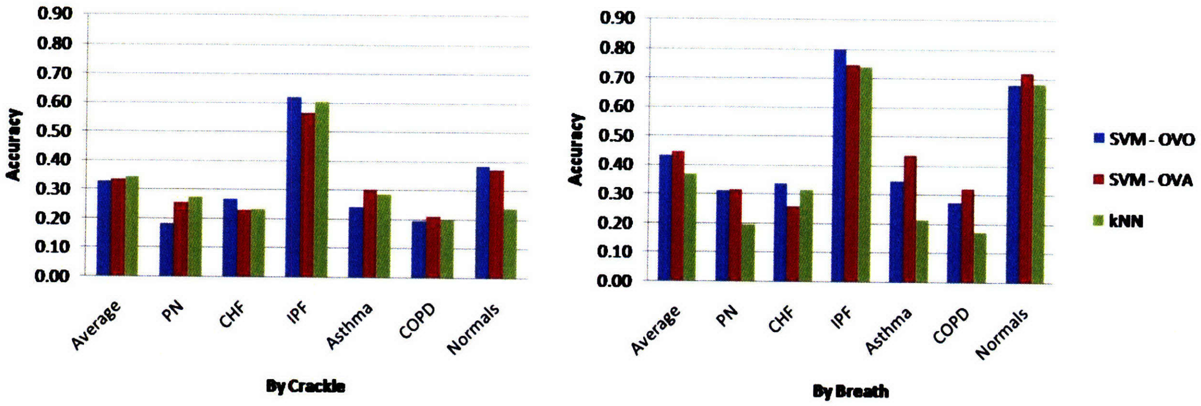


Figure 5.7 Multi-Class Accuracies for Method 1 (Left) and Method 2 (Right)

Using Method 1, only IPF crackles had any recognizable prediction power, and even so, it is only about 60% accurate. Overall accuracy is very poor. Using the per breath framework, the classification accuracies for IPF and normals are around 70%. This is fairly good recognition. In a multi-class classification, it is unlikely to be able to outperform the individual pair-wise comparison since there are more choices for every disease to make. Instead of the binary option, it can now be any one of six classes. In spite of this, the IPF and normals are still separable.

Surprisingly, for a very simplistic classifier, *k* nearest neighbor performs fairly well at nearly all choices for *k*. Clearly, as *k* gets larger, the ability to classify IPF and normals grows. This is due to an unknown group of features in multi-dimensional space where there is a bunching of these patients. It is more noticeable for higher values of *k* since the classes are likely intermixed. Some region has a much higher concentration of IPF and the algorithm picks up on it. The effects of varying *k* on Method 2 are shown in the Figure 5.8.

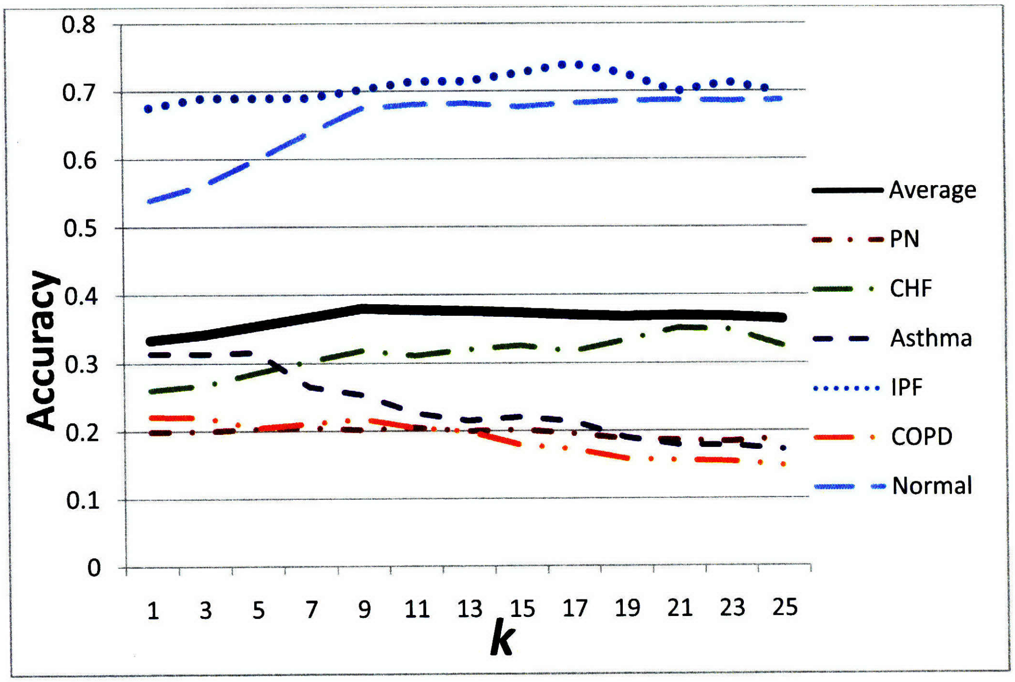


Figure 5.8 Effects of Varying *k* for kNN Classification

The thick black line is the overall accuracy. As *k* gets larger and larger, although the classification improves for the IPF and normal patients, it comes at the expense of the other diseases.

5.8 Summary

In this chapter, we presented the results and demonstrated the feasibility of computerized auscultation. We have shown that we can separate diseases very well through pair-wise comparisons. The best performing classes in the pair-wise comparisons were IPF and asymptomatic patients that we separated with accuracies near 85%. Still, all diseases displayed some amount of recognition performance. We also added clinical data to the acoustic feature sets that resulted in further increases in performance. Furthermore, we showed that IPF and normals can be distinguished very easily from all other diseases in a multi-class classification.

Chapter 6

Contributions, Applications, and Future Work

This chapter summarizes the contributions of this thesis and presents suggestions for future related research. We present general comments concerning the test results and applications of this technology. Finally, we provide recommendations for further work in the field.

6.1 Thesis Contributions

The goal of this thesis is to demonstrate the feasibility of a new diagnostic technique for cardiopulmonary disorders. We have shown that computerized auscultation by using a "smart" stethoscope can yield important diagnostic results. We have also shown that pair-wise comparisons can yield correct predictions with very high accuracy. In general, the best performing models were the ones that included all adventitious lung sounds. Crackles, wheezes, rhonchi, and squawks are all necessary sounds in order to diagnose patients with cardiopulmonary disorders. Furthermore, adding clinical information such as heart rate, respiratory rate, and temperature to the models increases performance significantly. Also, adding features related to the distribution of the crackles around the chest similarly improve classification performance.

This research has made the following contributions:

- Shows that multi-channel lung auscultation is a viable method for medical research.
- Shows that interstitial pulmonary fibrosis crackles are distinguishable from crackles of other diseases using acoustic analysis.
- Demonstrates that most pairs of diseases can be separated based on sounds, including asthma and chronic obstructive pulmonary disease. Pneumonia and congestive heart failure patients can be separated by incorporating acoustic and clinical data.
- Introduces a hybridized approach to data mining that combines data from multiple sources to make a diagnosis.
- Shows that interstitial pulmonary fibrosis and asymptomatic patients can be correctly classified when several diseases are possibilities.

6.2 Applications

It is not uncommon for doctors to misdiagnose patients even when they have traditionally strong diagnostic tests such as chest x-rays and computed tomography scans available. Furthermore, diagnoses can be hard to make particularly in the intensive care setting. In cases of doubt, the patient is often treated for both diseases which can be costly and harmful to the patient because of over medicating. The "smart" stethoscope will help resolve some of these indistinguishable comparisons.

Computerized auscultation via the "smart" stethoscope can be used in a variety of settings including remote telemedicine, in-home patient monitoring, and medical outreach. Remote telemedicine will be useful in any situation where a pulmonologist is not readily available. The "smart" stethoscope could provide either an initial diagnosis, or the results of the auscultation could be sent electronically to a doctor for further review. Some potential locations include oil rigs, embassies, forward operating bases, trans-oceanic ships, and any other location where a doctor and more advanced medical equipment may not be readily available.

In-home patient monitoring is another valuable potential application for this technology. Nurses often do not have the necessary training to perform auscultation, but they could easily be trained on how to administer the computerized auscultation using a "smart" stethoscope. The advantages of patient monitoring are numerous. After surgery or dismissal from the hospital a patient should still be monitored in case of complications. A "smart" stethoscope would provide a reliable and inexpensive means of monitoring these patients. Another benefit is that mildly ill patients in nursing homes or receiving in home care could be monitored. If their condition worsens, a nurse could be alerted to bring the patient to the Emergency Room. However, if their condition does not get worse, unnecessary trips to the emergency room could be prevented.

Finally, medical outreach will be transformed. In many developing countries, x-rays, and CT scans are unheard of outside the major cities. Doctors could travel to remote areas and administer the "smart" stethoscope and diagnose patients who would never have been able to get quality medical care. The stethoscope will be very cheap since it only involves a few microphones and a laptop computer. This portable technology can bring cheap, affordable health care to the masses.

6.3 Future Work

There are several opportunities for future work with respect to this thesis. First, genetic algorithms or some other feature selection algorithm could be applied to find the optimal subset of features for classification. Reducing the number of features will make the computer program more streamlined and also potentially yield better results. Multi-class neural networks should also be further explored because of their inherent properties which make them easily adaptable to the multi-class scenario.

The features used in the dataset could also be modified to reflect some of the ongoing research in defining adventitious lung sounds. Re-defining the features could improve classification performance. It also would be beneficial to collect the data with the purpose of performing computerized auscultation.

There are also many opportunities to extend this research from a medical perspective. We only consider five diseases in the study. More diseases could be added to the models to truly increase utility when making a diagnosis on an unseen patient. Furthermore, the same technology can be extended to not only adventitious lung sounds but also heart sounds. Heart murmurs, gallops, pleural rubs, and arrhythmias could all be modeled by this device. This could help for real-time analysis of aortic stenosis, heart disease, and other conditions.

Appendix A – Glossary of Acronyms

Avar	Amplitude Variability
BL	Bottom Left
BR	Bottom Right
CHF	Congestive Heart Failure
CT	Computed Tomography
CTC	Crackle Transmission Coefficient
COPD	Chronic Obstructive Pulmonary Disease
ERM	Empirical Risk Minimization
FN	False Negative
FP	False Positive
IPF	Idiopathic (Interstitial) Pulmonary Fibrosis
kNN	*k* Nearest Neighbor
MRI	Magnetic Resonance Imaging
OVA	One-vs-all
OVO	One-vs-one
PN	Pneumonia
RR	Respiratory Rate
SRM	Structural Risk Minimization
STG	Stethographics
SVM	Support Vector Machine
TEWA	Time Expanded Waveform Analysis
TL	Top Left
TN	True Negative
TP	True Positive
TR	Top Right
Tvar	Half Period Variability
VC	Vapnik Chervonenkis
ZXS	Zero Crossings

[This Page Intentionally Left Blank]

Appendix B – SVM Derivations

In this appendix, we extend the mathematical formulations for support vector machines. This section is intended for the mathematically inclined and interested reader.

Derivation of the Margin

In Chapter 3, we said the a separating hyperplane has this form: $\boldsymbol{w^T} \cdot \boldsymbol{x} + b = 0$. By adding a normalization constraint with respect to the data points, we can call this separator a *canonical hyperplane.* A hyperplane is in canonical form with respect to the data set if the following requirement is satisfied:

$$\min_{\forall i} |\boldsymbol{w^T} \cdot \boldsymbol{x_i} + b| = 1 \tag{A.1}$$

The distance $d(\boldsymbol{w}, \mathrm{b}, \boldsymbol{x})$ from any point $\boldsymbol{x_i}$ to the hyperplane is:

$$d(\boldsymbol{w}, \mathrm{b}, \boldsymbol{x_i}) = \frac{|\boldsymbol{w^T} \cdot \boldsymbol{x_i} + b|}{\|\boldsymbol{w}\|} \tag{A.2}$$

We also know that points on both side of the hyperplane must satisfy Equation A.1. The margin, M, is the distance between these two (or more) points that are the minimum distance away. Combining Equation A.1 and A.2 gives us the following derivation.

$$M = \min_{\boldsymbol{x_i}|y_i=+1} d(\boldsymbol{w}, \mathrm{b}, \boldsymbol{x_i}) + \min_{\boldsymbol{x_i}|y_i=-1} d(\boldsymbol{w}, \mathrm{b}, \boldsymbol{x_i}) \tag{A.3}$$

$$= \min_{\boldsymbol{x_i}|y_i=+1} \frac{|\boldsymbol{w^T} \cdot \boldsymbol{x_i} + b|}{\|\boldsymbol{w}\|} + \min_{\boldsymbol{x_i}|y_i=-1} \frac{|\boldsymbol{w^T} \cdot \boldsymbol{x_i} + b|}{\|\boldsymbol{w}\|} \tag{A.4}$$

$$= \frac{1}{\|\boldsymbol{w}\|} \left(\min_{\boldsymbol{x_i}|y_i=+1} |\boldsymbol{w^T} \cdot \boldsymbol{x_i} + b| + \min_{\boldsymbol{x_i}|y_i=-1} |\boldsymbol{w^T} \cdot \boldsymbol{x_i} + b|\right) \tag{A.5}$$

$$= \frac{1}{\|\boldsymbol{w}\|} * (1 + 1) \tag{A.6}$$

$$= \frac{2}{\|\boldsymbol{w}\|} \qquad \blacksquare \tag{A.7}$$

Dual Formulations to Solve Convex Optimization

Remember from Chapter 3 the following formulation of the optimization problem for the linearly separable dataset.

$$\mathrm{minimize}_{\mathbf{w},b} \ \frac{1}{2} \|\boldsymbol{w}\|^2 \tag{A.8}$$

$$\text{subject to} \quad y_i(\boldsymbol{w}^T \cdot \boldsymbol{x_i} + b) \geq 1 \quad \forall i = 1, 2, \ldots, N$$

To show how to solve this constrained minimization problem, we introduce the *Lagrangian Dual* formulation. Forming the dual problem has the several benefits. First, the constraints in formulation A.1 make the problem hard to solve. Instead of solving the problem this way, we introduce Lagrange multipliers, α_i, for each constraint and move the constraints themselves into the objective function. The original constraints in A.1 are replaced by constraints on the multipliers. Second, solving this type of problem in its dual form is typically more efficient sine it only involves dot products of vectors. Finally and most importantly, from the dual formulation we can derive the so-called *support vectors*.

To form the Lagrangian, we take the non-negative Lagrange multipliers and subtract them from the objective function. This gives the following Lagrangian:

$$\mathrm{L}(\boldsymbol{w}, \mathrm{b}, \boldsymbol{\alpha}) \triangleq \frac{1}{2}\, \|\boldsymbol{w}\|^2 - \textstyle\sum_{i=1}^{N} \alpha_i [y_i(\boldsymbol{w}^T \cdot \boldsymbol{x_i} + b) - 1] \qquad \text{(A.9)}$$

$$\equiv \frac{1}{2}\, \|\boldsymbol{w}\|^2 - \textstyle\sum_{i=1}^{N} \alpha_i\, y_i(\boldsymbol{w}^T \cdot \boldsymbol{x_i} + b) + \sum_{i=1}^{N} \alpha_i$$

To minimize the Lagrangian, the **α** vector is fixed, and the partial derivatives with respect to ***w*** and b must be equal to zero.

$$\frac{d L(\boldsymbol{w}, \mathrm{b}, \boldsymbol{\alpha})}{d\boldsymbol{w}} = \boldsymbol{w} - \textstyle\sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x_i} = 0 \qquad \text{(A.10)}$$

$$\frac{d L(\boldsymbol{w}, \mathrm{b}, \boldsymbol{\alpha})}{d b} = -\textstyle\sum_{i=1}^{N} \alpha_i y_i = 0 \qquad \text{(A.11)}$$

Therefore, we have the following conditions.

$$\boldsymbol{w} = \textstyle\sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x_i} \quad \text{and} \qquad \text{(A.12)}$$

$$\textstyle\sum_{i=1}^{N} \alpha_i y_i = 0 \qquad \text{(A.13)}$$

Substituting A.5 back into Equation A.2:

$$\mathrm{L}(\boldsymbol{w}, \mathrm{b}, \boldsymbol{\alpha})$$

$$= \frac{1}{2} \left\| \textstyle\sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x_i} \right\|^2 - \textstyle\sum_{i=1}^{N} \alpha_i y_i \left(\left(\sum_{j=1}^{N} \alpha_j y_j \boldsymbol{x_j} \right) \cdot \boldsymbol{x_i} + b \right) + \sum_{i=1}^{N} \alpha_i \qquad \text{(A.14)}$$

$$= \frac{1}{2} \textstyle\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i}\, \boldsymbol{x_j} - \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i}\, \boldsymbol{x_j} - b \sum_{i=1}^{N} \alpha_i y_i + \sum_{i=1}^{N} \alpha_i \qquad \text{(A.15)}$$

$$= -\frac{1}{2} \textstyle\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i}\, \boldsymbol{x_j} + \sum_{i=1}^{N} \alpha_i \qquad \text{(A.16)}$$

Let $D(\boldsymbol{\alpha})$ be the minimum value of the Lagrangian for a particular **α**. Then the following conditions exist:

$$D(\boldsymbol{\alpha}) = \begin{cases} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i\,\boldsymbol{x}_j + \sum_{i=1}^{N}\alpha_i & \text{if } \sum_{i=1}^{N}\alpha_i y_i = 0 \\ -\infty & \text{otherwise} \end{cases} \tag{A.17}$$

If the binding condition does not hold, b can increase or decrease to ±∞ causing the minimum to be unbounded. According to duality theory, we need to now maximize $D(\boldsymbol{\alpha})$ giving the following maximization problem:

$$\begin{aligned} &\text{maximize}_{\alpha} \quad -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i\,\boldsymbol{x}_j + \sum_{i=1}^{N}\alpha_i \\ &\text{subject to} \quad \sum_{i=1}^{N}\alpha_i y_i = 0 \\ &\qquad\qquad\qquad \alpha_i \geq 0 \qquad \forall i = 1,2,\ldots,N \end{aligned} \tag{A.18}$$

Here, since each dual variable α_i corresponds to one of the constraints in the original primal formulation, they take on special characteristics here. If $\alpha_i > 0$, then the constraint in the primal is active. On the other hand, if $\alpha_i = 0$, then the constraint is inactive. The constraint will only be active if the point lies on the margin of the optimal separating hyperplane. These are the support vectors. The training points that do not lie on the margin are unimportant and do not affect the orientation of the hyperplane. Furthermore, the optimal separating hyperplane is defined as a linear function of these support vectors.

$$\mathrm{w}^* = \sum_{\mathrm{x_i}\text{ is a s.v.}} \mathrm{\alpha_i y_i x_i} \tag{A.19}$$

Also, since the support vectors are often only a small proportion of the total data, the model can be represented using minimal computation expenditures and memory requirements.

[This Page Intentionally Left Blank]

Appendix C – Feature Definitions

Features are defined for inspiration and expiration	Num Crackles	The total number of crackles in a patient
	Cr/Breath	Average number of crackles per breath
	Med ZXS	Median number of zero crossings (Section 4.1.1)
	Med T1	Median T1 (Section 4.1.1)
	Med Freq	Median frequency (Section 4.1.1)
	Med T2/T1	Median T2/T1 (Section 4.1.1)
	Med Tvar	Median Tvar (Section 4.1.1)
	Med Timing	Median Timing (Section 4.1.1)
	Med CTC	Median CTC (Section 4.1.1)
	Med Ampl	Median amplitude (Section 4.1.1)
	Med A2/A1	Median A2/A1 (Section 4.1.1)
	Med A3/A1	Median A3/A1 (Section 4.1.1)
	Med Avar	Median Avar (Section 4.1.1)
	Med PosPolar	Median PosPolar (Section 4.1.1)
	LtoR Crackle Percent Diff	A percentage which measures the symmetry of the number of crackles on each side of the chest. It is calculated as the (#Crackles on the L - #Crackles on the R)/(Total # of crackles). 0% means there is no difference between sides and therefore an equal distribution. 100% means all crackles are on one side.
	Wz Rate	Percentage of breath cycle occupied by wheezing
	LtoR Wheeze Percent Diff	A percentage which measures the symmetry of the number of wheezes on each side of the chest. It is calculated as the (#Wheezes on the L - #Wheezes on the R)/(Total # of wheezes). 0% means there is no difference between sides and therefore an equal distribution. 100% means all wheezes are on one side.
	Freq Wz	Mean wheeze frequency
	Peak Ratio Wz	A measure of how sharp the wheeze is in frequency domain
	Timing Wz	When wheeze occurs during breath cycle
	WTC Wz	Wheeze transmission coefficient, similar to CTC
	Ampl Wz	Wheeze amplitude in dB
	Avar Wz	Variability of the amplitude of wheeze on mother channel throughout 20 seconds expressed in percent of wheeze amplitude.
	RMS	Average root mean square among all chest channels. RMS is a measure of sound power.
	STDev	Variation of RMS in the chest channels expressed as percent of average RMS
	lRMS/rRMS	Ratio of average RMS in left lung to averaged RMS in right lung. Expressed as a percentage.
	LtoR Rhonchi Percent Diff	A percentage which measures the symmetry of the number of rhonchi on each side of the chest. It is calculated as the (#Rhonchi on the L - #Rhonchi on the R)/(Total # of rhonchi). 0% means there is no difference between sides and therefore an equal distribution. 100% means all rhonchi are on one side.
	Duration	Inspiratory or expiratory duration in seconds

Features only defined for inspiration	Squawks/breath	The number of squawks heard per breath
	R1	Inspiratory duration/Expiratory duration in percent
	eRMS/iRMS	Ratio of expiratory RMS to inspiratory RMS in percent
	lR4	Left inspiratory R4 (ratio of low frequency energy (10Hz to 80Hz) to high frequency energy (80Hz to 500Hz)
	rR4	right inspiratory R4
	iStart	In-homogeneity of inspiratory start
	iEnd	In-homogeneity of inspiratory end
	lDR	Left inspiratory dynamic range (the difference between maximum and minimum sound amplitude)
	rDR :	Right inspiratory dynamic range
	LtoR Dynamic Range Percent	(lDR-rDR)/(l+r) Absolute difference between left and right inspiratory dynamic range expressed as percent of total.

Appendix D – Supplemental Results

SVM Classification of Breaths Crackle Only Data Set

	Pneumonia					
	Breath Analysis			Voting		
Neg Class	Sens	Spec	Acc	Sens	Spec	Acc
CHF	0.558	0.470	0.527	0.567	0.436	0.523
IPF	0.766	0.785	0.769	0.831	0.755	0.823
Asthma	0.740	0.522	0.705	0.785	0.545	0.740
COPD	0.644	0.470	0.590	0.658	0.470	0.597
Normals	0.690	0.870	0.753	0.739	0.930	0.792

	Congestive Heart Failure					
	Breath Analysis			Voting		
Neg Class	Sens	Spec	Acc	Sens	Spec	Acc
PN	0.470	0.558	0.527	0.436	0.567	0.523
IPF	0.784	0.815	0.790	0.848	0.802	0.841
Asthma	0.672	0.459	0.621	0.724	0.484	0.659
COPD	0.655	0.405	0.529	0.722	0.368	0.544
Normals	0.631	0.794	0.674	0.686	0.814	0.712

	Interstitial Pulmonary Fibrosis					
	Breath Analysis			Voting		
Neg Class	Sens	Spec	Acc	Sens	Spec	Acc
PN	0.785	0.766	0.769	0.755	0.831	0.823
CHF	0.815	0.784	0.790	0.802	0.848	0.841
Asthma	0.838	0.789	0.802	0.797	0.806	0.804
COPD	0.811	0.746	0.758	0.789	0.803	0.801
Normals	0.821	0.678	0.694	0.837	0.757	0.764

	Asthma					
	Breath Analysis			Voting		
Neg Class	Sens	Spec	Acc	Sens	Spec	Acc
PN	0.522	0.740	0.705	0.545	0.785	0.740
CHF	0.459	0.672	0.621	0.484	0.724	0.659
IPF	0.789	0.838	0.802	0.806	0.797	0.804
COPD	0.509	0.612	0.583	0.533	0.638	0.610
Normals	0.668	0.555	0.571	0.696	0.601	0.611

	COPD					
	Breath Analysis			Voting		
Neg Class	Sens	Spec	Acc	Sens	Spec	Acc
PN	0.470	0.644	0.590	0.470	0.658	0.597
CHF	0.405	0.655	0.529	0.368	0.722	0.544
IPF	0.746	0.811	0.758	0.803	0.789	0.801
Asthma	0.612	0.509	0.583	0.638	0.533	0.610
Normals	0.735	0.603	0.634	0.792	0.626	0.658

	Normals					
	Breath Analysis			Voting		
Neg Class	Sens	Spec	Acc	Sens	Spec	Acc
PN	0.870	0.690	0.753	0.930	0.739	0.792
CHF	0.794	0.631	0.674	0.814	0.686	0.712
IPF	0.678	0.821	0.694	0.757	0.837	0.764
Asthma	0.555	0.668	0.571	0.601	0.696	0.611
COPD	0.603	0.735	0.634	0.626	0.792	0.658

SVM Classification of Breaths Crackle and Distribution Data Set

	Pneumonia					
	Breath Analysis			Voting		
Neg Class	Sens	Spec	Acc	Sens	Spec	Acc
CHF	0.665	0.537	0.581	0.559	0.517	0.545
IPF	0.787	0.799	0.789	0.857	0.779	0.849
Asthma	0.841	0.402	0.770	0.873	0.411	0.789
COPD	0.771	0.314	0.629	0.801	0.317	0.641
Normals	0.827	0.781	0.797	0.864	0.858	0.860

	Congestive Heart Failure					
	Breath Analysis			Voting		
Neg Class	Sens	Spec	Acc	Sens	Spec	Acc
PN	0.537	0.665	0.581	0.517	0.559	0.545
IPF	0.802	0.786	0.800	0.879	0.743	0.858
Asthma	0.763	0.367	0.666	0.786	0.354	0.676
COPD	0.809	0.274	0.545	0.871	0.241	0.559
Normals	0.842	0.779	0.826	0.856	0.754	0.836

	Interstitial Pulmonary Fibrosis					
	Breath Analysis			Voting		
Neg Class	Sens	Spec	Acc	Sens	Spec	Acc
PN	0.799	0.787	0.789	0.779	0.857	0.849
CHF	0.786	0.802	0.800	0.743	0.879	0.858
Asthma	0.824	0.850	0.843	0.791	0.874	0.854
COPD	0.744	0.820	0.807	0.664	0.874	0.843
Normals	0.831	0.852	0.851	0.840	0.914	0.909

	Asthma					
	Breath Analysis			Voting		
Neg Class	Sens	Spec	Acc	Sens	Spec	Acc
PN	0.402	0.841	0.770	0.411	0.873	0.789
CHF	0.367	0.763	0.666	0.354	0.786	0.676
IPF	0.850	0.824	0.843	0.874	0.791	0.854
COPD	0.436	0.672	0.609	0.446	0.690	0.623
Normals	0.649	0.587	0.640	0.711	0.611	0.700

	COPD					
	Breath Analysis			Voting		
Neg Class	Sens	Spec	Acc	Sens	Spec	Acc
PN	0.314	0.771	0.629	0.317	0.801	0.641
CHF	0.274	0.809	0.545	0.241	0.871	0.559
IPF	0.820	0.744	0.807	0.874	0.664	0.843
Asthma	0.672	0.436	0.609	0.690	0.446	0.623
Normals	0.693	0.725	0.717	0.719	0.805	0.789

	Normals					
	Breath Analysis			Voting		
Neg Class	Sens	Spec	Acc	Sens	Spec	Acc
PN	0.781	0.827	0.797	0.858	0.864	0.860
CHF	0.779	0.842	0.826	0.754	0.856	0.836
IPF	0.852	0.831	0.851	0.914	0.840	0.909
Asthma	0.587	0.649	0.640	0.611	0.711	0.700
COPD	0.725	0.693	0.717	0.805	0.719	0.789

SVM Classification of Breaths Other Sounds Data Set

	Pneumonia					
	Breath Analysis			Voting		
Neg Class	Sens	Spec	Acc	Sens	Spec	Acc
CHF	0.599	0.639	0.613	0.599	0.651	0.616
IPF	0.851	0.776	0.843	0.873	0.720	0.856
Asthma	0.757	0.626	0.736	0.742	0.663	0.728
COPD	0.714	0.704	0.711	0.688	0.737	0.704
Normals	0.884	0.880	0.883	0.904	0.842	0.887

	Congestive Heart Failure					
	Breath Analysis			Voting		
Neg Class	Sens	Spec	Acc	Sens	Spec	Acc
PN	0.639	0.599	0.613	0.651	0.599	0.616
IPF	0.800	0.796	0.800	0.825	0.733	0.811
Asthma	0.776	0.713	0.760	0.762	0.711	0.748
COPD	0.751	0.681	0.717	0.710	0.686	0.698
Normals	0.847	0.819	0.840	0.867	0.791	0.853

	Interstitial Pulmonary Fibrosis					
	Breath Analysis			Voting		
Neg Class	Sens	Spec	Acc	Sens	Spec	Acc
PN	0.776	0.851	0.843	0.720	0.873	0.856
CHF	0.796	0.800	0.800	0.733	0.825	0.811
Asthma	0.911	0.856	0.871	0.858	0.859	0.859
COPD	0.834	0.846	0.845	0.793	0.856	0.847
Normals	0.861	0.824	0.828	0.842	0.829	0.830

	Asthma					
	Breath Analysis			Voting		
Neg Class	Sens	Spec	Acc	Sens	Spec	Acc
PN	0.626	0.757	0.736	0.663	0.742	0.728
CHF	0.713	0.776	0.760	0.711	0.762	0.748
IPF	0.856	0.911	0.871	0.859	0.858	0.859
COPD	0.657	0.736	0.715	0.675	0.721	0.709
Normals	0.804	0.824	0.821	0.762	0.833	0.825

	COPD					
	Breath Analysis			Voting		
Neg Class	Sens	Spec	Acc	Sens	Spec	Acc
PN	0.704	0.714	0.711	0.737	0.688	0.704
CHF	0.681	0.751	0.717	0.686	0.710	0.698
IPF	0.846	0.834	0.845	0.856	0.793	0.847
Asthma	0.736	0.657	0.715	0.721	0.675	0.709
Normals	0.845	0.932	0.912	0.814	0.936	0.913

	Normals					
	Breath Analysis			Voting		
Neg Class	Sens	Spec	Acc	Sens	Spec	Acc
PN	0.880	0.884	0.883	0.842	0.904	0.887
CHF	0.819	0.847	0.840	0.791	0.867	0.853
IPF	0.824	0.861	0.828	0.829	0.842	0.830
Asthma	0.824	0.804	0.821	0.833	0.762	0.825
COPD	0.932	0.845	0.912	0.936	0.814	0.913

SVM Classification of Breaths Full Data Set

	Pneumonia					
	Breath Analysis			Voting		
Neg Class	Sens	Spec	Acc	Sens	Spec	Acc
CHF	0.540	0.696	0.588	0.540	0.695	0.592
IPF	0.847	0.777	0.839	0.875	0.739	0.860
Asthma	0.763	0.617	0.739	0.750	0.647	0.731
COPD	0.708	0.675	0.698	0.681	0.709	0.690
Normals	0.891	0.896	0.893	0.910	0.867	0.898

	Congestive Heart Failure					
	Breath Analysis			Voting		
Neg Class	Sens	Spec	Acc	Sens	Spec	Acc
PN	0.696	0.540	0.588	0.695	0.540	0.592
IPF	0.806	0.783	0.804	0.840	0.742	0.825
Asthma	0.794	0.667	0.763	0.791	0.653	0.753
COPD	0.791	0.667	0.730	0.752	0.698	0.725
Normals	0.838	0.825	0.835	0.870	0.795	0.856

	Interstitial Pulmonary Fibrosis					
	Breath Analysis			Voting		
Neg Class	Sens	Spec	Acc	Sens	Spec	Acc
PN	0.777	0.847	0.839	0.739	0.875	0.860
CHF	0.783	0.806	0.804	0.742	0.840	0.825
Asthma	0.889	0.873	0.877	0.848	0.881	0.873
COPD	0.803	0.840	0.835	0.756	0.856	0.841
Normals	0.841	0.839	0.840	0.842	0.849	0.849

	Asthma					
	Breath Analysis			Voting		
Neg Class	Sens	Spec	Acc	Sens	Spec	Acc
PN	0.617	0.763	0.739	0.647	0.750	0.731
CHF	0.667	0.794	0.763	0.653	0.791	0.753
IPF	0.873	0.889	0.877	0.881	0.848	0.873
COPD	0.668	0.729	0.713	0.689	0.718	0.710
Normals	0.712	0.922	0.891	0.674	0.916	0.887

	COPD					
	Breath Analysis			Voting		
Neg Class	Sens	Spec	Acc	Sens	Spec	Acc
PN	0.675	0.708	0.698	0.709	0.681	0.690
CHF	0.667	0.791	0.730	0.698	0.752	0.725
IPF	0.840	0.803	0.835	0.856	0.756	0.841
Asthma	0.729	0.668	0.713	0.718	0.689	0.710
Normals	0.843	0.928	0.908	0.829	0.946	0.924

	Normals					
	Breath Analysis			Voting		
Neg Class	Sens	Spec	Acc	Sens	Spec	Acc
PN	0.896	0.891	0.893	0.867	0.910	0.898
CHF	0.825	0.838	0.835	0.795	0.870	0.856
IPF	0.839	0.841	0.840	0.849	0.842	0.849
Asthma	0.922	0.712	0.891	0.916	0.674	0.887
COPD	0.928	0.843	0.908	0.946	0.829	0.924

SVM Breath Classification with Voting Applied

		Voting Results for Method 2											
		Crackle Only			Crackle and Distribution			Other Sounds			Full Data		
		Sens	Spec	Acc	Sens	Spec	Acc	Sens	Spec	Acc	Sens	Spec	Acc
PN	CHF	0.567	0.436	0.523	0.559	0.517	0.545	0.599	0.651	0.616	0.540	0.695	0.592
	IPF	0.831	0.755	0.823	0.857	0.779	0.849	0.873	0.720	0.856	0.875	0.739	0.860
	Asthma	0.785	0.545	0.740	0.873	0.411	0.789	0.742	0.663	0.728	0.750	0.647	0.731
	COPD	0.658	0.470	0.597	0.801	0.317	0.641	0.688	0.737	0.704	0.681	0.709	0.690
	Normals	0.739	0.930	0.792	0.864	0.858	0.860	0.904	0.842	0.887	0.910	0.867	0.898
CHF	IPF	0.848	0.802	0.841	0.879	0.743	0.858	0.825	0.733	0.811	0.840	0.742	0.825
	Asthma	0.724	0.484	0.659	0.786	0.354	0.676	0.762	0.711	0.748	0.791	0.653	0.753
	COPD	0.722	0.368	0.544	0.871	0.241	0.559	0.710	0.686	0.698	0.752	0.698	0.725
	Normals	0.686	0.814	0.712	0.856	0.754	0.836	0.867	0.791	0.853	0.870	0.795	0.856
IPF	Asthma	0.797	0.806	0.804	0.791	0.874	0.854	0.858	0.859	0.859	0.848	0.881	0.873
	COPD	0.789	0.803	0.801	0.664	0.874	0.843	0.793	0.856	0.847	0.756	0.856	0.841
	Normals	0.837	0.757	0.764	0.840	0.914	0.909	0.842	0.829	0.830	0.842	0.849	0.849
Asthma	COPD	0.533	0.638	0.610	0.446	0.690	0.623	0.675	0.721	0.709	0.689	0.718	0.710
	Normals	0.696	0.601	0.611	0.711	0.611	0.700	0.762	0.833	0.825	0.674	0.916	0.887
COPD	Normals	0.792	0.626	0.658	0.719	0.805	0.789	0.814	0.936	0.913	0.829	0.946	0.924

[This Page Intentionally Left Blank]

References

[1] Anonymous. Medical Antiques Online. 2006. [Online]. Available: http://www.antiquemed.com [Accessed May 5, 2008].

[2] R. L. Murphy, "Computerized Multichannel Lung Sound Analysis," IEEE Engineering in Medicine and Biology Magazine, vol. 26, pp. 16-19, 2007.

[3] P. B. Ginsburg, "Controlling Health Care Costs," N Engl J Med, vol. 351, pp. 1591-1593, October 14. 2004.

[4] P. Forgacs, Lung Sounds. London, UK: Cassel, 1978, pp. 80.

[5] R. L. Murphy, A. Vyshedskiy, V. A. Power-Charnitsky, D. S. Bana, P. M. Marinelli, A. Wong-Tse and R. Paciej, " Automated Lung Sound Analysis in Patients With Pneumonia," Respir Care, vol. 49, pp. 1490-1497, 2004.

[6] R. L. Murphy Jr, S. K. Holford and W. C. Knowler, "Visual lung sound characterization by time-expanded waveform analysis," New England Journal of Medicine, vol. 296, pp. 968-971, April 28, 1977.

[7] R. L. Murphy Jr, E. A. Gaensler, S. K. Holford, E. A. Del Bono and G. Epler, "Crackles in the early detection of asbestosis," Am. Rev. Respir. Dis., vol. 129, pp. 375-379, Mar. 1984.

[8] A. R. A. Sovijarvi, F. Dalmasso, J. Vanderschoof, L. P. Malmberg, G. Righini and S. A. T. Stoneman, "Definition of Terms for Applications of Respiratory Sounds," Eur Respir Rev, vol. 10, pp. 597-610, 2000.

[9] R. L. Murphy Jr, E. A. Del Bono and F. Davidson, "Validation of an automatic crackle (rale) counter," Am. Rev. Respir. Dis., vol. 140, pp. 1017-1020, Oct. 1989.

[10] R. L. Murphy, "Localization of chest sounds with 3D display and lung soung mapping," US Patent 5 844 997, Dec. 1, 1998.

[11] R. L. Murphy, M. A. Murphy, G. Brockington and A. Vyshedskiy. A simplified introduction to heart and lung sounds: Interactive multimedia CD-ROM. [CD-ROM]. Version 2.0, 2006.

[12] A. Vyshedskiy, F. Bezares, R. Paciej, M. Ebril, J. Shane and R. Murphy, "Transmission of Crackles in Patients With Interstitial Pulmonary Fibrosis, Congestive Heart Failure, and Pneumonia," Chest, vol. 128, pp. 1468-1474, September 1. 2005.

[13] M. S. Niederman, "Pneumonia, including community-acquired and nosocomial pneumonia," in Baum's Textbook of Pulmonary Diseases ,7th ed.J. D. Crapo, J. L. Glassroth, J. B. Karlinsky and T. E. King, Eds. Lippincott Williams & Wilkins, 2003, pp. 1455.

[14] H. Simon. Pnuemonia diagnosis. NY Times Health. April 3, 2007. [Online]. Available: http://health.nytimes.com/health/guides/disease/pneumonia/diagnosis.html [Accessed May 5, 2008].

[15] G. Gandelman. Heart failure - symptoms, diagnosis, and treatment of heart failure. NY Times Health. July 17, 2006. Available: http://health.nytimes.com/health/guides/disease/heart-failure/in-depth-report.html [Accessed May 5, 2008].

[16] T. J. Desai and J. B. Karlinsky, "COPD: Clinical manifestations, diagnosis, and treatment," in Baum's Textbook of Pulmonary Diseases ,7th ed.J. D. Crapo, J. L. Glassroth, J. B. Karlinsky and T. E. King, Eds. Lippincott Williams & Wilkins, 2003, pp. 1455.

[17] N. R. Anthonisen and J. Manfreda, "Epidemiology of chronic obstructive pulmonary disease," in Baum's Textbook of Pulmonary Diseases ,7th ed.J. D. Crapo, J. L. Glassroth, J. B. Karlinsky and T. E. King, Eds. Lippincott Williams & Wilkins, 2003, pp. 1455.

[18] E. R. Sutherland, M. Kraft and J. D. Crapo, "Diagnosis and treatment of asthma," in Baum's Textbook of Pulmonary Diseases ,7th ed.J. D. Crapo, J. L. Glassroth, J. B. Karlinsky and T. E. King, Eds. Lippincott Williams & Wilkins, 2003, pp. 1455.

[19] Anonymous "Idiopathic Pulmonary Fibrosis: Diagnosis and Treatment . International Consensus Statement," Am. J. Respir. Crit. Care Med., vol. 161, pp. 646-664, February 1. 2000.

[20] S. Nagai and M. Kitaichi, "Idiopathic interstitial pnuemonias," in Baum's Textbook of Pulmonary Diseases ,7th ed.J. D. Crapo, J. L. Glassroth, J. B. Karlinsky and T. E. King, Eds. Lippincott Williams & Wilkins, 2003, pp. 1455.

[21] N. al Jarad, B. Strickland, G. Bothamley, S. Lock, R. Logan-Sinclair and R. Rudd, "Diagnosis of asbestosis by a time expanded wave form analysis, auscultation and high resolution computed tomography: a comparative study," Thorax, vol. 48, pp. 347-353, April 1. 1993.

[22] P. Bettencourt, E. Del Bono, D. Spiegelman, E. Hertzmark and R. Murphy Jr, "Clinical utility of chest auscultation in common pulmonary diseases," Am. J. Respir. Crit. Care Med., vol. 150, pp. 1291-1297, November 1. 1994.

[23] T. Kawamura, T. Matsumoto, N. Tanaka, S. Kido, Z. Jiang and N. Matsunaga, "Crackle Analysis for Chest Auscultation and Comparison with High-Resolution CT Findings," Radiation Medicine, vol. 21, pp. 258-266, 2003.

[24] N. Gavriely, M. Nissan, D. Cugell and A. Rubin, "Respiratory health screening using pulmonary function tests and lung sound analysis," Eur Respir J, vol. 7, pp. 35-42, January 1. 1994.

[25] T. Bergstresser, D. Ofengeim, A. Vyshedskiy, J. Shane and R. Murphy, "Sound transmission in the lung as a function of lung volume," J Appl Physiol, vol. 93, pp. 667-674, August 1. 2002.

[26] A. Kandaswamy, C. S. Kumar, R. P. Ramanathan, S. Jayaraman and N. Malmurugan, "Neural classification of lung sounds using wavelet coefficients," Computers in Biology and Medicine, vol. 34, pp. 523-537, 9. 2004.

[27] S. A. Taplidou and L. J. Hadjileontiadis, "Wheeze detection based on time-frequency analysis of breath sounds," Computers in Biology and Medicine, vol. 37, pp. 1073-1083, 8. 2007.

[28] D. E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning. Reading, MA: Addison-Wesley, 1989, pp. 412.

[29] I. Guler, H. Polat and U. Ergun, "Combining Neural Network and Genetic Algorithm for Prediction of Lung Sounds," J. Med. Syst., vol. 29, pp. 217-231, 2005.

[30] D. J. Hand, H. A. Mannila, P. A. Smyth, Principles of Data Mining, Cambridge, MA: The MIT Press, 2001, pp 578.

[31] G. F. Luger, Artificial Intelligence: Structures and Strategies for Complex Problem Solving. ,5th ed.Harlow, England ; New York: Addison-Wesley, 2005, pp. 903.

[32] T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning. New York: Springer, 2001, pp. 533.

[33] V. N. Vapnik, The Nature of Statistical Learning Theory, 1st ed. New York: Wiley, 1995, pp. 314.

[34] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," Informatica, vol. 31, pp. 249-268, 2007.

[35] R. O. Duda, P. E. Hart and D. G. Stork, Pattern Classification, 2nd ed. New York, NY: Wiley, 2000, pp. 654.

[36] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," Data Mining and Knowledge Discovery, vol. 2, pp. 121-167, 1998.

[37] S. S. Haykin, Neural Networks : A Comprehensive Foundation. ,2nd ed.Upper Saddle River, N.J.: Prentice Hall, 1999, pp. 842.

[38] Fix, E., Hodges, J.L. Jr., "Discriminatory analysis, non-parametric discrimination," USAF School of Aviation Medicine, Randolph Field, TX., Technical Report 4, 1951.

[39] Y. Lee, "Handwritten digit recognition using k nearest-neighbor, radial-basis function, and backpropagation neural networks," vol. 3, pp. 440-449, 1991.

[40] B. D. Ripley, Pattern Recognition and Neural Networks. Cambridge UK: Cambridge University Press, 1996, pp. 415.

[41] J. P. Vert, "Introduction to support vector machines and applications to computational biology," presented at the Human Genome Center, University of Tokyo, Japan, July 17-19, 2002.

[42] U. Krebel, "Pairwise classification and support vector machines," in Advances in Kernel Methods: Support Vector Learning ,1st ed.B. Scholkopf, C. J. C. Burges and A. J. Smola, Eds. Cambridge, MA: The MIT Press, 1999, pp. 255-268.

[43] R. Rifkin and A. Klautau, "In Defense of One-Vs-All Classification," J. Mach. Learn. Res., vol. 5, pp. 101-141, 2004.

[44] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed. San Francisco: Morgan Kaufman, 2005, pp. 560.

[45] T. Joachims, "Making large-scale SVM learning practical," in Advances in Kernel Methods - Support Vector Learning B. Scholkopf, C. J. C. Burges and A. J. Smola, Eds. MIT Press, 1999, pp. 169-184.

[46] T. Briggs. MATLAB/MEX interface to SVMlight v6.01. July 6, 2005. [Online] Available: http://webspace.ship.edu/thbrig/mexsvm.