Design and Validation of Mobile Kit and Machine Learning Algorithms for Pulmonary Disease Screening and Diagnosis

by

Daniel Chamberlain

Submitted to the Institute for Data, Systems, and Society and the Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degrees of Master of Science in Technology and Policy and Master of Science in Electrical Engineering and Computer Science at the MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2017

© Massachusetts Institute of Technology 2017. All rights reserved.
Design and Validation of Mobile Kit and Machine Learning Algorithms for Pulmonary Disease Screening and Diagnosis

by

Daniel Chamberlain

Submitted to the Institute for Data, Systems, and Society on January 13, 2017, in partial fulfillment of the requirements for the degrees of Master of Science in Technology and Policy and Master of Science in Electrical Engineering and Computer Science

Abstract

Pulmonary diseases are responsible for more than 15% of deaths worldwide. Much of this burden is concentrated in the developing world, where these diseases cause 19% of deaths. In much of the developing world, pulmonary diseases are underdiagnosed and misdiagnosed because the correct equipment is not available or health care is provided by workers with insufficient training.

To help improve the diagnosis of pulmonary disease, we built a pulmonary diagnostic kit that consists of an electronic stethoscope, an augmented reality peak flow meter, and an electronic questionnaire. Using this kit, we collected data from patients who visited the Chest Research Foundation, a pulmonary clinic in Pune, India.

Using the data collected from these patients, we pursued several avenues of research. First, we trained algorithms to automatically detect two adventitious breath sounds: wheezes and crackles. We used two approaches to detect these sounds: traditional signal processing methods and new techniques from deep semi-supervised learning. Both techniques showed moderate success at identifying wheezes and crackles.

Second, we evaluated the diagnostic potential of detecting wheezes and crackles and compared it to using signal processing analysis of lung sounds to directly detect pulmonary disease. We showed that this new technique leads to improved diagnostic accuracy. This finding indicates that future research should focus less on lung sound identification.

Third, we combined measurements from all three components of our kit to predict the diagnosis of patients with pulmonary disease. We showed that most of the diagnostic accuracy of the kit was provided by the peak flow meter and questionnaire combination. Together, these two devices were able to accurately detect patients with asthma and COPD.
After developing the diagnostic algorithms, we built an Android application to guide a user through the necessary data collection to arrive at a diagnosis. The application was designed to create questionnaires and data queries from an externally defined model definition file, allowing the application to be easily repurposed for different classification tasks in medicine and other fields.

Future research will expand the use of the pulmonary diagnostic kit to include additional pulmonary diseases and will test its use in a large-scale field study to determine its accuracy as a screening tool for asthma and COPD. If the results of future trials are consistent with the findings in this thesis, the kit and algorithm combination may provide useful information for improving diagnosis of pulmonary disease.

Thesis Supervisor: Richard R. Fletcher
Title: Research Scientist, D-Lab

Thesis Supervisor: Peter Szolovits
Title: Professor of Electrical Engineering and Computer Science
Acknowledgments

I would first like to thank my thesis advisor, Dr. Richard Fletcher, for his continuous support and guidance throughout my thesis. Without his wide-ranging expertise, it would have been impossible for this project to be successful.

I would also like to thank the members of our team at the Chest Research Foundation for working with us to develop and test the pulmonary diagnostic kit. Dr. Rahul Kodgule was an integral part of study design and execution and provided all of our team’s clinical expertise. Yogesh Thorat and Vandana Vincent tirelessly collected data from patients, enabling this entire project.

For their generous support throughout this research, I would like to thank the Tata Center for Technology and Design and the Vodafone Americas Foundation. I also wish to extend my deep gratitude to Dr. Peter Szolovits for his thoughtful and critical feedback on this thesis. For their efforts on different parts of this research, I would like to thank Vivek Miglani, John Mofor, Daniela Ganelin, and Michelle Lauer.

For his constant advice through this thesis, my enduring gratitude goes to my father. His expert advice and clinical knowledge have been of fundamental value to my graduate studies. Deep thanks also to my mother, who has tirelessly provided edits and feedback on my work through the years. Most importantly, I would like to thank my wife, Eleanor, for her constant support and guidance over the years. She is possessed of a great intelligence and immense work ethic, both of which inspired me throughout the development of this thesis.
## Contents

1 Introduction ........................................................................................................ 15  
  1.1 Pulmonary Disease ......................................................................................... 15  
      1.1.1 The Burden of Pulmonary Disease .................................................... 15  
      1.1.2 Diagnosis of Pulmonary Disease ....................................................... 16  
  1.2 Diagnostic Guidance Systems ........................................................................ 17  
      1.2.1 History of Diagnostic Guidance Systems ........................................ 18  
      1.2.2 Pulmonary Diagnostic Guidance Systems ....................................... 20  
          1.2.2.1 Clinical Algorithms ................................................................. 20  
          1.2.2.2 Machine Learning Algorithms ............................................. 21  
          1.2.2.3 Adventitious Lung Sound Analysis ......................................... 23  
  1.3 Project Overview ............................................................................................ 24  

2 The Mobile Diagnostic Kit .................................................................................. 27  
  2.1 Lung Sound Recording ................................................................................... 29  
      2.1.1 Digital Stethoscope ......................................................................... 29  
      2.1.2 Recording Application ..................................................................... 30  
  2.2 Augmented Reality Peak Flow Meter ........................................................... 31  
  2.3 Symptom and Risk Factor Questionnaire ..................................................... 35  

3 Identification of Adventitious Lung Sounds ......................................................... 37  
  3.1 Auscultation for Pulmonary Diseases ............................................................. 37  
      3.1.1 Types of Adventitious Lung Sounds ............................................... 38  
          3.1.1.1 Wheezes .............................................................................. 38
3.1.1.2 Crackles ........................................... 39
3.1.1.3 Other Adventitious Lung Sounds .............. 39
3.2 Data Collection and Methods .......................... 39
3.3 Data Analysis and Results ............................. 41
  3.3.1 Application of Signal Processing Techniques .... 41
     3.3.1.1 Features Used for Classification .......... 41
     3.3.1.2 Signal Processing Classification Approach .... 45
     3.3.1.3 Signal Processing Results .................. 47
  3.3.2 Application of Deep Learning Techniques ......... 48
     3.3.2.1 Semi-Supervised Learning .................. 48
     3.3.2.2 Data Preprocessing .......................... 49
     3.3.2.3 Deep Learning Classification Approach ....... 50
     3.3.2.4 Deep Learning Results ..................... 50
3.4 Discussion ............................................ 51
  3.4.1 Findings .......................................... 51
     3.4.1.1 Practical Considerations for Future Researchers . . 53
  3.4.2 Limitations ....................................... 53
  3.4.3 Future Work ....................................... 54
4 Using Lung Sound Analysis for Pulmonary Disease Diagnosis 57
  4.1 Diagnostic Utility of Identifying Adventitious Lung Sounds 57
     4.1.1 Adventitious Lung Sound Overlap Between Diseases .... 58
     4.1.2 Challenges in Accurately Identifying Adventitious Lung Sounds 59
  4.2 Lung Sound Analysis as an Alternative to Adventitious Lung Sound Identification .................................. 61
  4.3 Data Collection and Methods .......................... 62
  4.4 Data Analysis ......................................... 63
     4.4.1 Spatial Analysis of Adventitious Lung Sounds .... 63
     4.4.2 Signal Processing Lung Sound Analysis ........... 66
  4.5 Classification Algorithms and Results .................. 66
4.5.1 Classification Algorithms ........................................... 66
4.5.2 Classification Performance ....................................... 67
4.5.3 Feature Evaluation .................................................. 68
4.6 Discussion ................................................................ 71
  4.6.1 Findings .............................................................. 71
  4.6.2 Limitations ........................................................... 71
  4.6.3 Future Work .......................................................... 72

5 Diagnostic Guidance for Asthma and COPD 73
  5.1 Data Collection and Methods ....................................... 73
  5.2 Data Analysis ........................................................... 74
    5.2.1 Features for Classification ...................................... 75
    5.2.2 Classification Algorithms ....................................... 77
  5.3 Results ..................................................................... 79
    5.3.1 Detecting Unhealthy Patients ................................. 79
    5.3.2 Detecting Obstructive Diseases .............................. 81
    5.3.3 Detecting Asthma and COPD ................................. 81
    5.3.4 Combined Results ................................................. 82
    5.3.5 Feature Importance ............................................... 83
  5.4 Discussion .................................................................. 87
    5.4.1 Findings .............................................................. 87
    5.4.2 Limitations ........................................................... 88
    5.4.3 Future Work .......................................................... 88

6 Android Implementation of a Pulmonary Screening Tool 91
  6.1 Pulmonary Screening Application ................................. 91
    6.1.1 Machine Learning Models in the Application .......... 91
    6.1.2 Application Workflow ........................................... 94
  6.2 Integration with Other Applications .............................. 96
    6.2.1 Querying External Applications .............................. 96
      6.2.1.1 Peak Flow Meter Application .......................... 96
6.2.1.2 Lung Sound Recorder .................................. 97
6.2.2 Screening from an External Application .......... 97
6.3 Discussion .................................................. 98

7 Conclusions .................................................. 101
7.1 Contributions of this Work ................................. 101
  7.1.1 Machine Learning Contributions ....................... 101
  7.1.2 Clinical Contributions ................................ 103
  7.1.3 Mobile Application Contributions ..................... 104
7.2 Future Work .............................................. 105
List of Figures

2-1 The mobile diagnostic kit .................................................. 28
2-2 Electronic stethoscope .................................................... 30
2-3 Frequency response of the electronic stethoscope .................. 31
2-4 Screenshots from the data recording mobile application .......... 32
2-5 Peak flow meter modified with an augmented reality target ....... 33
2-6 A nurse using the Augmented Reality Peak Flow Meter .......... 33
2-7 Augmented reality application capturing a reading from the peak
flow meter ............................................................................. 34
2-8 Display of past peak flow meter readings in the augmented reality
application ............................................................................. 35
2-9 Bland-Altman plot showing the difference between augmented reality
readings and manual readings for the user testing task ............... 35

3-1 The different auscultation locations ...................................... 40
3-2 Spectrograms for the different types of lung sounds ............... 45
3-3 Best features from the denoising autoencoder by identification task . 51
3-4 Denoising autoencoder classification performance by identification task 52

4-1 Information lost by focusing on adventitious lung sound detection . 60
4-2 Explanation of map of spatial distribution of adventitious lung sounds 65
4-3 Maps of spatial distribution of wheezes for asthma and COPD ...... 65
4-4 Maps of spatial distribution of crackles for asthma and COPD ...... 65
4-5 Feature importance for detection of asthma and COPD using adventitious lung sound identification features ...................... 69
4-6 Feature importance for detection of asthma and COPD using lung sound analysis features ........................................... 70

5-1 Stages of the Asthma and COPD Identification Algorithm ........... 78
5-2 Feature coefficients for detection of unhealthy patients using the questionnaire and peak flow meter ......................... 84
5-3 Feature coefficients for detection of patients with obstructive disease using the questionnaire and peak flow meter .......... 85
5-4 Feature coefficients for differentiating patients with asthma and COPD using the questionnaire and peak flow meter ........... 86

6-1 Screening application overview ........................................... 92
6-2 Screenshots from the pulmonary screening application ............. 95
6-3 Connection between screening and peak flow meter applications ... 97
6-4 Connection between screening and lung sound recorder applications . 97
6-5 Connection between external applications and the screening application 98
List of Tables

1.1 Deaths and lost DALY from pulmonary diseases ........................................ 16
3.1 Number of lung sounds of each type by recording area ................................ 40
3.2 Descriptions of time domain features ......................................................... 42
3.3 Descriptions of frequency domain features .................................................. 43
3.4 Descriptions of time frequency domain features .......................................... 44
3.5 Description of time-frequency domain patch features ................................... 46
3.6 Adventitious identification performance by machine learning classifier ............ 47
3.7 Wheeze identification performance by recording location and machine learning classifier .......................................................... 47
3.8 Crackle identification performance by recording location and machine learning classifier .......................................................... 48
4.1 Summary of patients with each diagnosis ....................................................... 62
4.2 Percent of patients with lung sounds of each type by area and disease .............. 63
4.3 Description of features from adventitious lung sound identification ................. 64
4.4 Classification performance for detection of asthma and COPD using adventitious lung sound labels ................................................. 67
4.5 Classification performance for detection of asthma and COPD using signal processing lung sound features ........................................ 68
5.1 Summary of patients with each diagnosis ....................................................... 75
5.2 Description of questionnaire features .......................................................... 76
5.3 Classifier performance for detecting unhealthy patients by kit component ........................................ 80
5.4 Classifier performance for detecting obstructive diseases by kit component ........................................ 81
5.5 Classifier performance for differentiating asthma and COPD by kit component .................................... 82
5.6 Classifier performance for diagnosis patients using the questionnaire and peak flow meter ..................... 83
Chapter 1

Introduction

1.1 Pulmonary Disease

1.1.1 The Burden of Pulmonary Disease

Pulmonary diseases, or diseases of the lungs, are an increasing global health burden and significant cause of morbidity and mortality. In 2012, these diseases were responsible for more than 9.6 million deaths and 368 million lost disability-adjusted life years (DALYs) worldwide. Pulmonary diseases are projected to cause more than 12.5 million deaths in 2030. The burden of these diseases is particularly acute in the developing world, where pulmonary diseases are responsible for nearly 19% of deaths and 14% of lost DALYs [1].

The World Health Organization (WHO) defines six major types of pulmonary disease: tuberculosis (TB), lower respiratory infections, upper respiratory infections, cancers, chronic obstructive pulmonary disease (COPD) and asthma. Although TB is a lower respiratory infection, its impact is substantial enough to warrant its own category. The death and disability caused by each of these types is summarized in Table 1.1.
Table 1.1: Deaths and lost DALY from pulmonary diseases. Data derived from WHO Data Repository[1]

<table>
<thead>
<tr>
<th>Disease</th>
<th>Deaths (Percent of Worldwide Total)</th>
<th>Lost DALYs (Percent of Worldwide Total)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chronic obstructive pulmonary disease</td>
<td>3,102,606 (5.6%)</td>
<td>92,360,208 (3.4%)</td>
</tr>
<tr>
<td>Lower respiratory infections (not including TB)</td>
<td>3,051,318 (5.5%)</td>
<td>146,857,373 (5.3%)</td>
</tr>
<tr>
<td>Trachea, bronchus, lung cancers</td>
<td>1,599,311 (2.9%)</td>
<td>38,532,719 (1.4%)</td>
</tr>
<tr>
<td>Tuberculosis</td>
<td>934,838 (1.7%)</td>
<td>43,649,849 (1.6%)</td>
</tr>
<tr>
<td>Other respiratory diseases</td>
<td>551,158 (1.0%)</td>
<td>19,528,621 (0.7%)</td>
</tr>
<tr>
<td>Asthma</td>
<td>386,309 (0.07%)</td>
<td>25,202,293 (0.9%)</td>
</tr>
<tr>
<td>Upper respiratory infections</td>
<td>6,038 (0.01%)</td>
<td>2,072,253 (0.1%)</td>
</tr>
</tbody>
</table>

1.1.2 Diagnosis of Pulmonary Disease

To diagnose pulmonary diseases, physicians in the developed world rely on significant testing beyond the physical examination. For the diagnosis of asthma and COPD, for example, the American Thoracic Society (ATS) and European Respiratory Society (ERS) recommend conducting Pulmonary Function Testing (PFT). PFT consists of spirometry, body plethysmography, and a measurement of diffusing capacity for carbon monoxide (DLCO) [2]. To diagnose tuberculosis, physicians rely on a combination of skin tests, radiograph (x-ray) results, and sputum tests [3]. Upper and lower respiratory infections are diagnosed through a combination of information about symptoms, cultures, radiograph results, and other lab tests. Cancer is diagnosed through a combination of radiographs, high-resolution computed tomography (HRCT), and biopsy.

In many regions of the developing world, physicians are unable to use these diagnostic tests. This is because either the testing equipment is unavailable or there are no technicians trained to conduct the tests and interpret the results. As a result, a diagnosis is usually made based on clinical history and auscultation findings (i.e. listening with a stethoscope).
Pulmonary diagnostics in developing countries are complicated by the fact that many primary care doctors practice various forms of non-Western medicine, such as Ayurveda and Homeopathy [4]. The diagnosis of pulmonary disease is neglected in many schools of medicine. In addition, hospitals and clinics are overburdened, with insufficient physician staffing to see all patients, so diagnosis and care is often provided by nurses and health workers [5].

All of the aforementioned leads to significant misdiagnosis and underdiagnosis of pulmonary disease. In an international study of COPD, researchers found that only 26.4% of patients with COPD had undergone pulmonary function testing. More than 80% of patients with COPD were undiagnosed [6]. In India, a multi-center study found that more than 80% of child asthmatics were undiagnosed [7].

The underdiagnosis of asthma and COPD is particularly important because these are chronic conditions that have an ongoing impact on mortality and quality of life. Detecting them as early as possible is critical, because proper therapy can ameliorate the impact of these diseases. Much of the lung function deterioration for COPD occurs in the early stages of the disease. By identifying patients with COPD sooner, physicians and health workers can intervene to preserve lung function and reduce mortality [8]. Identifying asthma early is even more important because untreated asthma greatly increases the subsequent risk of developing COPD [9]. By helping patients control their asthma early, physicians could greatly reduce the future disease burden of COPD.

1.2 Diagnostic Guidance Systems

Diagnostic Guidance systems are tools used to improve the diagnostic capabilities of physicians and health workers. These systems differ from specific diagnostic tests because they dictate the order of a number of clinical investigations and combine multiple findings to arrive at a diagnostic recommendation. In general, these systems are an attempt to formalize the clinical decision-making process and reduce the risk of misdiagnosis.
1.2.1 History of Diagnostic Guidance Systems

The simplest diagnostic guidance systems take the form of the clinical algorithm. Clinical algorithms are typically represented as flow charts. For these algorithms, each step describes a question to ask or test to run on the patient. The flow chart outlines the overall sequence of questions and tests, which can change based on the answers in specific steps. After some number of steps, the flow chart can recommend a diagnosis of a particular disease or follow-up actions, such as different treatment options or referrals [10].

Szolovits and Pauker argue that while flowcharts work well for many types of routine diagnosis, more sophisticated techniques are necessary for many tasks in medicine. They describe the following set of attributes that render flowcharts inadequate for a particular decision-making task [11]:

- Multiple decision factors
- Uncertainty around decision factors
- Nonlinear importance of factors that may depend on each other
- Cost of gathering information is high

In these cases, more sophisticated diagnostic-guidance techniques are required.

In the 1970s, three approaches were used to develop diagnostic guidance systems in medicine. The first approach, expert systems, sought to codify the knowledge of experts into a series of rules which could be applied to determine diagnoses. This approach did not work well in medicine because the interactions between large databases of rules were hard to predict and often resulted in inaccurate diagnoses. The second approach was the development of disease matching algorithms to generate hypotheses about possible diseases, and then determine how well each hypothesis explains the available information. This approach was poorly adaptive to handling variation in the presentation or progression of diseases and so had limited utility in the clinical setting. To address this limitation, researchers
developed pathophysiological models of disease which would allow them to create causal models of disease and the presentation and progression of symptoms [12].

Over the subsequent 20 years, researchers continued to explore combinations of these three approaches, and a number of systems were developed. Shortliffe cites four early examples of diagnostic guidance systems in medicine. Two of these approaches were expert systems. MYCIN was an expert system designed to identify bacteria responsible for causing infection [13]. Internist-1 was an expert system designed to diagnose a wide variety of diseases that might present to a physician of internal medicine [14]. The Present Illness Program used the second approach to evaluate patients with edema [15]. CASNET was an example of the third approach, and linked a model of disease state to symptoms for the diagnosis of glaucoma patients [16].

In 1992, Shortliffe argued that although artificial intelligence in medicine had already been successful in tackling some clinical problems, widespread adoption was limited for a variety of reasons. The primary reason for lack of adoption was that algorithms were typically developed to target a single disease or class of diseases. However, physicians are faced with many different types of diseases and do not have the time to learn a new system for each disease [17]. To create successful systems, more focus needed to be placed on developing platforms that have broad utility within a consistent user interface and shared data.

In the last ten years, improvements in both computational power and data availability have led to advancements from the domain of machine learning [18]. Machine learning techniques are useful for uncovering hidden patterns in large datasets and using these discovered patterns to make future predictions. In medicine, this means that if enough data are available, machine learning techniques can be used to find combinations of lab measurements, symptoms, and risk factors that are indicative of different diseases. Szolovits notes two limitations of machine learning for diagnostics. First, these techniques often uncover only shallow relationships between variables of interest, neglecting complex interaction effects and failing to identify causal relationships. Second, these techniques often fail to build upon
the extensive body of knowledge developed by practitioners in the field [18].

Although these critiques are true, models developed using machine learning techniques can still be valuable in some medical situations. In some cases, the current standard for diagnosis is so low that learning even shallow relationships between variables of interest can lead to screening and diagnostic algorithms that are a large improvement over the status quo. In India, for example, where many of the medical visits are to primary care physicians and health workers, chronic conditions are often ignored because some physicians and health care workers do not have any experience diagnosing and managing these conditions. To improve the accuracy of any model, algorithms can be built collaboratively by engineers and physicians, leveraging the extensive experience and knowledge of physicians to determine which data to collect, how to process that data, which features to extract, and how to determine if the model is working correctly.

This approach works well with the recent resurgence in the use of clinical algorithms and guideline-based medicine. This resurgence has been caused by a desire to reduce variability in the quality of care provided by different practitioners [18]. These guidelines are often created to fit on a few sheets of paper and resemble the flowcharts developed in the early stages of diagnostic guidance research [19, 20, 21, 22]. If these guidelines are instead implemented on an electronic device, it becomes easier to create complicated guidelines that take into account the relationships between many different variables and customize the order of tests and questions before providing diagnostic guidance.

1.2.2 Pulmonary Diagnostic Guidance Systems

1.2.2.1 Clinical Algorithms

A number of clinical algorithms have been developed to improve the diagnosis of pulmonary diseases. Primary Care 101 is a step-by-step guide to the diagnosis of diseases often encountered by primary care physicians and health workers in South Africa. It is primarily symptom-based, but also includes some physiological mea-
measurements, such as heart rate, respiratory rate, and blood pressure. It was designed as a general-purpose diagnostic algorithm, but has sections on asthma, COPD, and tuberculosis [19]. Because the algorithm is general-purpose and implemented in paper form, it neglects core components of pulmonary disease diagnostics: analysis of lung function and auscultation.

The Allergic Rhinitis Management Pocket Reference was designed to facilitate the diagnosis and management of allergic rhinitis (AR) at the primary-care level. It provides both a list of symptoms suggestive of AR and symptoms that would indicate that AR-like symptoms are caused by some other condition. Then, if AR is suspected, the Pocket Reference provides flowcharts for both primary care physicians and specialists to confirm the diagnosis [20]. The AR Management Pocket Reference is useful for identifying allergic rhinitis, but it does not cover other pulmonary diseases.

The Institute for Clinical Systems Improvement has created diagnostic guidelines for both the diagnosis and treatment of respiratory illness and the diagnosis and management of asthma [21]. For respiratory illness, the system is primarily symptom-based. For asthma, the diagnostic criteria relies on spirometry and user-conducted auscultation [22]. Because these guidelines are disease-specific, it is still up to the user to determine when they are appropriate. In areas with limited pulmonary disease training, an overarching system needs to be created to guide the entire diagnostic process.

1.2.2.2 Machine Learning Algorithms

While there have been a few previous studies that sought to develop diagnostic guidance algorithms for pulmonary disease, this topic has been relatively understudied. Most of these studies tried to use information taken from lung sound recordings to estimate diagnoses. However, only one of them incorporated information about symptoms and risk factors that can be captured during a basic clinical exam. Additionally, while one of them used the results of pulmonary function testing to improve diagnosis, none of them tested the much less costly peak flow meter as an
alternative device.

Kahya et al. created a system with three classifications for individuals: healthy, obstructive airway disease, and restrictive airway disease [23]. The system used features extracted from recorded lung sounds. Features of the lung sounds included the coefficients of autoregressive models and the coefficients of Prony's method, a method of fitting damped sinusoids to a waveform. The classifiers had limited success, with the best combination of features and classification architecture giving a 71% accuracy in detecting the three types of patients.

In more recent work, Kahya et al. split recorded sound files into respiratory subphases using measurements taken simultaneously from a flow meter [24]. Each segment was fit to an autoregressive model and the system used these coefficients as well as information taken from the flow meter as inputs to an artificial neural network. The network achieved some success in identifying sound segments belonging to each disease class, but no results were presented to describe the accuracy of the system for individual patients.

Yamashita et al. created a system for differentiating between patients with emphysema and healthy subjects using analysis of lung sounds [25]. Recorded lung sounds were segmented, and the system determined whether each segment contained an adventitious lung sound. Then, if an adventitious lung sound was identified with enough certainty, patients were considered unhealthy and diagnosed with emphysema. This approach was able to accurately identify 89% of patients. However, because this study was limited to a single disease, there was no estimate of diagnostic accuracy in differentiating among diseases. The same lung sounds can occur across many diseases, so the accuracy of this method would likely decrease as more diseases are added to the study.

Sen et al. used vector autoregressive processes to estimate coefficients from a multichannel lung sound recording [26]. The coefficients were used in a support vector machine to determine whether patients were healthy, had bronchiectasis, or had interstitial lung disease. Their classification scheme did well for the recognition of interstitial pulmonary disease but did not work well for bronchiectasis.
Gavriely et al. explored combining analysis of lung sounds, a questionnaire, and measurements from PFT to identify patients with asthma, COPD, restrictive lung disease, congestive heart failure, and emphysema [27]. They were able to classify patients as having pulmonary disease with an accuracy of 87%, but did not attempt to perform disease-specific classification.

1.2.2.3 Adventitious Lung Sound Analysis

While there has been limited work on the development of diagnostic guidance algorithms, significant effort has been made to create systems and algorithms capable of automatically detecting adventitious lung sounds. In 2009, Hadjileontiadis published an overview of methods that have been used for lung sound analysis [28]. More recently, in 2013, Palaniappan et al. published a review detailing recent developments in automated lung sounds analysis [29]. Some of the most common methods for lung sound analysis are described here.

Starting in the 1950s, researchers recorded lung sounds and created spectrograms to analyze the time-frequency domain behavior of chest breathing sounds [30]. In the 1980s and 1990s, researchers analyzed subjects with wheezes, crackles, and no abnormalities to define typical characteristics that could aid in automatic detection [31, 32, 33, 34]. In addition to describing crackles, Piirila and Sovijarvi explored the clinical significance of crackle location within the breath phase [34].

As computerized analysis has become more accessible by advances in technology and computing power, researchers used a number of methods to automatically identify adventitious lung sounds. To identify wheezes, most prior work took advantage of the fact that wheezes are sustained musical sounds and researchers sought to detect them by identifying their unique characteristics in the frequency domain. Some researchers used both the Fourier Transform and Short-Time Fourier Transform to identify features that are indicative of wheezing [35, 36, 37]. As an alternative to sinusoidal basis functions, researchers explored the use of different wavelets and their associated coefficients to generate features for classification [38]. Some researchers also attempted to use features present in the time-domain signal.
to identify wheezes [39].

To identify crackles, which are short bursts of explosive sounds caused by the opening of small airways collapsed by fluid, past efforts focused on identifying the characteristic shapes of crackle sounds in both the time and frequency domains. Murphy et al. used combinations of amplitude, duration, and frequency of sound to automatically identify crackles [40]. Yi searched for high amplitude spikes in the time domain and compared the value of those peaks to the mean amplitude of the signal, identifying anything over a predetermined threshold as a crackle [35]. Yeginer and Kahya used a combination of amplitude, duration, and correlation with the Daubechies wavelet to identify crackle sounds [41].

While a number of methods have been proposed for both crackle and wheeze analysis, these methods have been tested on small numbers of patients and sound recording files that have been taken specifically for research. It is unclear whether these findings will generalize to recordings taken in more general clinical settings. No standard datasets exist, so it is difficult to compare results across studies. These algorithms often are not tested on data collected outside the research laboratory and, as a result, might not perform as well with lower-quality data collected from inexpensive microphones.

1.3 Project Overview

In this project, we sought to build on past efforts to create pulmonary diagnostic systems. Our goal was to develop tools that could be used by primary care physicians and health workers to screen for, and diagnose, pulmonary disease. The tools were targeted at these users because these individuals are responsible for the majority of patient-provider interactions in India and other regions of the developing world. Because these users have significant resource limitations, the cost of the tools was a major factor throughout the development of the diagnostic kit. To keep marginal costs low, we decided to rely as much as possible on software solutions.

To take advantage of equipment that is already in the field, we focused on
tools that could be deployed on a smartphone. In the developing world, individuals are already acquiring smartphones at a rapid pace because of the many capabilities they provide [42]. The computational and processing power of these devices can be used as the base for the mobile diagnostic kit described in the rest of this thesis.

This work focused on the identification of patients with asthma and COPD, although the approach is currently being applied to additional pulmonary diseases. Together, asthma and COPD are responsible for 5.7% of lost lives and 4.3% of lost DALYs worldwide each year. Because these illnesses are chronic, physicians and health workers could use the mobile diagnostic kit to stage large-scale, long-term screening campaigns to find suffering patients and ensure that they receive therapy. Additionally, collecting data from patients with chronic diseases simplifies the data collection process, because patients can be referred to a central facility to determine a reference standard diagnosis with little risk of change in their condition.

In Chapter 2, I describe three tools that we created to collect data from pulmonary disease patients: an electronic stethoscope, an augmented reality peak flow meter, and a digital questionnaire. In Chapter 3, I discuss methods that we developed to automatically analyze recorded lung sounds and detect wheezes and crackles. In Chapter 4, I propose new methods for lung sound analysis, moving beyond the identification of wheezes and crackles to tie aspects of recorded sounds directly to diagnosis. In Chapter 5, I present the screening and diagnostic algorithms developed to identify patients with asthma and COPD using data collected with the mobile diagnostic kit. In Chapter 6, I describe an implementation of the screening and diagnostic algorithms in a standalone application for the Android platform. In Chapter 7, the final chapter, I discuss the potential of the diagnostic kit and the remaining steps that must be taken before it can be deployed to primary care physicians and health workers.
Chapter 2

The Mobile Diagnostic Kit

In this chapter, I evaluate how lung sounds could be used for diagnosis. I describe some of the limitations associated with using adventitious lung sound identification and propose a new approach: using machine learning and signal processing methods that can be used to tie lung sound analysis directly to disease diagnosis. I present data on the frequency of lung sounds within patients with different diseases and show that the discriminatory power of the presence of adventitious lung sounds is limited. Next, I propose methods to tie analysis of lung sounds directly to diagnosis, without attempting to identify adventitious lung sounds. I evaluate these methods on data collected from patients in Pune, India. Finally, I close by discussing the implications of these findings. To improve the diagnosis of pulmonary disease, we developed a mobile diagnostic kit that could be deployed in areas that are lacking in the tools traditionally used to diagnose pulmonary disease. In this chapter, I describe the design, development and validation of the components of the diagnostic kit.

The mobile diagnostic kit was designed to capture three types of information. First, the kit records lung sounds using a custom-built electronic stethoscope. Second, the kit measures peak expiratory flow, which provides important clues about lung function and volume. Third, the kit captures information about patient symptoms and risk factors for different pulmonary diseases. A picture of the kit is shown in Figure 2-1.
Figure 2-1: The mobile diagnostic kit
We selected these three types of data after consulting with our clinical partners at the Chest Research Foundation in Pune, India. While we considered mobile implementations of traditional PFT devices, we identified these three data collection tools as less complex and inexpensive alternatives that could potentially provide accurate diagnoses. Although they do exist, we decided not to include mobile spirometers because they are costly and require significant training before they can be accurately administered. In addition, spirometry requires a large effort on the part of the patient to obtain meaningful measurements and consequently cannot be used on children or the elderly.

2.1 Lung Sound Recording

2.1.1 Digital Stethoscope

The first component of the kit was an electronic stethoscope, which allows information to be captured about a patient’s lung sounds. To construct an electronic stethoscope, an Indian-sourced stethoscope was retrofitted with an electret microphone. The stethoscope selected was the Micro-Tone model by Malhotra Surgical Industries, which sells for approximately US $20 and is commonly available in India. The microphone used was an omnidirectional electret condenser microphone (CUI Inc. CMC-2742WBL-25L).

The stethoscope pneumatic tubing was cut halfway between the head-piece and the earpieces and the earpieces were discarded. The headpiece was removed from the tubing and a standard two-conductor insulated wire was threaded through the tubing. The microphone was soldered to the two conductor wire on the headpiece side of the tubing and the other side of the wire was soldered to a 3.5mm male Mono plug. The microphone was pushed into the open tubing until it was far enough in for the headpiece to be reinserted. After the headpiece was reinserted, the electronic stethoscope was complete. A splitter was constructed to split the microphone and headphone connections on the Android phone. This allowed
headphones to be connected at the same time as the stethoscope. A picture of the completed stethoscope is shown in Figure 2-2.

We evaluated the stethoscope by measuring its frequency response. We generated continuous wave sinusoids in MATLAB on a PC with frequencies from 60 to 1500 Hz and played them through a wide-band Bose Wave Music System. The sound level was measured on the mobile phone using the Android application "Sound Level Meter". We selected this application because it is accurate enough to meet OSHA's type 2 instrument standard [43]. The sound level of the produced sound files was corroborated with a baseline reading from a commercial sound level meter (YFE YF-20), which read 80 dB(a) at 500 Hz. The frequency response of the stethoscope is shown in Figure 2-3. The frequency response was relatively flat over the lung sound range of interest (75-1000 Hz [33]), so the stethoscope was able to record clinically relevant sounds.

2.1.2 Recording Application

We designed an Android application to interface with the mobile stethoscope and record lung sounds for analysis. The application is named "Lung Sound Recorder" and can be found in the Android Play Store [44]. Screenshots from the application
are shown in Figure 2-4. The application includes a patient log-in screen and a graphical interface that enables the user to select body locations and record sounds. Recordings can be captured from up to eleven locations.

After the user selects a recording location, the mobile application presents the user with a recording screen that displays a real-time plot of the sounds being recorded by the microphone. As the sound is being recorded, it is played back through connected headphones. Each recorded sound file is thirty seconds long, sampled at a rate of 44100 Hz, and saved in WAV format. The user can record and play back sounds multiple times and select which files to save to the phone. Currently, patient data is stored by patient ID number instead of patient name, because the sound files are not encrypted and we assumed that different medical facilities will have different protocols for protecting patient data.

2.2 Augmented Reality Peak Flow Meter

The second component of the kit was a peak flow meter, which is an inexpensive mechanical device that is used to assess the volume and strength of a patient’s lungs. These devices measure a patient’s peak expiratory flow rate (PEFR). Previous research has shown the PEFR measurements can be used to screen for asthma and COPD [45]. However, because the peak flow meter is an analog device, the readings need to be digitized for analysis. To increase the reliability of digitization
Figure 2-4: Screenshots from the data recording mobile application. From the top left are the log-in screen, the recording location selection for both the front and back, and a sample real-time waveform plotted as a sound file is recorded.
and minimize the effects of illiteracy on the process, we developed an augmented reality application to automatically digitize peak flow meter readings.

To prepare the peak flow meter for digital capture, a custom-printed sticker was attached as an augmented reality target. The augmented reality target is a pattern that is readily distinguishable from the background world. The modified peak flow meter is shown in Figure 2-5. A nurse using the Augmented Reality Peak Flow Meter is shown in Figure 2-6.

We developed the augmented reality application for Android smartphones using the Vuforia™ Augmented Reality SDK. The application allows the user to
enter their age, height, and gender. Using this information, the application computes the personalized reference values for comparison with the user’s measured PEFR. In order to capture a reading, the user aims the smartphone at the peak flow meter. The application automatically detects the augmented reality target and searches for the red measurement indicator. The position of the indicator is compared to the target and the lateral distance between them is used to determine the measurement value. The detected measurement is displayed on the screen. Underneath the peak flow meter indicator, the application displays green, yellow, and red bars that correspond to the healthy, potentially unhealthy, and unhealthy measurement zones defined by the American Lung Association. The application is shown capturing a reading in Figure 2-7.

Once the measurement has been detected, the user clicks “Continue” and the reading is saved to the device. Image data is not saved, thus minimizing memory use. Each reading is saved with a timestamp and all data is saved in CSV format. The user is given the option to export all of their readings using any of the sharing options enabled on their device (text message, email, etc.) The user may also view a plot of their historical readings in the application. The application display of past readings is shown in Figure 2-8.

To validate the augmented reality approach, the peak flow meter and application were provided to a doctor and two nurses in India. The subjects were provided with one minute of training in the use of the device. The peak flow meter was set to ten randomly selected readings for each subject and subjects were asked
Figure 2-8: Display of past peak flow meter readings in the augmented reality application

Figure 2-9: Bland-Altman plot showing the difference between augmented reality readings and manual readings for the user testing task
to digitize the readings.

A Bland-Altman plot of the results is shown in Figure 2-9. The mean error was 5.5 L/min with a root mean squared error of 10.3 L/min. The peak flow meter ranges from 60 to 840 L/min with healthy values for 30-year-old men at roughly 600 L/min. In this device, the largest errors occurred at the highest levels of PEFR, where accuracy is less important because all readings are within healthy ranges.

2.3 Symptom and Risk Factor Questionnaire

The final component of the kit was a questionnaire that captured information about a patient’s symptoms and risk factors for pulmonary disease. To design the ques-
tionnaire, we reviewed previous efforts to produce symptom-based diagnostic guidelines [46, 47, 20, 19, 21, 48]. These sources were used to create a comprehensive questionnaire that captures information about the onset, duration, and progress of breathlessness, coughing, chest pain, fever, and nasal symptoms. The questionnaire also captures information about demographics, medical history, and risk factors such as smoking history or family history of allergies.

The original version of the questionnaire was designed to be comprehensive and captures more information than was required for the standalone Android application. As patient data was collected, each question was evaluated using techniques from machine learning to determine its contribution to the final diagnosis. Only questions that provided diagnostic utility were implemented in the final Android version of the questionnaire.
Chapter 3

Identification of Adventitious Lung Sounds

In this chapter, I describe algorithms that can be used to automatically identify adventitious lung sounds. I begin by describing pulmonary auscultation and the different types of adventitious lung sounds. Then, I discuss two different approaches for creating features to identify wheezes and crackles. I evaluate these methods on lung sounds collected from patients in Pune, India. Finally, I compare the two approaches and make recommendations for future research.

3.1 Auscultation for Pulmonary Diseases

Auscultation, the process of listening to a patient’s heart or lung sounds, is one of the oldest and most common medical procedures still in use today. This technique was significantly advanced through the invention of the stethoscope by Rene Laennec 200 years ago. Recently, auscultation has become useful not only for point-of-care diagnostics, but also as a tool for remote patient monitoring and telemedicine [49].

Unlike in other fields of medicine, where many technologies exist to assess patient health, pulmonology still relies on auscultation as the most common method for routinely testing and monitoring a patient’s pulmonary health. Unfor-
Fortunately, auscultation is limited by its subjectivity; the results are subject to the clinician’s auditory acuity, individual training, and interpretation of sounds.

When a physician auscultates a patient, they listen for the presence of adventitious lung sounds. Adventitious lung sounds are sounds that are superimposed on the sounds of healthy breathing. Although these sounds can give important information about the presence of disease, some providers have difficulty detecting them [50]. Primary care physicians are less able to accurately detect these sounds than pulmonary specialists. Because of this, there is an opportunity to develop tools that can improve the detection of adventitious lung sounds at the primary and sub-primary care levels.

3.1.1 Types of Adventitious Lung Sounds

There are many conflicting definitions of the different classes of adventitious lung sounds (Section 4.1.2). Two of the most prevalent adventitious lung sounds are wheezes and crackles. Other lung sounds include stridor, pleural rubs, rhonchus, and squawks.

3.1.1.1 Wheezes

Wheezes are harmonic lung sounds caused by obstructed airways that vibrate as air passes through them [51]. Gavriely et al. proposed that wheezes are caused when the speed of airflow reaches a critical velocity and induces fluttering in the airways [32]. According to Sarkar et al., without airflow limitation fluttering cannot occur and so wheezes will not be produced [52]. Sarkar describes wheezes as “musical notes”. Wheezes can be caused by local obstructions or inflammation. Wheezes tend to occur in larger airways because airflow speed is not sufficient in small airways. Both asthma and COPD create generalized airway obstructions so wheezes can be heard at many different auscultation locations [52].
3.1.1.2 Cracks

Crackles are short, explosive breath sounds that are caused by the rapid opening of abnormally closed airways [34]. According to Piirila, crackles are associated with a number of pulmonary diseases, including pulmonary fibrosis, bronchiectasis, COPD, pulmonary edema, and pneumonia. In the early stages of disease, crackles are only heard in the lower parts of the lungs, but they become more prevalent in the upper airways as diseases progress [34].

3.1.1.3 Other Adventitious Lung Sounds

While there is general agreement about the definition of wheezes and crackles, there are many different definitions of less prevalent adventitious lung sounds. Sarkar et al. identify stridor and pleural rubs as two additional types of lung sounds [52]. A stridor "is a loud, high-pitched, musical sound produced by upper respiratory tract obstruction". Pleural rubs, caused by inflammation of the pleura, are "non-musical, short explosive sounds" that are characterized by their "grating, rubbing, creaky, or leathery" qualities [52].

Bohadana et al. identify rhonchus and squawks as additional adventitious lung sounds [53]. Rhonchus are "musical, low-pitched" sounds. Squawks are short wheezes immediately preceded by crackles [53]. Melbye notes that the term rhonchus is used interchangeably with low-pitched wheeze [54].

3.2 Data Collection and Methods

The electronic stethoscope was used to collect lung sound data from 204 patients at the Chest Research Foundation in Pune, India. First, the physician manually auscultated each patient at eleven recording locations (Figure 3-1) and noted the presence of any abnormal lung sounds. Technicians then used the stethoscope and Android application described in Chapter 2 to record thirty seconds of lung sounds from the eleven recording locations on the trachea, chest, and back. The number of abnormal lung sounds for each area are shown in Table 3.1.
Figure 3-1: The different auscultation locations

Table 3.1: Number of lung sounds of each type by recording area

<table>
<thead>
<tr>
<th>Area</th>
<th>Normal</th>
<th>Wheeze</th>
<th>Crack</th>
<th>Wheeze and Crack</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>289</td>
<td>41</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>289</td>
<td>37</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>288</td>
<td>42</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>282</td>
<td>37</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>282</td>
<td>39</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>270</td>
<td>40</td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>265</td>
<td>39</td>
<td>27</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>269</td>
<td>44</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>266</td>
<td>45</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>284</td>
<td>41</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>284</td>
<td>43</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>
3.3 Data Analysis and Results

As described in Section 1.2.2.3, there has been much prior work on the development of algorithms to automatically detect adventitious lung sounds. The traditional approach has been to collect a set of patient data and develop signal processing methods capable of detecting adventitious lung sounds. However, because many of the past studies had small numbers of subjects, it is difficult to assess whether the results would generalize.

To answer this question, we engineered a set of signal processing features inspired by past work in the field. Then, we used machine learning methods to train models capable of automatically identifying wheezes and crackles. We assessed the ability to generalize to new data by using standard machine learning techniques.

In addition to using this traditional feature engineering approach, we used recently developed methods from deep learning to learn a set of features from our dataset. In other domains, researchers have found that these algorithms can learn features directly from the data that are useful for classification tasks [55]. This approach was tested because it held promise for ready applicability to a wide-range of sound analysis tasks.

3.3.1 Application of Signal Processing Techniques

3.3.1.1 Features Used for Classification

To create features for classifying the different lung sounds, we analyzed the recorded sound files and engineered features that were capable of discriminating between the different lung sounds. Many of these features were influenced by or taken directly from prior research on the identification of abnormal lung sounds. The features extracted using signal processing techniques can be broken into three categories: time domain, frequency domain, and time-frequency domain.
Table 3.2: Descriptions of time domain features

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Feature Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spike Detector Count</td>
<td>First, the sound signal was converted to the envelope. Then, the number of times that the signal amplitude at a specific point was more than two times larger than the mean amplitude in the 0.04 second surrounding window was recorded as the number of spikes.</td>
</tr>
<tr>
<td>Local Relative Amplitude Standard Deviation</td>
<td>First, the sound signal was converted to the envelope. Then, the ratio between the signal amplitude at each point to the mean amplitude in the 0.04 second surrounding window was computed. The standard deviation of this relative amplitude was used as a feature.</td>
</tr>
<tr>
<td>Renyi Entropy [39]</td>
<td>$H_q(x) = -\log \sum_{i=1}^{n} x_i^q$ for each time index $i$ in the sound file.</td>
</tr>
<tr>
<td>Zero Crossing Standard Deviation [39]</td>
<td>First, the time intervals between zero crossings were computed. Then, the standard deviation of these intervals was used as a feature in the model.</td>
</tr>
<tr>
<td>Zero Crossing Standard Deviation Over Mean [39]</td>
<td>The same as the previous feature, but divided by the mean time interval between zero crossings.</td>
</tr>
</tbody>
</table>

**Time Domain Features** Time domain features were the result of analysis performed directly on the raw values of the recorded sound file. Descriptions of the time domain features are shown in Table 3.2. Some of these features were described by Aydore et al. and implemented again here [39].

**Frequency Domain Features** Frequency domain features were developed by analyzing the result of the Fourier Transform. For many of these features, the frequency domain was split into three bands: low frequencies (100-250 Hz), middle frequencies (250-800 Hz) and high frequencies (800-1600 Hz). Descriptions of the frequency domain features are shown in Table 3.3. Some of these features were described by Aydore et al. and implemented again here [39].
Table 3.3: Descriptions of frequency domain features

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Feature Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most Powerful Frequencies</td>
<td>The most powerful frequencies in each of the three frequency bands.</td>
</tr>
<tr>
<td>Relative Power of Most Powerful Freqencies</td>
<td>The relative power of the most powerful frequency in each band divided by the average power of that band.</td>
</tr>
<tr>
<td>Ratios of Mean Power between Bands</td>
<td>The ratios between the mean powers in each of the frequency bands.</td>
</tr>
<tr>
<td>Ratio between f50 and f90 [39]</td>
<td>The ratio between the frequency which was at the 90th percentile for cumulative power and the frequency which was at the 50th percentile for cumulative power.</td>
</tr>
</tbody>
</table>

**Time-Frequency Domain Features**  Time-frequency domain features were developed by analyzing the result of the Short Time Fourier Transform. The spectrogram was computed with a Hamming window of size 0.04 seconds (160 points) and 50% overlap. Then, the spectrogram powers were converted to a log scale. Some of the time-frequency domain features were computed using the mean and variance of time and frequency bins within the spectrogram. These features were computed on both the entire spectrogram and a subset of the spectrogram (250-1000 Hz) previously identified to contain wheeze information. Descriptions of these time-frequency domain features are shown in Table 3.4.

After analyzing the spectrogram, it became apparent that both wheezes and crackles can appear as localized instances of higher than average power. Wheezes appeared as small wave-like patterns. Crackles appeared as vertical bars 3-2. Features that focused on finding these local deviations were developed. The spectrogram was split into patches of fixed width and height. Then, comparisons were made between the power within different patches. Three types of patches were used: square, vertical, and horizontal. Each square patch had dimensions of five frequency bins and five time bins. Vertical patches contained five time bins and all of the frequency bins. Horizontal patches contained five frequency bins and all of the time
Table 3.4: Descriptions of time frequency domain features

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Feature Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectrogram Variance</td>
<td>The variance in power levels across all time-frequency bins in the spectrogram.</td>
</tr>
<tr>
<td>Maximum Variance by Time</td>
<td>The variance of power within each time window was computed. Then, the maximum value across the spectrogram was recorded.</td>
</tr>
<tr>
<td>Maximum Variance by Frequency</td>
<td>The variance of power within each frequency was computed. Then, the maximum value across the spectrogram was recorded.</td>
</tr>
<tr>
<td>Mean Variance by Time</td>
<td>The variance of power within each time window was computed. Then, the mean value across the spectrogram was recorded.</td>
</tr>
<tr>
<td>Mean Variance by Frequency</td>
<td>The variance of power within each frequency was computed. Then, the mean value across the spectrogram was recorded.</td>
</tr>
<tr>
<td>Frequency with highest variance</td>
<td>The variance within each frequency was computed. Then, the frequency with the highest variance was recorded.</td>
</tr>
<tr>
<td>Frequency with highest variance com-</td>
<td>The mean and variance within each frequency was computed. Then, the frequency with the highest variance to mean ratio was recorded.</td>
</tr>
<tr>
<td>pared to mean</td>
<td></td>
</tr>
</tbody>
</table>
Figure 3-2: Spectrograms for the different types of lung sounds

3.3.1.2 Signal Processing Classification Approach

For each adventitious lung sound, we used 80% of the patients as the training dataset and 20% of the patients as a test dataset. This random selection was stratified on whether the patient had the lung sound, so that class sizes were balanced between the training and test sets. The presence of an adventitious lung sound was determined by the physician through manual auscultation at the point of care.

First, a classification model for each adventitious lung sound was trained on all of the recorded lung sounds. This approach combined the data from all recording locations to create the largest possible dataset. Second, sounds were grouped into recordings of the upper or lower airways and separate classifiers were trained on each group. This approach takes into account differences in sounds recorded on different parts of the body. If more data were available, separate classifiers would be created for each recording area, but there were not sufficient record-
**Table 3.5: Description of time-frequency domain patch features**

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Feature Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum value of mean power in patch</td>
<td>The mean value of power in each patch was computed. Then, the maximum was recorded.</td>
</tr>
<tr>
<td>Maximum value of mean power in patch normalized</td>
<td>The mean value of power in each patch was computed. Then, the maximum was computed. The maximum was normalized by the mean power across the spectrogram.</td>
</tr>
<tr>
<td>by mean value in spectrogram</td>
<td></td>
</tr>
<tr>
<td>Variance of mean power in patch</td>
<td>The mean value of power in each patch was computed. Then, the variance was recorded.</td>
</tr>
<tr>
<td>Skewness of of mean power in patch</td>
<td>The mean value of power in each patch was computed. Then, the skewness was recorded.</td>
</tr>
<tr>
<td>Kurtosis of mean power in patch</td>
<td>The mean value of power in each patch was computed. Then, the kurtosis was recorded.</td>
</tr>
</tbody>
</table>

For each set of recording areas, we tested three different types of classification algorithms: logistic regression models with L1 regularization, support vector machines with the radial basis function kernel, and gradient boosting machines. Five-fold cross validation was used to determine the best value of model parameters for each classification scheme. In order to ensure that regularization did not improperly penalize features because they occur on different scales, each feature was normalized so that it had zero mean and unit standard deviation.

For each set of recording locations and classification methodology, we created a receiver-operating-characteristic curve and associated area under the curve. We repeated the entire process of model training and evaluation 100 times with different randomly selected test sets. This repetition allowed analysis of the variability of the final results and their dependence on sampling error. This provides a better estimate of generalizability to previously unseen data. We report the median and inter-quartile range (IQR) of the area under the curve (AUC) values.
### Table 3.6: Adventitious identification performance by machine learning classifier. Performance is reported as median (inter-quartile range).

<table>
<thead>
<tr>
<th>Lung Sound</th>
<th>Classifier</th>
<th>Logistic Regression</th>
<th>Support Vector Machine</th>
<th>Gradient Boosting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wheeze</td>
<td></td>
<td>0.76(0.75-0.78)</td>
<td>0.77(0.75-0.79)</td>
<td>0.73(0.71-0.74)</td>
</tr>
<tr>
<td>Crackle</td>
<td></td>
<td>0.75(0.73-0.78)</td>
<td>0.79(0.75-0.81)</td>
<td>0.78(0.75-0.81)</td>
</tr>
</tbody>
</table>

### Table 3.7: Wheeze identification performance by recording location and machine learning classifier. Performance is reported as median (inter-quartile range).

<table>
<thead>
<tr>
<th>Area</th>
<th>Classifier</th>
<th>Logistic Regression</th>
<th>Support Vector Machine</th>
<th>Gradient Boosting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper Body</td>
<td></td>
<td>0.70(0.67-0.72)</td>
<td>0.74(0.72-0.77)</td>
<td>0.70(0.68-0.73)</td>
</tr>
<tr>
<td>Lower Body</td>
<td></td>
<td>0.74(0.72-0.77)</td>
<td>0.72(0.70-0.75)</td>
<td>0.72(0.69-0.74)</td>
</tr>
</tbody>
</table>

#### 3.3.1.3 Signal Processing Results

**Classification of recordings from all recording areas**  Three types of models were trained to recognize each of the adventitious lung sounds included in this study. These models were trained on a dataset that contained sounds from all of the recording locations. The results from these models are shown in Table 3.6. For wheeze identification, the support vector machine outperforms both of the other algorithms. For crackle identification, the three algorithms perform roughly the same.

**Classification of recordings grouped by upper and lower airways**  The same three types of models were trained on two subsets of the overall dataset. The recordings were grouped into those recorded from the upper and lower airways. The results from these models are shown in Table 3.7 and Table 3.8. For wheeze identification, splitting the dataset into two groups leads to a decrease in performance. For crackle identification, splitting the dataset into two groups leads to improved performance in both groups.
Table 3.8: Crackle identification performance by recording location and machine learning classifier. Performance is reported as median (inter-quartile range).

<table>
<thead>
<tr>
<th>Area</th>
<th>Logistic Regression</th>
<th>Support Vector Machine</th>
<th>Gradient Boosting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper Body</td>
<td>0.85(0.75-0.96)</td>
<td>0.92(0.88-0.95)</td>
<td>0.93(0.85-0.97)</td>
</tr>
<tr>
<td>Lower Body</td>
<td>0.77(0.74-0.79)</td>
<td>0.81(0.77-0.83)</td>
<td>0.79(0.75-0.82)</td>
</tr>
</tbody>
</table>

3.3.2 Application of Deep Learning Techniques

3.3.2.1 Semi-Supervised Learning

In order to address the problem of learning generalizable features, neural network algorithms, and more recently deep learning, now provide a means to automatically explore and discover the underlying structure of a dataset and extract important features. For large datasets where it is not possible to manually label all data, unsupervised learning techniques can be used.

By representing the learning problem as a layered system of neurons which connect inputs to outputs through dynamic weighted connections, it is possible to create an autoencoder algorithm, which will discover the underlying structure of a dataset and map it to a lower-dimensional representation. In its simplest form, an autoencoder will learn the PCA representation of a dataset [56]. A denoising autoencoder (DA) works in a similar fashion, but uses randomly corrupted input data to find a lower dimensional representation. Input data is corrupted and passed to the DA, which then seeks to recreate the original uncorrupted data. This random corruption helps to remove the effects of random noise from the dataset and forces the algorithm to learn better features. By using layers of denoising autoencoders, robust high level features can be learned from a dataset. This technique has been used to successfully extract important features in image classification tasks [55].

When features extracted using unsupervised learning methods are combined with a supervised learning algorithm, the approach is called “semi-supervised” learning. While a few previously-published studies have explored the use of neural
networks for lung sound classification, these studies have used the neural networks only for supervised learning, not for unsupervised learning and feature generation [57, 58, 59].

3.3.2.2 Data Preprocessing

Because of computational constraints, we split the thirty-second sound files into smaller segments. We chose four seconds, because in most cases this would capture at least one respiratory cycle. To reduce boundary effects, we created four-second files with 50% overlap. This resulted in 11,627 four-second sound files.

Because these sound files were subsets of the longer files, there was no guarantee that each short sound file would contain the same adventitious lung sounds heard during manual auscultation. To obtain labels for the shorter sound files, we randomly selected 890 to be reviewed and labeled by the physician. Of these, 390 were found to have inaudible breath sounds, and these were excluded from subsequent analysis.¹

To prepare the data for the DA, we low-pass filtered each signal with a Hamming window filter of order 200 with a cutoff of 1600 Hz and downsampled the recordings to 4000 Hz. Then, each sound file was converted to a spectrogram using the Short-Time Fourier Transform (STFT). For the STFT, we used a Hamming window of length 80 with 50% overlap.

To adjust for differences between the recording locations, we normalized each location separately. Across all recordings for each area, we normalized each frequency bin to have zero mean and unit standard deviation. Each spectrogram was then downsampled by a factor of four to reduce the computation required to train the autoencoder.

The DA was built with three layers of 50 neurons. Each layer used masking noise to corrupt the input, with corruptions levels of 0.1, 0.2, and 0.3 in the

first, second, and third layers, respectively. To compute reconstruction error, we used binary cross-entropy. The DA was implemented in Python using the Theano library [60, 61]. The DA was trained using the entire set of labeled and unlabeled sound files.

### 3.3.2.3 Deep Learning Classification Approach

We trained two support vector machine (SVM) classifiers: one to identify wheezes and one to identify crackles. Both classifiers used a radial basis function kernel. To identify the best subset of autoencoder features, we used greedy forward feature selection. The performance of each classifier was evaluated by computing the Receiver Operating Characteristic (ROC) curve and associated Area Under the Curve (AUC) for 50 randomly generated sets of 5-fold cross-validation splits. To create the ROC curves, the distance from each point in the validation fold to the decision boundary was computed.

### 3.3.2.4 Deep Learning Results

#### Best features from the denoising autoencoder

For each of the classification tasks, we selected the best ten features from the denoising autoencoder. These features are shown in Figure 3-3. Each feature plot shows the weight provided to each pixel in the original spectrogram. Yellow indicates that the neuron has large negative weights for those input pixels. White indicates that the neuron has large positive weights for those input pixels.

The features learned by the DA for both classification tasks contain clearly visible wheeze and crackle events. The vertical yellow and white events are crackle representations learned by the DA. The yellow and white events that are limited in frequency range and repeat horizontally are wheeze representations learned by the DA.

#### Classification performance

If both classifiers were designed to have a true positive rate of 0.9, then the false positive rates would be 0.36 and 0.56 for wheezes.
Figure 3-3: Best features from the denoising autoencoder by identification task and crackles, respectively. ROC curves for both classifiers are shown in Figure 3-4. We obtained ROC curves with AUCs of 0.86 and 0.74 for wheezes and crackles, respectively.

3.4 Discussion

3.4.1 Findings

We demonstrated two approaches to detect adventitious lung sounds. The first approach used traditional techniques from signal processing to engineer features that can be used to identify wheezes and crackles. The second approach used unsupervised learning techniques to learn a feature representation. This feature representation captures the structure inherent in the sound file spectrograms and can use it to identify wheezes and crackles.

Both approaches perform reasonably well. The first approach does a bet-
ter job identifying crackles, while the second approach does a better job identifying wheezes. These algorithms were tested on a larger corpus of lung sounds than prior studies. Additionally, the sounds collected in this study were collected by a research assistant with a low-cost stethoscope and Android phone; they were not collected by a pulmonologist with an expensive electronic stethoscope. These results more accurately represent the performance of a device used by primary care physicians or health workers in the field.

When we tested signal processing techniques, we tested aggregating lung sounds across the body or splitting them into two groups: upper and lower airways. For wheezes, the aggregate model outperformed the separate upper and lower airway models. For crackles, the upper and lower airway models outperformed the aggregate model. This indicates that wheezes are more similar between different recording locations than crackles. It also makes sense that the crackle identification model had trouble with identifying crackles in the lower airways. Hearing adventitious lung sounds in the lower airways is more challenging because accurate detection of breath sounds requires patient cooperation in taking deep breaths.

Although the two methods perform similarly, it is more difficult to use results from the second approach to answer the question of interest: Does the patient have a wheeze or crackle? When we attempted to use predictions from multiple four-second segments to identify wheezes and crackles in a thirty-second recording, the accuracy of the method decreased. We were not able to achieve accurate
predictions for the thirty-second sound files. As a result, we believe that, given our current dataset, the engineered feature approach will outperform unsupervised feature learning.

3.4.1.1 Practical Considerations for Future Researchers

While designing these methods, we encountered two implementation questions that we did not anticipate. By highlighting them here, we hope that future researchers will consider these questions before engaging in data collection.

First, the protocol for lung sound recording affects how likely it is to capture an adventitious lung sound. Adventitious lung sounds may appear on only a subset of breath cycles, so a longer recording is more likely to capture an adventitious lung sound from a patient. However, if a patient is auscultated for a long time or multiple successive auscultations are conducted, crackles can disappear as the lungs temporarily expand. Additionally, the strength of inhalation and exhalation can affect the likelihood of hearing an adventitious lung sound. Because of these factors the data collection protocol needs to be carefully designed to maximize the likelihood of capturing an adventitious lung sound during each recording.

Second, ground truth labeling for adventitious lung sounds is difficult to obtain. As has previously been demonstrated in the literature, interrater reliability for the detection of adventitious lung sounds is low. This presents challenges to machine learning research, as the success of many methods depends on high-quality labeled examples. Although the process is time-consuming, for future work we recommend asking multiple physicians to label each sound and using the amount of agreement between physicians to weight training examples in the model.

3.4.2 Limitations

There are a few caveats to these results. First, the adventitious lung sound labels were provided by a single pulmonologist. As previously discussed, there can be significant variation between physicians in their abilities to recognize adventitious
lung sounds. While the physician who provided these labels is a pulmonary disease
specialist with more than twenty years of experience, the labels could be improved
by increasing the number of physicians who review them.

Second, we did not explore using the feature learning approach on the full
length sound files. Because of computational constraints, we were limited to analyz-
ing smaller sound files. If the entire sound files had been processed as a single unit,
we would not have needed to aggregate multiple predictions. Without this step, it
is possible that feature learning would have outperformed feature engineering.

Third, for the feature engineering approach we used physician labels from
manual auscultation done before the sound file recordings. There can be some tem-
poral variation in the presentation of adventitious lung sounds, so sounds could
have disappeared or emerged between the two sessions. This is especially true of
crackles, where repeated auscultation can lead to a temporary opening of the air-
ways. In the future, we recommend simultaneous recording and manual auscul-
tation so that there is no temporal variation.

3.4.3 Future Work

We have identified a few avenues for future research. First, the results need to be
evaluated on a larger patient dataset. This will ensure that they generalize across
providers, patients and recording locations. Most of our sound files were recorded
by a single technician, so other users may have differences in technique that affect
the quality of the sound recording. Additionally, these files were mostly recorded
at a single location, so it will be important to ensure that the results generalize to
locations with different levels of background noise.

The signal processing approach could be improved further by creating
recording-area specific models. For now, the aggregate wheeze model outperforms
the upper and lower airway models, but this is only because the amount of inter-
area variation is smaller than the intra-area variation in recordings. As more data
is collected from each area, all important information will be available in recordings
from that area and so moving to area-specific models will perform at least as well
as an aggregate model.

We also think that additional work could be done to improve the unsupervised learning approach. Although the unsupervised learning approach did not perform as well as the engineered feature approach in this analysis, it will continue to improve. Two modifications will lead to improvements in the unsupervised approach. First, the inclusion of additional labeled and unlabeled data will allow the stacked denoising autoencoder to learn more robust features and better tie those features to the adventitious lung sounds. Second, moving from autoencoders to convolutional autoencoders will help improve the methods by making the recognition algorithms time invariant. Traditional autoencoders focus on finding structures in the same location in multiple images and using that information for classification. A convolutional autoencoder can convolve a structure across the image to find it in any location. In image recognition tasks, the typical approach is to use two-dimensional convolution to make the image recognition invariant in both the x and y axes. In sound processing, a one-dimensional convolution should be used to make the sound recognition task invariant in time.
Chapter 4

Using Lung Sound Analysis for Pulmonary Disease Diagnosis

In this chapter, I evaluate how lung sounds could be used for diagnosis. I describe some of the limitations associated with using adventitious lung sound identification and propose a new approach: using machine learning and signal processing methods that can be used to tie lung sound analysis directly to disease diagnosis. I present data on the frequency of lung sounds within patients with different diseases and show that the discriminatory power of the presence of adventitious lung sounds is limited. Next, I propose methods to tie analysis of lung sounds directly to diagnosis, without attempting to identify adventitious lung sounds. I evaluate these methods on data collected from patients in Pune, India. Finally, I close by discussing the implications of these findings.

4.1 Diagnostic Utility of Identifying Adventitious Lung Sounds

The diagnostic utility of identifying adventitious lung sounds is limited for two reasons. First, there is overlap between the lung sounds produced by different diseases. Second, there is large variation among physicians in the ability to accurately de-
tect adventitious lung sounds and disagreement about what to call them. In other
words, the inter-rater and intra-rater reliability for auscultation of abnormal breath
sounds is poor.

4.1.1 Adventitious Lung Sound Overlap Between Diseases

The biggest challenge associated with using adventitious lung sound identification
for diagnosis is the limited sensitivity it provides for identifying specific diseases.
When auscultating patients to find adventitious lung sounds, physicians hope to
learn information that will improve their diagnosis of a pulmonary disease. Detection
of adventitious lung sounds can be thought of as a diagnostic test, with the
presence of adventitious lung sounds giving a positive result. As with all diagnos-
tic tests, physicians can use Bayes Rule to incorporate new information into their
diagnosis (Equations 4.1 and 4.2).

\[
p(disease|sound) = \frac{p(sound|disease)p(disease)}{p(sound|disease)p(disease) + p(sound|disease)p(disease)}
\]

\[
p(disease|\neg sound) = \frac{p(sound|disease)p(disease)}{p(sound|disease)p(disease) + p(sound|disease)p(disease)}
\]

In Equations 4.1 and 4.2, there are two quantities that are sufficient and
typically used to describe the performance of a diagnostic test or symptom: the
sensitivity, \( p(sound|disease) \), and the specificity, \( p(\neg sound|disease) \). Given the sensi-
tivity and specificity, the other quantities in the equations can be calculated if the
baseline probabilities of \( p(sound) \) and \( p(disease) \) are known.

In order for adventitious lung sound identification to be useful for diagno-
sis, the sensitivity and specificity of the test should be high. For screening, either
the sensitivity or specificity of the test should be high, and the test can be com-
bined with another test with complementary performance. For example, a common
diagnostic strategy is to first screen using a test with high sensitivity (but correspondingly low specificity). This will ensure that few patients are missed, but will result in many false positives because the specificity is low. Combining this test with a second test of higher specificity will help eliminate the false positives in the second stage of testing. If neither sensitivity nor specificity is high, then the presence of adventitious lung sounds will not provide useful information about disease.

Sensitivity and specificity change depending on the diagnostic task. If the goal is only to identify unhealthy individuals, then the specificity of using adventitious lung sounds is high; the likelihood that a healthy individual will have no abnormal lung sounds is high. However, these quantities are lower for separating patients with asthma from patients with COPD; both diseases produce similar adventitious lung sounds so both the sensitivity and specificity of detecting an adventitious lung sound are low.

The problem with using adventitious lung sound identification for the diagnosis of pulmonary disease is illustrated in Figure 4-1. There are many different diseases that cause each of the adventitious lung sounds. As a result, only limited information can be gained by detecting the lung sounds.

4.1.2 Challenges in Accurately Identifying Adventitious Lung Sounds

The second challenge associated with using adventitious lung sound identification for diagnosis is the difficulty in assigning ground truth labels to potential adventitious lung sounds. In order to train doctors or algorithms to identify adventitious lung sounds, it is important to agree on standard definitions for adventitious lung sounds and conclusively label particular recordings. However, there is disagreement in the medical community over what the classes of lung sounds should be. In addition, even if physicians agree on the classes of lung sounds, the ability of physicians to agree on the same label for a particular sound is not guaranteed.

Sarkar et al. describe how lung sounds were originally defined as either
Sound caused by local obstruction

Sound caused by asthma

Sound caused by COPD

Sound caused by pulmonary fibrosis

Sound caused by bronchiectasis

Sound caused by pneumonia

Local obstruction

Asthma

COPD

Pulmonary fibrosis

Bronchiectasis

Pneumonia

Figure 4-1: Information lost by focusing on adventitious lung sound detection

continuous or interrupted [52]. In 1976, they were redefined as fine crackles, coarse crackles, wheezes, or rhonchi. In addition to these classifications, Bohadana et al. also describe stridor, pleural rub, and squawks [53]. In 2014, Melbye et al. argued that the classification scheme should be simplified, focusing on two classes of sounds: wheezes and crackles [54]. With significant disagreement among doctors about how to classify lung sounds, it is difficult to use lung sound identification for diagnosis.

Even when physicians agree on the classes of lung sounds, there is significant variation between physicians in their labeling of adventitious lung sounds. Elphick et al. found kappa scores of 0.07 and 0.36 for wheezes and crackles, respectively [62]. Schilling et al. computed an interrater reliability of 76% for the detection of abnormal lung sounds [63]. Melbye et al. found that kappa scores were below 0.4 for all adventitious sounds except inspiratory wheeze [54]. As a result, multiple physicians detect different lung sounds in the same recording or auscultation location. This means that multiple physicians will listen to the same sound file
and identify different adventitious lung sounds.

Finally, lung sounds may change with repeated auscultation. For example, crackles heard on initial auscultation may disappear after the patient takes several deep inspirations. Similarly, the ability to detect wheezes will depend on the depth of inspiration and resultant volume of expired air.

Because of these three factors, it is difficult to assign ground truth labels to particular sound files. Without agreed-upon labels, it is challenging to use the presence of these sounds for diagnosis.

4.2 Lung Sound Analysis as an Alternative to Adventitious Lung Sound Identification

By focusing on detecting specific classes of lung sounds, researchers may discard information that could be useful for the diagnosis of disease (Figure 4-1). This approach makes sense when sound analysis is done by physicians, who have varying ability to hear and identify lung sounds and cannot do fine-grained signal processing analysis. However, if sound files are being recorded and analyzed electronically, algorithms need not be limited to the identification of adventitious lung sounds. Instead, signal processing features from the sound files can be tied directly to different diseases.

This approach follows the same line of inquiry as past research by Sovijärvi et al., which found that the shape of the crackle waveform could be used to discriminate between different pulmonary diseases [64]. By moving beyond the identification of adventitious lung sounds, the team was able to capture information that was more pertinent for the diagnosis of different diseases.

In addition to improving the ability to discriminate between different pulmonary diseases, this approach sidesteps the issue of ground-truth labeling. Patient diagnoses are determined using Pulmonary Function Testing, and, unlike adventitious lung sound identification, are not affected by the subjective nature of a
Table 4.1: Summary of patients with each diagnosis

<table>
<thead>
<tr>
<th>Variable</th>
<th>Asthma</th>
<th>COPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>104</td>
<td>50</td>
</tr>
<tr>
<td>Sex (% Male)</td>
<td>0.44</td>
<td>0.86</td>
</tr>
<tr>
<td>Age (yrs) mean±sd</td>
<td>52±13</td>
<td>65±8</td>
</tr>
<tr>
<td>Weight (kg) mean±sd</td>
<td>61±13</td>
<td>53±12</td>
</tr>
<tr>
<td>Smoking (%)</td>
<td>0.18</td>
<td>0.76</td>
</tr>
</tbody>
</table>

physician’s ear. As a result, linking lung sound features directly to diagnosis may improve the accuracy and utility of these algorithms.

4.3 Data Collection and Methods

At the time of writing, data had been collected from 104 patients with asthma and fifty patients with COPD from the Chest Research Foundation in Pune, India. These patients were a subset of the overall study, which collected data from 237 consecutive pulmonary disease patients who visited CRF for medical care and 88 healthy subjects who were recruited from the families and friends of visiting patients. Diagnosis was made after each patient underwent Pulmonary Function Testing. Data collection is ongoing, and the targeted patient sample size is 500 patients. A summary of the asthma and COPD patients is shown in Table 4.1.

Two methods were used to collect lung sound data. First, the physician manually auscultated each patient at eleven recording locations (Figure 3-1) and noted the presence of any adventitious lung sounds at each auscultation location. Technicians then used the stethoscope and Android application described in Chapter 2 to record thirty seconds of lung sounds from the eleven recording locations on the trachea, chest, and back. The percent of patients with each adventitious lung sound for each area by disease are shown in Table 4.2.
Table 4.2: Percent of patients with lung sounds of each type by area and disease

<table>
<thead>
<tr>
<th>Area</th>
<th>Wheezes</th>
<th>Crackles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Asthma</td>
<td>COPD</td>
</tr>
<tr>
<td>1</td>
<td>0.11</td>
<td>0.16</td>
</tr>
<tr>
<td>2</td>
<td>0.09</td>
<td>0.16</td>
</tr>
<tr>
<td>3</td>
<td>0.12</td>
<td>0.18</td>
</tr>
<tr>
<td>4</td>
<td>0.10</td>
<td>0.11</td>
</tr>
<tr>
<td>5</td>
<td>0.11</td>
<td>0.13</td>
</tr>
<tr>
<td>6</td>
<td>0.10</td>
<td>0.13</td>
</tr>
<tr>
<td>7</td>
<td>0.12</td>
<td>0.13</td>
</tr>
<tr>
<td>8</td>
<td>0.12</td>
<td>0.18</td>
</tr>
<tr>
<td>9</td>
<td>0.12</td>
<td>0.16</td>
</tr>
<tr>
<td>10</td>
<td>0.11</td>
<td>0.18</td>
</tr>
<tr>
<td>11</td>
<td>0.12</td>
<td>0.16</td>
</tr>
</tbody>
</table>

4.4 Data Analysis

In order to compare the diagnostic value of adventitious lung sound identification to signal processing lung sound features, we developed a set of machine learning features from both the lung sound labels and from signal processing methods on the recorded sound files. Then, we used these features to train classification models and evaluated their performance on identifying patients with asthma and COPD.

4.4.1 Spatial Analysis of Adventitious Lung Sounds

To evaluate the diagnostic potential of adventitious lung sound identification, we used the physician-provided lung sound labels from manual auscultation. Because the distribution of adventitious sounds across the lungs can provide important information about the type and extent of pulmonary disease [65, 34, 52], we developed a number of features looking at the presence and distribution of adventitious lung sounds in the patient. These features are summarized in Table 4.3. Each of these features were calculated for three adventitious lung sound categories: wheeze; crackle; and abnormal, which was true for an area if a wheeze or a crackle was de-
Table 4.3: Description of features from adventitious lung sound identification

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Feature Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presence of adventitious sound in each area</td>
<td>Whether the physician identified an adventitious sound in each of the 11 auscultation locations.</td>
</tr>
<tr>
<td>Presence of adventitious sound in any area</td>
<td>Whether the physician heard an adventitious sound in any auscultation area.</td>
</tr>
<tr>
<td>Count of adventitious sounds</td>
<td>The number of auscultation locations where the physician heard an adventitious sound.</td>
</tr>
<tr>
<td>Presence of adventitious sound in the upper body</td>
<td>Whether the physician heard an adventitious sound in the auscultation areas located in the upper body.</td>
</tr>
<tr>
<td>Count of adventitious sounds in the upper body</td>
<td>The number of auscultation locations in the upper body where the physician heard an adventitious sound.</td>
</tr>
<tr>
<td>Presence of adventitious sound in the lower body</td>
<td>Whether the physician heard an adventitious sound in the auscultation areas located in the lower body.</td>
</tr>
<tr>
<td>Count of adventitious sounds in the lower body</td>
<td>The number of auscultation locations in the lower body where the physician heard an adventitious sound.</td>
</tr>
</tbody>
</table>

To aid in the analysis of adventitious lung sounds, we developed a new method for visualizing their spatial extent. An explanation of the method is shown in Figure 4-2. In the diagram, each square represents the presence or absence of an adventitious lung sound at each recording location. The numbers in each location correspond to the eleven recording locations we used in our lung sound recording (Figure 3-1). A map could be generated for each patient and compared to standard maps generated for each type and stage of different pulmonary diseases.

Example maps for patients in our dataset are shown in Figure 4-3 and Figure 4-4. Because we had only a small number of patients and did not do any grading based on disease severity, the maps generated from this study have limited utility for the diagnosis of pulmonary disease. However, the methodology of analyzing the spatial distribution of lung sounds could be used in future research.
Figure 4-2: Explanation of map of spatial distribution of adventitious lung sounds. Each square represents one of the eleven recording locations on the body. If an adventitious lung sound is present at that site, the square is marked as red. If no lung sound is present, the square remains white.

Figure 4-3: Maps of spatial distribution of wheezes for asthma and COPD. Coloring represents percentage of patients with a wheeze in each area.

Figure 4-4: Maps of spatial distribution of crackles for asthma and COPD. Coloring represents percentage of patients with a crackle in each area.
4.4.2 Signal Processing Lung Sound Analysis

These features were originally developed to build algorithms capable of detecting wheezes and crackles and then reused to directly predict diagnoses. A complete description of these features can be found in Section 3.3.1.1. Each feature was computed over the entire 30 second recording.

4.5 Classification Algorithms and Results

4.5.1 Classification Algorithms

We used 80% of the patients as the training dataset and 20% of the patients as a test dataset. This random selection was stratified by whether the patient had asthma or COPD, so that class sizes were balanced between the training and test sets.

For each patient, sound files were recorded from eleven auscultation locations. This resulted in eleven different sets of lung sound analysis features. Each area was considered as a separate set of machine learning features, with no information shared between recordings on the same patient. This resulted in twelve feature sets for each patient: one set of adventitious lung sound identification features and eleven sets of features for the eleven recording locations.

Then, for each set of machine learning features, we tested three different types of classification algorithms: logistic regression models with L1 regularization, support vector machines with the radial basis function kernel, and gradient boosting machines. Five-fold cross validation was used to determine the best value of model parameters for each classification scheme. In order to ensure that regularization did not improperly penalize features because they occur on different scales, each feature was normalized so that it had zero mean and unit standard deviation. Before training the algorithms, features with zero variance across the training dataset were excluded.

For each set of features and classification methodology, we created a
Table 4.4: Classification performance for detection of asthma and COPD using adventitious lung sound labels. Performance is reported as median (inter-quartile range).

<table>
<thead>
<tr>
<th>Classifier</th>
<th>AUC</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>0.56(0.50-0.59)</td>
<td>0.11(0.00-0.22)</td>
<td>1.00(1.00-1.00)</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>0.61(0.56-0.66)</td>
<td>0.33(0.22-0.44)</td>
<td>0.90(0.85-0.95)</td>
</tr>
<tr>
<td>Gradient Boosting</td>
<td>0.59(0.56-0.63)</td>
<td>0.22(0.11-0.33)</td>
<td>0.95(0.85-1.00)</td>
</tr>
</tbody>
</table>

receiver-operating-characteristic curve. Additionally, we computed the values for sensitivity and specificity that minimized the total probability of error.

We repeated the entire process of model training and evaluation 100 times with different randomly selected test sets. This repetition allowed analysis of the variability of the final results and their dependence on sampling error. This provides a better estimate of generalizability to previously unseen data. We reported the area under the curve (AUC) values at the 25th percentile, median, and 75th percentile.

In addition to evaluating the classification accuracy of the different feature sets, we explored the value of different features for diagnostic accuracy. We selected the best performing feature sets and examined which features contributed to the final classifications.

### 4.5.2 Classification Performance

Three types of models were trained using the features derived from adventitious lung sound identification. The results from these models for the classification of asthma and COPD are summarized in Table 4.4. All three models give roughly equivalent results, with Support Vector Machines slightly outperforming the other two approaches. All three models performed poorly, indicating that detecting the presence of wheezes and crackles is not sufficient for the diagnosis of asthma and COPD.

The same three types of models were trained on the signal processing fea-
Table 4.5: Classification performance for detection of asthma and COPD using signal processing lung sound features. Performance is reported as median AUC (inter-quartile range).

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Area 1</th>
<th>Area 2</th>
<th>Area 3</th>
<th>Area 4</th>
<th>Area 5</th>
<th>Area 6</th>
<th>Area 7</th>
<th>Area 8</th>
<th>Area 9</th>
<th>Area 10</th>
<th>Area 11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>0.69(0.63-0.75)</td>
<td>0.67(0.61-0.72)</td>
<td>0.63(0.58-0.70)</td>
<td>0.62(0.58-0.67)</td>
<td>0.65(0.59-0.69)</td>
<td>0.64(0.59-0.69)</td>
<td>0.73(0.66-0.79)</td>
<td>0.66(0.61-0.70)</td>
<td>0.66(0.61-0.72)</td>
<td>0.67(0.63-0.70)</td>
<td>0.68(0.65-0.75)</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>0.71(0.67-0.74)</td>
<td>0.68(0.62-0.73)</td>
<td>0.67(0.63-0.74)</td>
<td>0.63(0.58-0.72)</td>
<td>0.59(0.55-0.66)</td>
<td>0.72(0.68-0.76)</td>
<td>0.72(0.68-0.78)</td>
<td>0.67(0.62-0.72)</td>
<td>0.67(0.62-0.72)</td>
<td>0.68(0.62-0.72)</td>
<td>0.72(0.68-0.77)</td>
</tr>
<tr>
<td>Gradient Boosting</td>
<td>0.71(0.64-0.76)</td>
<td>0.69(0.64-0.74)</td>
<td>0.72(0.67-0.76)</td>
<td>0.63(0.58-0.67)</td>
<td>0.66(0.61-0.71)</td>
<td>0.66(0.61-0.71)</td>
<td>0.74(0.67-0.79)</td>
<td>0.70(0.65-0.76)</td>
<td>0.72(0.66-0.77)</td>
<td>0.72(0.67-0.76)</td>
<td>0.74(0.69-0.79)</td>
</tr>
</tbody>
</table>

Features created from the sound files in each auscultation location. The results from these models for the classification of asthma and COPD are summarized in Table 4.5. In general, Gradient Boosting Machines provided the best results, although the large inter-quartile ranges make conclusive determinations of the best model impossible. Additionally, there was some variation between recording locations, with Areas 7 and 11 providing the best results for distinguishing between these diseases.

4.5.3 Feature Evaluation

The most important features from the gradient boosting lung sound identification model are shown in Figure 4-5. Although the adventitious lung sound identification model did not perform well overall, it did learn some patterns that confirmed our findings about the difference between auscultation sites. Two of the top five features were associated with abnormal lung sounds in Area 11. Many of the rest of the features captured the variation between adventitious lung sounds in the upper and lower airways. This confirms that the spatial distribution of lung sounds can
Figure 4-5: Feature importance for detection of asthma and COPD using adventitious lung sound identification features. The lines of the box indicate the 25, 50, and 75 percentile values each feature weight across the bootstrapped iterations. Feature importance is computed using Gini importance [66].

provide information about asthma and COPD.

The most important features from the gradient boosting signal processing model trained on Area 11 are shown in Figure 4-6. These features show that some features from the FFT of the entire sound file (maximum power and variance) are useful for differentiating asthma and COPD. Some of the other top features were those comparing the horizontal patches and those focusing just on the subset of the spectrogram identified as useful for finding wheezes (250-1000 Hz). One feature proposed by Aydore for the detection of wheezes, f50 over f90 ratio [39], can be used to directly classify asthma and COPD.
Figure 4-6: Feature importance for detection of asthma and COPD using lung sound analysis features. The lines of the box indicate the 25, 50, and 75 percentile values each feature weight across the bootstrapped iterations. Feature importance is computed using Gini importance [66]. WS prefix means feature was derived from a spectrogram limited to 250-1000 Hz.
4.6 Discussion

4.6.1 Findings

The classification algorithms proposed in this section showed that signal processing features extracted from lung sounds recorded at a single auscultation site show promise for the diagnosis of asthma and COPD. These features provide diagnostic guidance that is superior to an algorithm that detects wheezes and crackles. This means that auscultating at a few sites and analyzing the sounds using signal processing could outperform auscultating at many locations and identifying wheezes and crackles.

This result makes sense when the percentages of patients with adventitious lung sounds from each disease are considered. There is significant overlap between asthma and COPD, meaning that the presence of wheezes and crackles alone will not have discriminatory power. Instead, more fine-grained analysis is required.

All of the lung sounds in this analysis were recorded using the low-cost electronic stethoscope described in Chapter 2. These sounds were recorded by a technician, not a physician. Auscultation could become a task for technicians, freeing up physician labor for more complex tasks.

4.6.2 Limitations

There are a few caveats to these findings. First, the lung sound labels were provided by a single pulmonologist. If there were errors in the sound file labels, the accuracy of diagnosis from the labels could be reduced.

Second, these results were tested on only two diseases. As a result, they do not show that there is no value in the identification of lung sounds. Sovijärvi et al. showed that by locating crackles within the inspiratory cycle, they were able to detect different pulmonary diseases [64]. However, we find that simply detecting wheezes and crackles was not effective for the diagnosis of asthma and COPD.
4.6.3 Future Work

These results indicate a promising future direction for research. The features developed here are only an initial exploration of how lung sound analysis could be used for pulmonary disease diagnosis. The models could likely be improved by including other features that have been used for lung sound classification or general sound analysis tasks.

Additionally, future work is required to verify that these results replicate with other patient populations and other diseases. The patient population studied had only asthma or COPD, so it is important to ensure that these features are useful for discriminating between other pulmonary diseases. Given the richness of the feature set, it is likely that this method will continue to outperform lung sound identification.
Chapter 5

Diagnostic Guidance for Asthma and COPD

In this chapter, I describe a diagnostic algorithm for asthma and COPD. This algorithm was created as a first step towards a general-purpose system for the diagnosis of pulmonary disease. I begin by describing the data collected from patients at a pulmonary clinic in India. Then, I propose a set of machine learning features that were derived from measurements made using the diagnostic kit. I use these features to train machine learning models to identify four types of patients: healthy, non-obstructive, asthma, and COPD. I evaluate each component of the kit separately and discuss the implications of my findings for deploying a mobile diagnostic kit.

5.1 Data Collection and Methods

At the time of writing, data had been collected from 325 subjects at the Chest Research Foundation in Pune, India. The patient population consists of two subgroups: 237 consecutive pulmonary disease patients who visited CRF for medical care and 88 healthy subjects who were recruited from the families and friends of visiting patients. Data collection is ongoing, and the targeted patient sample size is 500 patients.

For each subject, a technician recorded lung sounds from eleven standard-
ized recording locations on the thorax. An experienced pulmonologist auscultated at the same eleven locations with a traditional stethoscope and indicated the presence of wheezes or crackles. This auscultation served as our reference standard.

Each subject recorded three peak flow meter measurements. The technician administered the questionnaire and recorded the responses.

To determine subject diagnoses, each subject underwent a clinical exam, followed by complete examination in the Pulmonary Function Testing lab (spirometry, body plethysmography, pulmonary gas diffusion, and impulse oscillometry). If required by the normal course of medical care, an x-ray was obtained and the results were used to assist in diagnosis. In order to identify allergic rhinitis, the pulmonologist followed ARIA guidelines [20]. Asthma and COPD were diagnosed according the GOLD report and ERS/ATS guidelines [48, 2].

Thirty-seven subjects in the dataset were excluded from analysis. Ten subjects had no age recorded. Thirty subjects had no peak flow meter value recorded. One subject had no weight recorded. A summary of the patients included in the study with each diagnosis is shown in Table 5.1.

One hundred and fifty-four patients were diagnosed with asthma or COPD. If a patient was diagnosed with asthma or COPD, they were considered to be an asthma and COPD patients regardless of any additional diagnoses. No patients were diagnosed with both asthma or COPD.

Seventy-three patients had a diagnosis other than asthma or COPD. Twenty-nine of these patients had allergic rhinitis. Seventeen of these patients had interstitial lung disease. Seven patients had post-TB lung damage and the remaining patients had pulmonary fibrosis, cardiac problems or respiratory infections.

5.2 Data Analysis

In order to develop algorithms capable of detecting pulmonary disease, we needed to extract a number of features from the patient data. These features were extracted from data collected with the three kit components. Then, these features
Table 5.1: Summary of patients with each diagnosis

<table>
<thead>
<tr>
<th>Variable</th>
<th>Healthy</th>
<th>Other</th>
<th>Asthma</th>
<th>COPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>98</td>
<td>73</td>
<td>104</td>
<td>50</td>
</tr>
<tr>
<td>Sex (% Male)</td>
<td>0.56</td>
<td>0.45</td>
<td>0.44</td>
<td>0.86</td>
</tr>
<tr>
<td>Age (yrs) mean±sd</td>
<td>35±12</td>
<td>52±16</td>
<td>52±13</td>
<td>65±8</td>
</tr>
<tr>
<td>Weight (kg) mean±sd</td>
<td>62±12</td>
<td>61±12</td>
<td>61±13</td>
<td>53±12</td>
</tr>
<tr>
<td>Smoking (%)</td>
<td>0.07</td>
<td>0.16</td>
<td>0.18</td>
<td>0.76</td>
</tr>
</tbody>
</table>

were used to train machine learning models to recognize the different diseases in our dataset.

5.2.1 Features for Classification

The most straightforward features were those developed from the responses to the questionnaire. While the questionnaire asked for details about symptoms and risk factors for pulmonary disease, this instantiation of the algorithm only used answers to the questions about whether symptoms and risk factors were present. The features derived from these responses are described in Table 5.2.

The next set of features used was derived from the peak flow meter readings. Clinical recommendations state that the peak flow meter should be used three times and the maximum reading should be used as the measured value. In addition, the measured value should be compared to the population predicted value for a given patient’s gender, age, and height. Both the maximum value and percent of population predicted value were used as features. The equations used to compute the population predicted values are shown in Equation 5.1 and Equation 5.2.

\[
\text{Predicted Value Male} = -1.807 \times \text{Age (years)} + 3.206 \times \text{Height (cm)} \quad (5.1)
\]

\[
\text{Predicted Value Female} = -1.454 \times \text{Age (years)} + 2.368 \times \text{Height (cm)} \quad (5.2)
\]
### Table 5.2: Description of questionnaire features

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Feature Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>Patient’s sex</td>
</tr>
<tr>
<td>Age</td>
<td>Patient’s age</td>
</tr>
<tr>
<td>Breathlessness</td>
<td>Does the patient experience breathlessness?</td>
</tr>
<tr>
<td>Cough</td>
<td>Does the patient experience cough?</td>
</tr>
<tr>
<td>Chest Pain</td>
<td>Does the patient experience chest pain?</td>
</tr>
<tr>
<td>Fever</td>
<td>Does the patient have a fever?</td>
</tr>
<tr>
<td>Nasal Symptoms</td>
<td>Does the patient have any nasal symptoms?</td>
</tr>
<tr>
<td>Family History of COPD</td>
<td>Does the patient have a family history of COPD?</td>
</tr>
<tr>
<td>Family History of Allergies</td>
<td>Does the patient have a family history of allergies?</td>
</tr>
<tr>
<td>Personal History of Allergies</td>
<td>Does the patient have a personal history of allergies?</td>
</tr>
<tr>
<td>Exposure to Biomass Cooking</td>
<td>Is the patient currently exposed or has the patient been previously exposed to biomass cooking?</td>
</tr>
<tr>
<td>Smoking History</td>
<td>Does the patient currently smoke or has the patient previously been a smoker?</td>
</tr>
<tr>
<td>Tobacco Chewing History</td>
<td>Does the patient currently chew tobacco or has the patient previously chewed tobacco?</td>
</tr>
<tr>
<td>Alcohol Consumption History</td>
<td>Does the patient consume alcohol or has the patient previously consumed alcohol?</td>
</tr>
</tbody>
</table>
To incorporate the lung sound data, two different approaches were used. First, to test the performance of a theoretically perfect lung sound classifier, the pulmonologist's auscultation labels were used directly for classification. These features were the same as the features described in Section 4.4.1. Second, lung sound analysis features were extracted directly from the lung sound recordings. These features were extracted from the right interscapular (area 11) recording location. The features extracted were described previously in Section 3.3.1.1. This recording area was chosen because it provided the best results on the diagnostic task described in Section 4.5.2.

5.2.2 Classification Algorithms

The algorithm to identify patients with asthma and COPD was designed as three stages of machine learning classification. The first stage detected whether the patient was healthy or unhealthy. If they were unhealthy, the second stage detected whether the patient had an obstructive airway disease. If the patient did have an obstructive airway disease, the third stage determined whether that disease was asthma or COPD. This multi-stage approach was selected because it allows decision makers to more easily specify the acceptable level of trade-offs between false positives and false negatives at each stage of classification. This would allow the user to require a higher certainty before declaring patients healthy versus deciding whether a patient had asthma instead of COPD. The three stages are shown in Figure 5-1.

We used 80% of the data as the training dataset and 20% of the data as a randomly selected held-out test dataset. This random selection was stratified by the outcome variable, so that class sizes were properly balanced between the training and test sets.

All three stages of the algorithm were designed as logistic regression models with L1 regularization. Logistic regression models were used because the interpretation of these models is intuitive and they can be easily explained to physicians. L1 regularization was used because this technique identifies only those features most important for accurate classification. This will minimize the amount of data
that needs to be collected to achieve an accurate diagnosis. Five-fold cross validation was used to determine the best value for the regularization parameter. In order to ensure that the regularization parameter did not improperly penalize features because they occur on different scales, each feature was normalized so that it had zero mean and unit standard deviation.

The first stage of the classifier was trained on the entire training set. The second stage of the model was trained only on unhealthy patients. The third stage of the model was trained only on patients with asthma or COPD. The performance of the three-stage approach was evaluated using the test set. For each stage, a receiver-operating-characteristic curve was created. Additionally, the values for sensitivity and specificity that minimize the total probability of error were computed. After evaluating each stage separately, the entire system was tested to determine the classification accuracy of patients with asthma and COPD. Each single-stage classifier used the minimum probability of error threshold. A confusion matrix was computed to determine classification accuracy by patient type.
In addition to testing the performance of an algorithm trained on data from all three components of the kit, models were trained with each combination of one or two components of the diagnostic kit. This allowed an evaluation of each component of the kit, so that the simplest required diagnostic kit and algorithm could be deployed.

The entire process of model training and evaluation was repeated 100 times with different randomly selected test sets. This repetition allowed analysis of the variability of the final results and their dependence on sampling error. This provides a better estimate of generalizability to previously unseen data. For the evaluation of single stage classifiers, AUC values at the 25\(^{th}\) percentile, median, and 75\(^{th}\) percentile were reported. For the combined system confusion matrix, the 25\(^{th}\) percentile, median, and 75\(^{th}\) percentile for each cell were computed.

5.3 Results

5.3.1 Detecting Unhealthy Patients

The first stage of the classification algorithm was trained to separate healthy from unhealthy patients. The results for different component combinations are shown in Table 5.3. The questionnaire results alone achieve almost perfect classification across the bootstrap iterations. The peak flow meter also does well, but both the sound analysis and sound identification are not able to adequately classify healthy and unhealthy patients. Because the questionnaire is almost perfect, adding other components does not greatly improve performance. When combined, the sound identification and sound analysis perform better than either separately. Adding either of them to the peak flow meter or questionnaire does not lead to meaningful increases in performance.
Table 5.3: Classifier performance for detecting unhealthy patients by kit component. Components are Sound Analysis (SA), Sound Identification (SI), Questionnaire (QU), and Peak Flow Meter (PF). Performance is reported as median (interquartile range).

<table>
<thead>
<tr>
<th>Components Used</th>
<th>AUC</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA</td>
<td>0.75(0.72-0.79)</td>
<td>0.76(0.66-0.85)</td>
<td>0.69(0.56-0.79)</td>
</tr>
<tr>
<td>SI</td>
<td>0.66(0.63-0.68)</td>
<td>0.31(0.27-0.36)</td>
<td>1.00(1.00-1.00)</td>
</tr>
<tr>
<td>PF</td>
<td>0.87(0.84-0.90)</td>
<td>0.74(0.65-0.80)</td>
<td>0.89(0.78-0.94)</td>
</tr>
<tr>
<td>QU</td>
<td>1.00(0.99-1.00)</td>
<td>0.98(0.95-1.00)</td>
<td>1.00(0.95-1.00)</td>
</tr>
<tr>
<td>SI, SA</td>
<td>0.83(0.78-0.86)</td>
<td>0.73(0.63-0.80)</td>
<td>0.82(0.69-0.88)</td>
</tr>
<tr>
<td>PF, SA</td>
<td>0.86(0.83-0.89)</td>
<td>0.69(0.64-0.78)</td>
<td>0.88(0.81-0.94)</td>
</tr>
<tr>
<td>PF, SI</td>
<td>0.90(0.87-0.92)</td>
<td>0.74(0.67-0.82)</td>
<td>0.94(0.89-1.00)</td>
</tr>
<tr>
<td>QU, SA</td>
<td>0.99(0.98-1.00)</td>
<td>0.96(0.93-1.00)</td>
<td>1.00(0.94-1.00)</td>
</tr>
<tr>
<td>QU, SI</td>
<td>1.00(0.99-1.00)</td>
<td>0.95(0.95-0.98)</td>
<td>1.00(1.00-1.00)</td>
</tr>
<tr>
<td>QU, PF</td>
<td>0.99(0.99-1.00)</td>
<td>0.97(0.95-0.97)</td>
<td>1.00(0.94-1.00)</td>
</tr>
<tr>
<td>PF, SI, SA</td>
<td>0.88(0.86-0.91)</td>
<td>0.74(0.66-0.80)</td>
<td>0.94(0.81-0.94)</td>
</tr>
<tr>
<td>QU, SI, SA</td>
<td>1.00(0.99-1.00)</td>
<td>0.97(0.95-1.00)</td>
<td>1.00(1.00-1.00)</td>
</tr>
<tr>
<td>QU, PF, SA</td>
<td>1.00(0.99-1.00)</td>
<td>0.97(0.94-0.97)</td>
<td>1.00(1.00-1.00)</td>
</tr>
<tr>
<td>QU, PF, SI</td>
<td>1.00(0.99-1.00)</td>
<td>0.97(0.95-0.97)</td>
<td>1.00(1.00-1.00)</td>
</tr>
<tr>
<td>QU, PF, SI, SA</td>
<td>1.00(0.99-1.00)</td>
<td>0.97(0.94-0.97)</td>
<td>1.00(1.00-1.00)</td>
</tr>
</tbody>
</table>
Table 5.4: Classifier performance for detecting obstructive diseases by kit component. Components are Sound Analysis (SA), Sound Identification (SI), Questionnaire (QU), and Peak Flow Meter (PF). Performance is reported as median (interquartile range).

<table>
<thead>
<tr>
<th>Components Used</th>
<th>AUC</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA</td>
<td>0.65 (0.60-0.70)</td>
<td>0.64 (0.54-0.82)</td>
<td>0.62 (0.46-0.77)</td>
</tr>
<tr>
<td>SI</td>
<td>0.69 (0.65-0.72)</td>
<td>0.94 (0.26-0.97)</td>
<td>0.36 (0.29-1.00)</td>
</tr>
<tr>
<td>PF</td>
<td>0.87 (0.83-0.88)</td>
<td>0.70 (0.63-0.81)</td>
<td>0.85 (0.77-0.92)</td>
</tr>
<tr>
<td>QU</td>
<td>0.79 (0.74-0.83)</td>
<td>0.84 (0.68-0.90)</td>
<td>0.69 (0.54-0.79)</td>
</tr>
<tr>
<td>SI, SA</td>
<td>0.72 (0.68-0.77)</td>
<td>0.66 (0.51-0.84)</td>
<td>0.75 (0.50-0.83)</td>
</tr>
<tr>
<td>PF, SA</td>
<td>0.83 (0.79-0.88)</td>
<td>0.72 (0.60-0.80)</td>
<td>0.82 (0.73-0.91)</td>
</tr>
<tr>
<td>PF, SI</td>
<td>0.89 (0.86-0.92)</td>
<td>0.78 (0.70-0.85)</td>
<td>0.83 (0.83-0.92)</td>
</tr>
<tr>
<td>QU, SA</td>
<td>0.74 (0.70-0.79)</td>
<td>0.77 (0.64-0.86)</td>
<td>0.67 (0.58-0.77)</td>
</tr>
<tr>
<td>QU, SI</td>
<td>0.86 (0.80-0.90)</td>
<td>0.87 (0.74-0.94)</td>
<td>0.69 (0.62-0.85)</td>
</tr>
<tr>
<td>QU, PF</td>
<td>0.88 (0.86-0.91)</td>
<td>0.78 (0.70-0.85)</td>
<td>0.85 (0.77-0.92)</td>
</tr>
<tr>
<td>PF, SI, SA</td>
<td>0.85 (0.80-0.89)</td>
<td>0.76 (0.68-0.84)</td>
<td>0.80 (0.70-0.90)</td>
</tr>
<tr>
<td>QU, SI, SA</td>
<td>0.81 (0.75-0.85)</td>
<td>0.82 (0.71-0.89)</td>
<td>0.73 (0.55-0.82)</td>
</tr>
<tr>
<td>QU, PF, SA</td>
<td>0.86 (0.81-0.91)</td>
<td>0.76 (0.68-0.84)</td>
<td>0.82 (0.73-0.91)</td>
</tr>
<tr>
<td>QU, PF, SI</td>
<td>0.92 (0.89-0.95)</td>
<td>0.81 (0.78-0.89)</td>
<td>0.92 (0.83-0.92)</td>
</tr>
<tr>
<td>QU, PF, SI, SA</td>
<td>0.88 (0.83-0.92)</td>
<td>0.84 (0.72-0.92)</td>
<td>0.80 (0.70-0.90)</td>
</tr>
</tbody>
</table>

5.3.2 Detecting Obstructive Diseases

The second stage of the classification algorithm was trained to separate patients with obstructive airway diseases from patients with non-obstructive airway diseases. The results for different component combinations are shown in Table 5.4. At this level of classification, the peak flow meter is the best performing individual component, followed by the questionnaire. Once again, both of the auscultation-related components do not achieve good performance. The best performing combination is the questionnaire, peak flow meter, and sound identification.

5.3.3 Detecting Asthma and COPD

The final stage of the classification algorithm was trained to separate patients with asthma from patients with COPD. The results for different component combina-
Table 5.5: Classifier performance for differentiating asthma and COPD by kit component. Components are Sound Analysis (SA), Sound Identification (SI), Questionnaire (QU), and Peak Flow Meter (PF). Performance is reported as median (interquartile range)

<table>
<thead>
<tr>
<th>Components Used</th>
<th>AUC</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA</td>
<td>0.69(0.62-0.74)</td>
<td>0.67(0.44-0.78)</td>
<td>0.74(0.57-0.89)</td>
</tr>
<tr>
<td>SI</td>
<td>0.55(0.53-0.58)</td>
<td>0.20(0.10-0.30)</td>
<td>0.95(0.86-1.00)</td>
</tr>
<tr>
<td>PF</td>
<td>0.85(0.80-0.89)</td>
<td>0.88(0.75-0.88)</td>
<td>0.79(0.68-0.89)</td>
</tr>
<tr>
<td>QU</td>
<td>0.91(0.88-0.93)</td>
<td>0.80(0.70-0.90)</td>
<td>0.86(0.76-0.90)</td>
</tr>
<tr>
<td>SI, SA</td>
<td>0.66(0.61-0.73)</td>
<td>0.61(0.33-0.78)</td>
<td>0.75(0.60-0.90)</td>
</tr>
<tr>
<td>PF, SA</td>
<td>0.80(0.73-0.85)</td>
<td>0.71(0.71-0.86)</td>
<td>0.72(0.61-0.83)</td>
</tr>
<tr>
<td>PF, SI</td>
<td>0.84(0.78-0.89)</td>
<td>0.75(0.62-0.88)</td>
<td>0.84(0.74-0.89)</td>
</tr>
<tr>
<td>QU, SA</td>
<td>0.93(0.89-0.96)</td>
<td>0.89(0.78-1.00)</td>
<td>0.79(0.72-0.89)</td>
</tr>
<tr>
<td>QU, SI</td>
<td>0.90(0.86-0.94)</td>
<td>0.80(0.78-0.90)</td>
<td>0.86(0.76-0.90)</td>
</tr>
<tr>
<td>QU, PF</td>
<td>0.90(0.85-0.94)</td>
<td>0.88(0.72-0.91)</td>
<td>0.84(0.74-0.95)</td>
</tr>
<tr>
<td>PF, SI, SA</td>
<td>0.79(0.71-0.86)</td>
<td>0.71(0.57-0.86)</td>
<td>0.78(0.65-0.89)</td>
</tr>
<tr>
<td>QU, SI, SA</td>
<td>0.92(0.89-0.95)</td>
<td>0.89(0.78-1.00)</td>
<td>0.84(0.74-0.89)</td>
</tr>
<tr>
<td>QU, PF, SA</td>
<td>0.92(0.88-0.95)</td>
<td>1.00(0.86-1.00)</td>
<td>0.83(0.72-0.89)</td>
</tr>
<tr>
<td>QU, PF, SI</td>
<td>0.90(0.85-0.92)</td>
<td>0.88(0.75-0.91)</td>
<td>0.84(0.74-0.91)</td>
</tr>
<tr>
<td>QU, PF, SI, SA</td>
<td>0.90(0.84-0.94)</td>
<td>0.86(0.86-1.00)</td>
<td>0.83(0.67-0.94)</td>
</tr>
</tbody>
</table>

tions are shown in Table 5.5. Both the questionnaire and peak flow meter are useful for discriminating between patients with asthma and COPD. The best combination is the questionnaire combined with sound analysis.

5.3.4 Combined Results

Based on the results from the previous section, it is clear that the questionnaire and peak flow meter provide the greatest diagnostic value for each stage of classification. While both the sound analysis and sound identification do lead to small increases in median AUC for the detecting obstructive diseases and differentiating asthma and COPD, this performance increase is within the range of sampling noise.

To verify that this result held for the entire classification system, we tested the questionnaire and peak flow meter in all three stages of classification. In this
Table 5.6: Classifier performance for diagnosis patients using the questionnaire and peak flow meter. Values reported are the median and interquartile range for the confusion matrix across bootstrap iterations.

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Healthy</td>
</tr>
<tr>
<td>Healthy</td>
<td>18(17-18)</td>
</tr>
<tr>
<td>Nonobstructive</td>
<td>1(0-2)</td>
</tr>
<tr>
<td>Asthma</td>
<td>0(0-1)</td>
</tr>
<tr>
<td>COPD</td>
<td>0(0-0)</td>
</tr>
</tbody>
</table>

assessment, the classifiers worked in sequence, classifying patients and removing them from further evaluation. For example, an individual marked as healthy was not considered by the obstructive disease classifier or the asthma-COPD classifier. The results are shown in Table 5.6. For each stage of classification, we chose the threshold that produced the minimum probability of error as the class cutoffs.

The combined algorithms accurately classify almost all of the patients correctly. No asthma or COPD patients are classified as healthy. Some asthmatics are classified as non-obstructive or COPD, so work could be done to improve the accuracy of detecting asthma patients.

5.3.5 Feature Importance

The feature coefficients for the three stages of classification are shown in Figures 5-2, 5-3, and 5-4. For the first stage of the model (Figure 5-2), positive coefficients mean that feature is associated with an increased likelihood of being unhealthy. We can see that coughing, breathlessness, and nasal symptoms are all associated with unhealthy patients. The maximum peak flow meter reading over reference is negatively associated with being unhealthy, so lower values indicate unhealthy patients.

For the second stage of the model (Figure 5-3), positive coefficients are associated with having an obstructive disease. The most important feature is max-
Figure 5-2: Feature coefficients for detection of unhealthy patients using the questionnaire and peak flow meter. The lines of the box indicate the 25, 50, and 75 percentile values each feature weight across the bootstrapped iterations. The feature coefficients are beta values from the logistic regression model. Positive coefficients indicate that the feature is associated with increased likelihood of being unhealthy.

imum peak flow meter reading over reference, with low values indicating high likelihood of an obstructive disease. Breathlessness and a history of allergies are also associated with having an obstructive disease.

For the third stage of the model (Figure 5-4), positive coefficients are associated with having COPD. The most important features are smoking and age, both of which are associated with having COPD.
Figure 5-3: Feature coefficients for detection of patient with obstructive disease using the questionnaire and peak flow meter. The lines of the box indicate the 25, 50, and 75 percentile values each feature weight across the bootstrapped iterations. The feature coefficients are beta values from the logistic regression model. Positive coefficients indicate that the feature is associated with increased likelihood of having an obstructive airway disease.
Figure 5-4: Feature coefficients for differentiating patients with asthma and COPD using the questionnaire and peak flow meter. The lines of the box indicate the 25, 50, and 75 percentile values each feature weight across the bootstrapped iterations. The feature coefficients are beta values from the logistic regression model. Positive coefficients indicate that the feature is associated with increased likelihood of having COPD.
5.4 Discussion

5.4.1 Findings

The classification algorithms proposed in this chapter were able to accurately identify patients with asthma and COPD from a subject population that also contained patients with allergic rhinitis, post-TB pulmonary impairment, interstitial lung disease, ischemic heart disease, and healthy individuals without lung disease. Although the original diagnostic kit consisted of three components (electronic stethoscope, peak flow meter, questionnaire), almost all of the diagnostic value was provided by the peak flow meter and questionnaire. This finding agrees with past research, which found that a few questions and the peak flow meter can be used as a screening tool for these diseases [67]. While the analysis of recorded lung sounds and identification of adventitious lung sounds did have some diagnostic value on its own, this value did not lead to large improvements in diagnostic accuracy when combined with the questionnaire and peak flow meter. All of the important diagnostic information was captured by these two tools.

The feature importance results show that the model has learned relationships that agree with medical knowledge. As shown by Kodgule et al. [67], the peak flow meter can be used to identify obstructive airway diseases. The model also agrees with the GOLD report [48], which finds that age and smoking are risk factors for COPD.

These results indicate that the kit could be deployed with only the questionnaire and peak flow meter. The electronic stethoscope is the most expensive and difficult-to-use part of the kit, so removing it from the final implementation would simplify production and increase the ease of use. As long as the patient population is limited to the diseases captured in this study, this research indicates that the kit will work well to identify patients with asthma and COPD.
5.4.2 Limitations

This finding is limited to the diseases studied in this patient population. None of the patients presented with pneumonia, tuberculosis, or lung cancers. If these diseases had been captured in this study, the presence of abnormal lung sounds would likely have had more diagnostic value. For example, pleural rubs can indicate inflammation of the pleura, likely caused by pneumonia. Bronchial breathing in the lower lungs can indicate a cavity caused by tuberculosis. All of these diseases can result in lung sound asymmetries between the left and right lungs. If patients with these conditions were included in the study, the diagnostic value of both the spatial distribution and presence of abnormal lung sounds may have justified the inclusion of the electronic stethoscope results in the diagnostic kit.

These findings are limited by the classification algorithms used. In order to maintain the transparency of the final diagnostic algorithms, we used only logistic regression. However, in Chapter 4, we showed that gradient boosting machines were the best type of model to use with lung sound analysis features to detect asthma and COPD. As a result, these features might have been more useful had a more complicated classification algorithm been used.

5.4.3 Future Work

Future work could improve three aspects of this result. First, data could be collected from additional subjects. This would improve the assessment of the model by reducing the uncertainty associated with the AUC estimates. If data is collected from patients with additional pulmonary diseases, it will allow researchers to ensure that the model will generalize to populations that include these additional types of patients.

Second, additional features could be derived from questions on the questionnaire. Data has already been collected on the onset, periodicity, and severity of the symptoms assessed. This information was not included in the final model because any improvement in the results on the current patient dataset would not
have been distinguishable. As the patient sample size increases and disease coverage expands, these features should be assessed for their classification performance.

Third, the thresholds for each of the models could be updated to reflect the prevalence of diseases and expected costs of misclassification. The current thresholds were chosen to produce the minimum probability of error, but these may not be the best thresholds for deployment. The best model threshold should be chosen by determining costs for each type of misclassification and combining that information with disease prevalence to determine the threshold that minimizes the Bayesian cost equation (Equation 5.3, where $C_{01}$ is the cost of choosing class 0 when class 1 is correct, $p(0|1)$ is the probability of classifying class 1 as class 0, and $p(1)$ is the prevalence of class 1). The choice of threshold will depend on whether the tool is being used for screening or diagnosis. If used for screening, the cost of misclassification would be lower because follow-up testing will be used to confirm diagnoses.

The proper threshold should be evaluated for each use case individually, as the availability of follow-up testing will vary and disease prevalences will change based on the target population. Any deployment should include a short pilot program to determine the optimal threshold and adjust the models for any site and population-specific idiosyncrasies.

$$E_{\text{Cost}} = C_{01}p(0|1)p(1) + C_{10}p(1|0)p(0)$$

(5.3)
Chapter 6

Android Implementation of a Pulmonary Screening Tool

In Chapter 5, I described an algorithm to screen for asthma and COPD. In order for this algorithm to have an impact on the state of disease diagnosis, it needed to be packaged into a deployable form. In this chapter, I describe the implementation of that algorithm in an Android application that could be used by primary care physicians and health workers to screen for asthma and COPD.

6.1 Pulmonary Screening Application

The Pulmonary Screening application is designed to receive a machine learning model as input and automatically build a screening application around it. An overview of the application is shown in Figure 6-1. The application is a data-gathering tool for the underlying machine learning model, so it is important to explain how models are stored before going through the application workflow.

6.1.1 Machine Learning Models in the Application

For the purposes of this Android application, a model consists of a set of features and information on how to use feature values to make a prediction. Because the
initial implementation of the models is limited to logistic regression, the only information required to make predictions is the model intercept and a coefficient for each feature.

To describe each feature, the model contains information about how to query the user for feature values. Features are defined as one of four different categories: categorical, continuous, derived, and external.

Categorical and continuous features can both be obtained by asking the user questions within the application. Within the model, each feature is associated with text that is used to query the user. Categorical features also contain a list of categories that the user can choose from.

Derived features are feature values that can be obtained through a manipulation of other features. An example of this is the “Peak Flow Meter % Predicted”, which is obtained by dividing the maximum peak flow meter reading by the population predicted value. Each derived feature has a list of dependencies that defines all of the features that must be obtained before it can be calculated. In the current implementation of the application, the calculation of each derived feature is hardcoded into the application.

External features are feature values that are obtained from external applications. An example of this is the maximum peak flow meter value obtained from the Augmented Reality Peak Flow Meter application. These features launch the external application, passing any information necessary for the application to function. Once the external application is finished, it returns a set of feature values.
In the current implementation of the application, each external feature is handled separately using custom activities.

Models are stored as XML files. When a model file is needed by the application, it is handled by the ModelParser class. The ModelParser class converts a model file into two outputs: an ArrayList of FeatureRequests and a Model object. The FeatureRequest ArrayList contains a list of features required by the model to make predictions. These are contained by subclasses of FeatureRequest corresponding to the four categories previously described. The features are ordered in descending order by the absolute value of their coefficient size, so that more important features are listed first.

The Model object holds all of the feature coefficients and the model intercept. It has two primary methods that are used by the application: predictClassProbabilities, predictClassProbabilitiesLowerBounds. Each of these methods accepts a HashMap of feature names and feature values.

The first method, predictClassProbabilities, makes a prediction of the probabilities for each class given the provided feature values. It assumes that any unknown features do not affect the final class probabilities.

The second method, predictClassProbabilitiesLowerBounds, makes a prediction of the lower bound probabilities for each class. It evaluates the lowest possible likelihood for each class, assuming the most pessimistic values for unknown features. For continuous values, this is three standard deviations below the mean. For categorical features, this is whichever categorical outcome has the biggest impact on reducing each class probability. These are pessimistic predictions for each class, so the values will not sum to one.

In this version of the Android application, there are three models. Each model corresponds to one of the classification tasks described in Figure 5-1. These models sit behind the scenes and the user does not interact directly with them.
6.1.2 Application Workflow

When the application is launched, the user has the option of screening a new patient or reviewing the results of a previous patient (Figure 6-2a). If the user wishes to review the results from a previous patient, they are asked to enter a patient ID number. Once they do, the ResultsActivity is launched and the results from a previous screening are displayed.

If they choose to screen a new patient, the RegistrationActivity is launched. In the RegistrationActivity, the user is asked for a patient ID number and some basic demographic information (Figure 6-2b).

Once that information is entered, the application loads the three classification models into memory. For each model, the application stores an ArrayList of FeatureRequests, ordered by their importance for the classification task. The first classification task is to determine whether the patient is unhealthy. The application looks at the ArrayList of FeatureRequests for the classification task and uses the first entry to query the user for information. This process is repeated for each feature in the ArrayList. Example continuous and categorical features are shown in Figure 6-2c and Figure 6-2d. An example of an external feature is shown in Figure 6-2e.

After each feature value is obtained, the application queries the Model class using to determine the pessimistic probability for each class using the predictClassProbabilitiesLowerBounds method. If the pessimistic likelihood is more than 50% for either class, the model assumes that the final prediction will not change. If the patient has been classified as healthy, the application starts the ResultsActivity.

If the patient is classified as unhealthy, the application begins to ask questions from the next model, obstructive vs. non-obstructive. The entire process is repeated, moving on to the final model if the patient is classified as obstructive or moving onto the ResultsActivity if the patient is classified as non-obstructive. Once the final model determines whether the patient has asthma or COPD, the ResultsActivity is launched.
Pulmonary Screening

(a) Application Launch

Patient ID

(b) Patient Registration

Breathlessness

(c) Categorical Question

Does the patient experience breathlessness?

No

(d) Continuous Question

Weight

Patient Age (years)

(e) External Feature

What is the patient's weight in kg?

75

(f) Patient Results

Max PFM Reading

What was the patient's maximum peak flow meter reading in three trials (L/min)?

Patient 1 likely has COPD.

The model predicts that the patient has a 72.4% chance of having COPD. The following features contributed to this determination:
- The patient's PFM reading was 60 L/min
- The patient is or was a smoker
- The patient experiences breathlessness

Figure 6-2: Screenshots from the pulmonary screening application
By using the pessimistic model probability as a cut-off for class probability, the screening application is able to reduce the number of questions required before making a determination without sacrificing prediction accuracy. For patients on the class margins, questions will be asked until the model is certain. For clear instances of a particular class, the model will only need to ask a few questions before moving on to the next model or the final result.

The ResultsActivity shows the likelihood of each stage of classification (Figure 6-2f). In addition, it shows the three features that contributed most to the determination in each stage of classification. The user is able to flip between summary pages for each level of classification.

6.2 Integration with Other Applications

The screening application was designed to connect with other Android applications. There are two types of connections. First, the screening application can ask other applications to return feature values. In the current version of the application, this is used to get peak flow meter readings and features from lung sound analysis. Second, the screening application can be run by an external application. This allows pulmonary screening to be included in a larger application offering that addresses different public health needs.

6.2.1 Querying External Applications

6.2.1.1 Peak Flow Meter Application

The peak flow meter application is used to capture readings from the peak flow meter. The application itself is described in more detail in Section 2.2.

A diagram of the interaction between the screening application and the peak flow meter application is shown in Figure 6-3. The screening application sends an intent to the MeasurePeakFlowActivity within the peak flow meter app. The intent includes the patient’s sex, height, and age. The intent also indicates whether
the peak flow meter app should show the instructions page. When the MeasurePeakFlowActivity in the peak flow meter app is finished, it returns the peak flow meter reading.

### 6.2.1.2 Lung Sound Recorder

The lung sound recorder application is used to record lung sounds using an electronic stethoscope. The application itself is described in more detail in Section 2.1. A diagram of the interaction between the screening application and the lung sound recorder application is shown in Figure 6-4. The screening application sends an intent to the BodyActivity in the lung sound recorder application. The intent includes an array of booleans indicating which areas need to be recorded. When the lung sound recorder is finished, it returns the recorded sound data. Features are extracted from the sound data within the screening application.

### 6.2.2 Screening from an External Application

The screening application can be run from an external application by sending an intent to the ScreeningActivity (Figure 6-5). The intent needs to include a patient ID number and the patient’s age and sex. When it finishes, the activity will return a set of feature values collected by the application and the predictions for each of the three outcomes.
6.3 Discussion

We have created an Android application that walks a user through the screening process for asthma and COPD. The application was designed to be flexible, so that changes in the underlying machine learning models would not require modification of the Android code.

This flexibility is apparent in any of the model features that are obtained by asking the user a question. The feature is defined in an XML file and the Android application automatically builds an activity to ask the user the associated question.

This same flexibility was not included for externally obtained and derived features. In this version of the application, both of these types of features must be hardcoded into ScreeningActivity. For external features, this is because external applications will have different input requirements. To prevent passing all patient data to external applications, we chose to specifically pass only the information required by each external application. As a result, the call for each external application was coded separately.

For derived features, it would be possible to increase the application’s flexibility by allowing XML definition of derived features. This would allow the application to read in a formula specifying how a derived feature can be obtained from other feature values. At present, there is only a single derived feature in the application (Max Peak Flow % Predicted), so it was hard-coded into the application.

Because of the application design, it can be easily modified to work for any machine learning predictions that rely heavily on questionnaires. This means it can be extended to other pulmonary diseases or other classes of disease. After the
model XML files are changed, the application will automatically generate questions and make predictions for the new outcomes of interest.

In the future, the application could be improved by enabling more flexible access with other applications. Application access could become a general-purpose activity, with the application to call, variables to pass, and information to request specified in the model XML file. This would allow for more flexible reuse in other applications.
Chapter 7

Conclusions

In this thesis, I described a pulmonary diagnostic kit and associated algorithms that could be used to screen for asthma and COPD. The kit consisted of three components: an electronic stethoscope, a peak flow meter, and a symptom and risk factor questionnaire. The kit used an Android device to collect data and make predictions using techniques from machine learning. This project made contributions in machine learning, mobile application development and clinical research.

7.1 Contributions of this Work

7.1.1 Machine Learning Contributions

In the course of developing the screening algorithms, we explored methods for automatically detecting adventitious lung sounds. This was a natural first step for analysis, because significant prior work by multiple researchers has sought to create methods for detecting adventitious lung sounds. In most of the past studies, sample sizes were small and sound files were collected using expensive electronic stethoscopes. This study is unique because it is one of the largest to date and the sounds were recorded using a stethoscope that costs less than $30 USD. Additionally, unlike many past studies, we evaluated performance using a test set, ensuring that my results will generalize new data. For crackle detection, I presented algorithms
capable of accurately detecting crackles in both the upper and lower airways. The methods were less successful at detecting wheezes.

In addition to developing a number of signal processing features for the detection of abnormal lung sounds, we used new methods from deep learning to learn a feature representation from the data. These features were used to detect wheezes and crackles in four-second segments of the full recordings. These methods showed promise, but limitations on physician time prevented a full analysis of their potential. The physician labor required to listen to and label these sound files resulted in a much smaller number of labeled examples than is typical when using deep learning methods.

In addition to developing methods to identify lung sounds, we also demonstrated that lung sound identification may not be the best approach if the goal is to accurately diagnose disease. As an alternative, I processed recorded sound files using methods from signal processing to create a set of machine learning features. Then, we used these features to directly predict whether a patient had asthma or COPD. This method worked better than an approach that relied on a pulmonologist's identification of wheezes and crackles at eleven different recording locations. This approach is especially promising as the range of diseases expands, because there are significant overlaps in the sounds produced by different diseases. By analyzing the sound files directly, richer data can be extracted, with greater diagnostic value.

This finding has value beyond the diagnosis of pulmonary disease. When machine learning researchers approach a new problem, they often use the previous experience of domain experts to guide their inquiry. However, the decision-making of clinicians may be influenced by technical limitations that do not apply as new methods are developed. In this case, the human ear is not capable of discerning fine-grained information about the frequency content of lung sounds. Instead, physicians have created classes of adventitious lung sounds that can provide clues about different diseases. As new analysis techniques are applied, researchers can skip identifying adventitious lung sounds and instead tie aspects of lung sounds directly to
disease diagnosis.

7.1.2 Clinical Contributions

The kit and algorithm combination was able to accurately detect patients with asthma and COPD from a patient population that contained healthy subjects and patients with a number of other pulmonary conditions. The final instantiation of the algorithm was implemented in an Android application. The application links to the two other Android applications described in this thesis (Lung Sound Recorder and Augmented Reality Peak Flow Meter) to gather data and identify pulmonary disease.

This project addresses a critical need in the developing world, where pulmonary diseases are underdiagnosed because of both a lack of trained physicians and a lack of diagnostic equipment. Currently, less than 20% of patients with COPD worldwide have been diagnosed and more than 80% of childhood asthmatics in India are undiagnosed [8, 9]. This kit can be provided to health workers, who can use it to screen for asthma and COPD in underserved populations.

While evaluating the kit, we showed that the majority of diagnostic value for detecting asthma and COPD can be obtained using only the peak flow meter and questionnaire. These results agree with past work, which found that a combination of a questionnaire and a peak flow meter is useful for separating patients with asthma from patients with COPD [67]. However, the results published here have broader implications. We found that the combination of a questionnaire and a peak flow meter can not only separate patients with asthma from patients with COPD, but that this combination is able to accurately detect patients with these diseases in a patient pool with a wide range of diseases.

While we found that both lung sound identification and lung sound analysis provided small improvements for detecting subsets of the diseases studied, the improvement was not large enough to be distinguishable from sampling error. Given these results, a kit that consisted of only the questionnaire and peak flow meter could be used to accurately screen for these diseases. As the patient sample
size increases and more diseases are included, it will be important to continue to assess the relative contribution of each kit component.

For now, this finding simplifies deployment, because no manufacturing is required to produce either the peak flow meter or electronic questionnaire. Additionally, health workers will not need to be trained in the proper use of the electronic stethoscope.

As more data is collected and these algorithms improve, it will be important to compare the diagnostic accuracy of traditional PFT methods with mobile tools. There may come a point when diagnostic tools that are less expensive and do not require as much training are a viable alternative to the current diagnostic approach.

### 7.1.3 Mobile Application Contributions

In order to implement the screening and diagnostic algorithm on the mobile phone, we developed an Android application that can automatically build a questionnaire from a machine learning model. The application is flexible, so that changes in the underlying machine learning models would not require modification of the Android code.

Because of the application design, it can be easily modified to work for any machine learning predictions that rely heavily on questionnaires. This means it can be extended to other diseases or domains besides healthcare.

The application is able to interface with other Android applications in two ways. First, the application can ask other applications for data, so that external sensors can be managed through individual applications. Second, the application can be used within another Android application, allowing this pulmonary diagnostic and screening tool to be embedded in preexisting mobile health platforms.
7.2 Future Work

To further increase the value of the kit, we are currently using it to collect data from patients with more varied pulmonary conditions. This will allow us to determine if the application can be used as a general-purpose pulmonary screening tool.

At the same time, a large-scale field study is being planned to collect data on the classification performance on a larger patient population. This study will evaluate two aspects of the kit. First, it will test how the diagnostic accuracy of the algorithms changes when data is collected by less skilled practitioners and health workers. Second, it will assess the value of the kit for general-purpose screening, instead of the performance on patients that present to a pulmonary clinic.

Because the kit is based around an Android phone, it can be readily deployed on devices that are already being purchased by individuals in the target populations. As a result, this kit is both inexpensive and scalable. If the results of future trials are consistent with the findings in this thesis, the kit and algorithm combination may provide useful information for improving diagnosis of pulmonary disease.
Bibliography


109


Todo list