A Human-Machine Framework for the Classification of Phonocardiograms

by

William Callaghan

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2018

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This thesis includes first-authored material that has been submitted to conference proceedings published by the Association for Computing Machinery (ACM). The conference paper from which I have adapted content is the following:

William Callaghan, Joslin Goh, Michael Mohareb, Andrew Lim and Edith Law. 2018. MechanicalHeart: A Human-Machine Framework for the Classification of Phonocardiograms. Submitted for publication in *Proceedings of the ACM on Human-Computer Interaction*, CSCW, (2018)

Abstract

In this thesis, we present and evaluate a framework for combining machine learning algorithms, crowd workers, and experts in the classification of heart sound recordings. The development of a hybrid human-machine framework for heart sound recordings is motivated by the past success in utilizing human computation to solve problems in medicine as well as the use of human-machine frameworks in other domains. We describe the methods that decide when and how to escalate the analysis of heart sound recordings to different resources and incorporate their decision into a final classification. We present and discuss the results of the framework which was tested with a number of different machine classifiers and a group of crowd workers from Amazon's Mechanical Turk. We also provide an evaluation of how crowd workers perform in various different heart sound analysis tasks, and how they compare with machine classifiers. In addition, we investigate how machine and human analysis are effected by different types of heart sounds and provide a strategy for involving experts when these methods are uncertain. We conclude that the use of a hybrid framework is a viable method for heart sound classification.

Acknowledgements

The past few years have been an experience that I will never forget and it would have not been the same without the following people. First, I would like to thank my family, to which this thesis is dedicated to, for their continued love and support. I would also like to thank my supervisor, Edith Law, for taking me in as a graduate student and for teaching me about what it takes to do research and to be a graduate student, which often involved learning to persist through difficult times.

In addition, I would like to thank the following faculty members for which I had the pleasure of interacting with in varying capacities throughout my graduate studies: Daniel Vogel, Edward Lank, Gladimir Baranoski, Jim Wallace, Michael Terry and Robin Cohen. A special mention goes to Joslin Goh for educating me on good experimental design and analysis.

A special thank you to Dr. Michael Mohareb from the William Osler Health System for providing the expert annotation of murmurs in our subset of heart sounds. As well, thank you to the people involved in curating the open-source dataset of heart sounds and for running the Computing in Cardiology Challenge to advance the field of heart sound analysis.

Of course, my time as a graduate student would not have been the same without my HCI lab members and friends. Throughout the past few years we have celebrated good times, and been there for each other during the difficult times, and I thank you all for being here on this journey with me.

Dedication

To my Mom (Julie), Dad (Keith), Stepdad (Graham) and my two sisters Kristine and Nicole.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Cardiovascular disease continues to be the leading cause of death worldwide [63]. In 2015, an estimated 17.7 million people died from cardiovascular disease, representing 31% of all global deaths [63]. In clinical practice, the physical examination of a patient is one of the first steps in evaluating their cardiovascular system [50]. Auscultation, the act of listening to sounds originating from the internal organs, is an important part of this process and may reveal pathological cardiac conditions such as arrhythmia, heart failure, and more [50, 82]. It is often the first step in disease evaluation, serving as a guide for further examination, and thus plays an important role in the early detection of cardiovascular disease [50].

Automated analysis of heart sounds, including heart sound classification, has been widely studied since the original work by Gerbarg *et al* (1963) [23]. Although many of these methods have demonstrated the ability to detect abnormalities, they are often done so on unrealistic, clean data [50]. The process of automated heart sound analysis is also challenging, as the frequency of the fundamental heart sounds, murmurs and respiration overlap significantly, making separation of normal and abnormal heart sounds difficult in both the frequency and time domains [50]. These factors motivated the creation of a large open access database and a challenge to develop robust heart sound classification algorithms [50, 14].

Crowdsourcing is an approach that enlists the help of humans to solve challenging problems that are difficult for automated approaches to complete with accuracy and precision [43, 70]. In medicine, crowdsourcing analysis of medical data is in its infancy, however there are a number of studies that have already shown its promise [69, 56, 51, 20, 89, 60].

The use of the crowd, or some human input, has also been leveraged to support machine learning (ie. active learning), where a learning algorithm can request labels for unlabelled instances from an oracle, and incorporate its feedback into the learning process [74]. Active learning has been applied to problems in areas such as biosignal classification [90, 45], speech recognition [93, 29, 84, 28], image classification [33] and text classification [92, 83]. Both the use of crowdsourcing in a medical context and the diverse set of problems being solved by combining humans and machines, motivates the exploration of these techniques in the heart sound classification space.

1.1 Thesis Objectives

In this thesis, we introduce and evaluate a framework for combining machine learning algorithms, crowd workers, and experts in the classification of heart sound recordings. We use a query strategy inspired by active learning to determine how to escalate the analysis of heart sound recordings to different resources and incorporate their decision into a final classification. Specifically, we aim to answer the following questions:

- What are non-expert, paid, crowd workers performance on various heart sound analysis tasks? Can these crowd workers accurately classify heart sounds as normal or abnormal? Can they identify the presence/absence of regions that are indicative of abnormalities?
- How do the two different heart sound analysis tasks compare in performance? Can the identification task serve as a proxy for heart sound classification? Can the information from both tasks be combined to achieve better performance in heart sound classification than the two alone?
- How does crowd-based heart sound classification compare to machine classification? Are there different types of heart sounds that are easier or more difficult to analyze?
- How do we combine machine classifiers and the crowd in a framework to better classify heart sounds? How do we determine when to involve an expert?

1.2 Contributions

Overall, this thesis focuses on a hybrid human-machine framework for heart sound classification. We explore methods for escalating heart sound analysis from a machine classifier to the crowd and then to an expert if needed. We then investigate ways of incorporating their analyses into a final classification. In addition, we evaluate how different crowd-based heart sound analysis tasks can be used to classify heart sounds and how machine and human analysis are effected by different types of heart sounds. As a result, we contribute:

- A framework for binary heart sound classification that utilizes input from machines, crowd workers, and experts (if required). This framework comes to a final classification of a given heart sound based on who has analyzed the heart sound, their level of uncertainty and a threshold of acceptable uncertainty. Our hybrid framework achieves greater performance than a baseline classifier alone, and utilizes less expert resources while achieving similar performance, when compared to a framework that does not use the crowd.
- A characterization of how different crowd-based heart sound analysis tasks can be used to classify heart sounds, and how the crowd performs in each of them. This includes a comparison of a binary classification task (normal or abnormal) and a murmur detection task, which is used as a proxy for classification but also provides the evidence behind such decision. We also illustrate how these two views of heart sound normality can be used in conjunction to come to a final classification.
- A number of extensions to the audio annotator framework initially developed by Cartwright *et al* (2017)[11]. Extensions include common audio analysis functionality like zoom, pan/scroll and volume controls, in addition to more advanced features like the ability to define rules and contextual information to guide work, and a separate viewer that allows users to compare multiple audio clips to one another.

In addition, this thesis answers each of the research questions mentioned above, discusses the applicability and integration of such a framework into real-world scenarios, and proposes how such a framework can be extended to support other types of bioacoustic signal analysis.

1.3 Outline

The remainder of the thesis is organized as follows:

Chapter Two describes the background theory in the domain of heart sound analysis. This includes an overview of how heart health is evaluated in the clinic by listening to heart sounds, and the criteria that is used in differentiating between normal and abnormal heart sounds. An overview of the work in automated heart sound classification, human computation in medicine and audio analysis, and hybrid human-machine frameworks is also presented.

Chapter Three introduces and describes the various components of the hybrid human-machine framework for heart sound classification. This includes the machine classifiers used, the interface developed for human annotation, the policy designed for querying humans and the rule(s) for incorporating their analyses into the final classification result. A detailed description of the data and methods used in evaluating the framework is also presented.

Chapter Four examines the results from the experiments that were ran to evaluate the hybrid human-machine framework as a whole, as well as its individual components. Applicability and integration into real-world scenarios is also discussed.

Chapter Five concludes this thesis by summarizing the work and outlining potential ways of extending the research in the future.

Chapter 2

Background

2.1 Heart Anatomy and Physiology

The following section summaries the heart's anatomy and physiology as described in *Auscultation Skills: Breath & Heart Sounds* [17, Chapter 1] unless otherwise noted.

The primary functions of the heart are to pump deoxygenated blood to the lungs, and the returning oxygenated blood throughout the body. The heart is divided into four chambers (left/right atrium and ventricle), which contain unidirectional valves to control blood flow.

To maintain adequate blood circulation, the heart must generate enough pressure to pump the blood throughout the body. This is accomplished by rhythmically contracting and relaxing the heart muscle, through electrical activation by cardiac cells. This pushes blood through the chambers of the heart and around the body, "as a result of the opening and closing of heart valves"[50].

2.2 Heart Auscultation and Basic Heart Sounds

The mechanical action of the heart, including the pumping of blood between the chambers of the heart, and the opening and closing of heart values to facilitate this process, gives rise

to vibrations which are audible on the chest wall [17, 50]. An audio or graphical recording of these vibrations (Figure 2.1) is referred to as a heart sound recording or phonocardiogram (PCG).

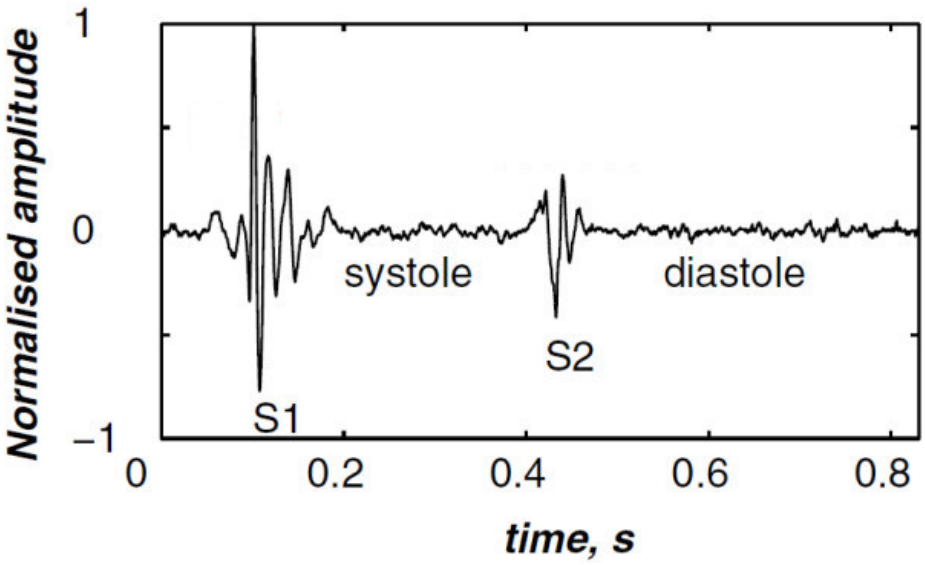


Figure 2.1: A phonocardiogram (PCG) of a normal heart sound [19].

Listening for specific heart sounds can give an indication of the heart's health [50]. In clinical practice, the physical examination of a patient is one of the first steps in evaluating their cardiovascular system [50]. Auscultation, the act of listening to sounds originating from the internal organs, is an important part of this process and may reveal pathological cardiac conditions such as arrhythmia, heart failure, and more [50, 82]. It is often the first step in disease evaluation, serving as a guide for further examination, and thus plays an important role in the early detection of cardiovascular disease [50].

A normal functioning heart produces two basic heart sounds: S1 and S2, and are essentially the "lub" and "dub" that most people think of when they hear a heart beat [17]. Immediately following S1 and lasting until S2 is Systole, and from S2 until the following S1 is known as Diastole. These four stages make up the cardiac cycle. Other sounds may be present such as the third (S3), and fourth (S4) heart sounds, clicks, snaps, or heart murmurs [17, 50].

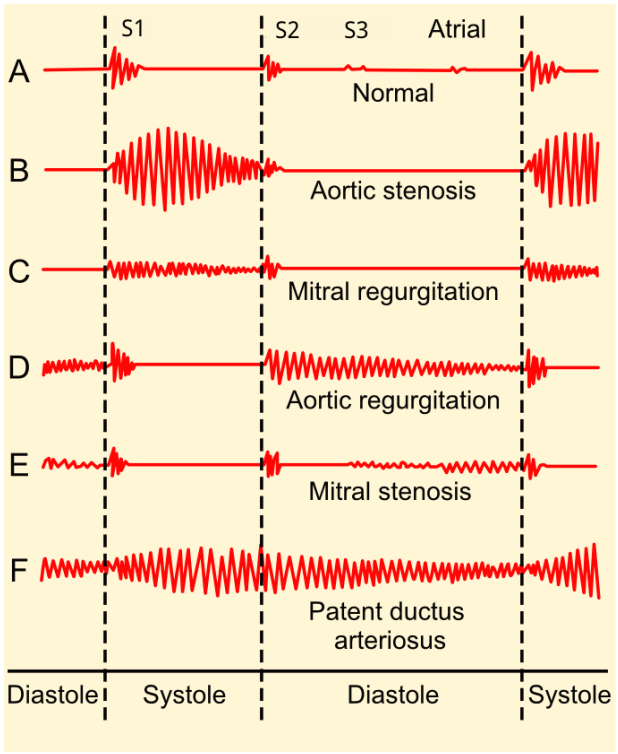


Figure 2.2: Illustration of normal and abnormal heart sounds. Adapted from Madhero (2010)[52].

2.3 Abnormal Heart Sounds

A heart murmur refers to an abnormal heart sound with "an underlying physiologic pathology,"[17] often caused by turbulent blood flow due to abnormal valves. Murmurs are characterized by:

Volume

The volume or intensity of a heart sound is often graded on a scale from 1-6, where 1 indicates the murmur is barely audible and 6 indicates the murmur can be heard without the use of a stethoscope [82]. This intensity is related to the amount of blood flow propagating through the valves but can also be effected by a patient's body type [17].

Pitch

The pitch of a heart sound is also related to the amount of blood flow, which creates vibrations of a given frequency [17]. Lower and higher pitched sounds are produced in result of slower and faster blood flow, respectively [82].

Configuration

The configuration or pattern of a murmur refers to the shape of the given sound on a PCG. [17]. Such patterns include crescendo (increasing intensity), decrescendo (fading), crescendo-decrescendo (and vice-versa), or uniform/plateau (equal intensity throughout the murmur) [17, 82]. In Figure 2.2 we can see an example of some of these configurations. The aortic stenosis condition shows a crescendo-decrescendo murmur whereas a uniform/plateau murmur is shown in mitral regurgitation.

Quality

Describing the quality of a murmur is open to interpretation, however, common words for describing the sound include blowing, harsh, or rumbling.

Timing and Duration

Murmurs can be classified based on where they occur in the cardiac cycle [17]. A murmur that occurs between S1 and S2 is known as systolic [82]. A diastolic murmur can occur between the end of S2 and the beginning of the next S1. Continuous murmurs are heard throughout Systole and into Diastole. Furthermore, murmurs can be described as where they occur within a given stage [17]. For example, a systolic murmur can be early, mid, late, or holosystolic (throughout Systole) [17]. Examples of such murmurs are present in Figure 2.2. We can see that aortic stenosis and mitral regurgitation are systolic murmurs whereas aortic regurgitation and mitral stenosis are diastolic murmurs.

Location

Auscultation is often performed at five pre-defined areas on the chest [82]. A murmur's location is defined as the region on the chest where the murmur is heard best, and is often the place of maximum intensity [17]. It is also usually correlated with "the underlying location of the valve that's responsible for producing the murmur," [17]. For example, instances of mitral regurgitation are best heard at the heart's apex or mitral area [82].

2.4 Automated Heart Sound Classification

Automated heart sound classification has shown to be challenging as the frequency of the fundamental heart sounds, murmurs and respiration overlap significantly, making separation of normal and abnormal heart sounds difficult in both the frequency and time domains [50]. Gerbarg *et al* (1963)[23] were the first to publish on automated methods in heart sound classification (specifically the classification of mitral regurgitation) by means of a decision-making process based upon the ratio of signal power and energy in different stages of the cardiac cycle.

Since then, much work has been done in the area of automated heart sound classification (see Table 2.1). However, Liu *et al* (2016)[50] argues that many of these investigations are unrealistic because of their use of high-quality recordings with pronounced features, not often seen in real-world recordings. As a result, they created a large database of heart sound recordings obtained from both real-world clinical and non-clinical environments, containing both clean and very noisy recordings. The PhysioNet/Computing in Cardiology (CinC) 2016 Challenge was then created to develop algorithms robust to these environments, that could classify heart sounds as normal or abnormal [14].

Liu *et al* (2016)[50] provides an extensive overview of prominent previous work in the field of heart sound classification. We include an adapted version of their summary in Table 2.1 with the addition of more recent work. From this review and the description of the top 20 entries in the CinC challenge [15], the most popular features are extracted via wavelet, frequency, time, time-frequency, and mel-frequency cepstral coefficient (MFCC) based methods. We provide a description of these common features in Appendix B. In addition, the most widely-used heart sound classification models have shown to be Neural Networks, Support Vector Machines (SVM), Hidden Markov Models (HMM), and

Classification Method	Author	Features
Neural Network	Akay *et al* (1994) [3]	Wavelet
	Liang and Hartimo (1998) [49]	Wavelet
	Uguz (2012) [86]	Wavelet
	Bhatikar *et al* (2005) [7]	Frequency
	Sepehri *et al* (2008) [73]	Frequency
	De Vos and Blanckenberg (2007) [18]	Time-Frequency
	Potes *et al* (2016) [67]†	Time-Frequency
	Kay and Agarwal (2016) [35]	Time-Frequency
	Grzegorczyk *et al* (2016) [27]†	Time, Frequency
	Nilanon *et al* (2016) [61]†	Spectral, MFCC
	Rubin *et al* (2016) [71]†	MFCC
	Maknickas and Maknickas (2017) [54]	MFSC
Support Vector Machine	Ari *et al* (2010) [4]	Wavelet
	Zheng *et al* (2015) [96]	Wavelet
	Patidar *et al* (2015) [66]	Wavelet
	Maglogiannis *et al* (2009) [53]	Frequency
	Gharehbaghi *et al* (2015) [24]	Frequency
	Goda and Hajas (2016) [25]†	Time, Frequency, Wavelet
	Ortiz *et al* (2016) [64]†	Time, MFCC, DTW
Hidden Markov Model	Wang *et al* (2007) [88]	MFCC
	Chauhan *et al* (2008) [12]	MFCC
	Saracoglu (2012) [72]	Frequency
k Nearest Neighbors	Bentley *et al* (1998) [6]	Wavelet
	Quiceno-Manrique *et al* (2010) [68]	Time-Frequency
	Avendano-Valencia *et al* (2010) [5]	Time-Frequency
	Bobillo (2016) [9]	Time-Frequency
Ensemble	Homsi *et al* (2016) [30]†	Frequency, Statistical, Wavelet
	Vernekar *et al* (2016) [87]†	Time, Frequency, MFCC
	Zabihi *et al* (2016) [94]†	Time, Frequency, Time-Frequency
Rule-Based	Langley and Murray (2016) [41]†	Wavelet
Random-Forest	Singh-Miller and Singh-Miller (2016) [78]†	Spectral

† Indicates a CinC challenge published paper.

Table 2.1: Summary of previous heart sound classification methods. Adapted from Liu *et al* (2016)[50] and Clifford *et al* (2017) [15]. Also contains papers published from the CinC challenge.

Clustering-based methods [50, 15]. The machine classifiers we use in our framework are given a more detailed description in Section 3.3.

2.5 Crowdsourcing and Human Computation

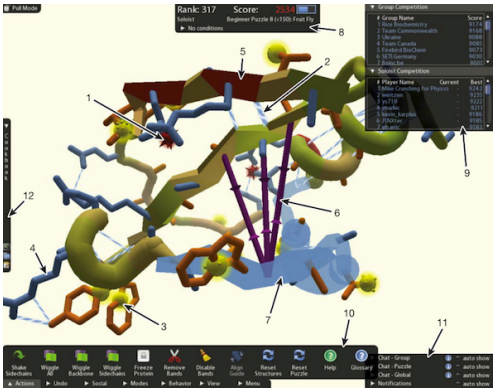
Crowdsourcing is an approach that enlists the help of humans to solve challenging problems that are currently unsolved or difficult for automated approaches to complete with accuracy and precision [43, 70]. On crowdsourcing platforms (e.g., Amazon Mechanical Turk [1]), people, henceforth referred to as crowd workers, often perform short microtasks such as image labelling and classification, audio transcription, or surverys, in exchange for small amounts of compensation [21, 70].

2.5.1 Crowdsourcing Medical Data Analysis

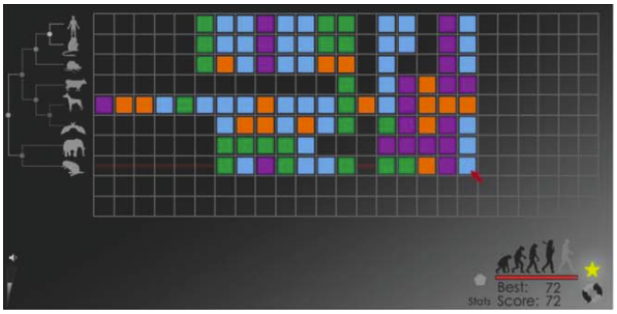
In the field of biology and medicine, there have been a number of successful crowdsourcing experiments to date, often times, including members that have no background in the field [16]. In bioinformatics, Foldit (Figure 2.3) is a game that allows players to manipulate protein structures with the goal of finding it's native conformation, or lowest energy [16]. Players receive a score for their solution based on achieving the lowest energy conformation. A leader board is even present to entice players to continue to optimize their solution(s). For traditional computation approaches to protein prediction problems, the search space is very large, with small protein structures having "on the order of 1,000 degrees of freedom," [16]. With the use of human spatial reasoning to explore the search space, FoldIt has outperformed state-of-the-art prediction systems, and aided in the discovery of important protein structures that have been unsolved for decades [16, 36, 26].

Phylo is another game, whose goal is to aid in solving the problem of large multiple sequence alignment (MSA) [34, 26]. The goal of MSA is to align nucleotides (the building blocks of DNA) from common ancestors to aid in the study of evolution and gene function [34]. In Phylo (Figure 2.3), DNA sequences are represented as rows of color-coded blocks which players can slide horizontally (pushing their neighbors if necessary) [34]. The goal is to find a configuration that "maximizes conservation across columns while minimizing

[1] https://www.mturk.com



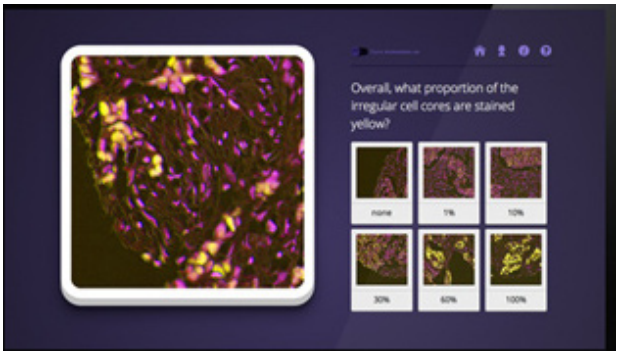
(a) FoldIt [16]



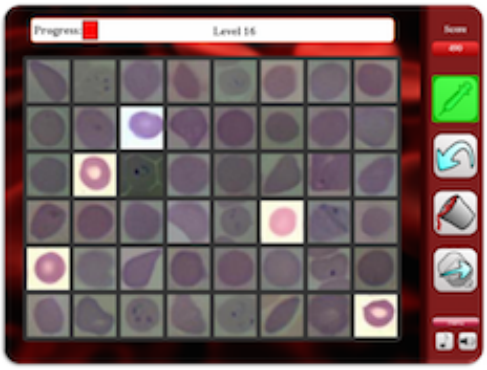
(b) Phylo [34]

Figure 2.3: Crowdsourcing in bioinformatics.

the number of gaps," [34]. Within the first seven months of its deployment, over 254,000 puzzles were solved, resulting in a 70% improvement over the computationally-generated sequence alignments [34, 26].



(a) CellSlider[a][20]



(b) Malaria Diagnosis Game [56]

[a]Image Source: [1]

Figure 2.4: Examples of applications for crowdsourcing biomedical analysis.

In diagnostic medicine, Mavandadi *et al* (2012)[56] and Luengo-Oroz *et al* (2012)[51] crowdsourced the analysis of red blood cell smears (Figure 2.4) to assist in the identification of malarial infection, and achieved expert-level performance [26]. When it comes to a malaria diagnosis, confirmation of a negative diagnosis can take up to twenty minutes

for an expert [51]. In addition, completely automated approaches are not as robust due to the variable appearance of parasites and image quality [51]. To test the feasibility of crowdsourcing in analysis of blood smears, Mavandadi *et al* (2012) [51] gave non-experts a grid of images collected by light microscopy, containing samples of red blood cells. The goal was to classify which samples were healthy or infected. Diagnostic decisions made by non-expert participants in Mavandadi *et al* (2012)[56] were within 1.25% of those made by a medical professional. Similarly, in Luengo-Oroz *et al* (2012)[51] non-experts aided in malaria infection detection, but instead were required to specify the location of the malarial parasites in the microscopic image. In this case, the aggregate crowd achieved a parasite counting accuracy of over 99%.

In CellSlider [20] (Figure 2.4), non-experts were used to identify cancerous cells and score estrogen receptor expression (associated with survival) in images of breast cancer tumor cores, with high accuracy. Given a sub-image of a tissue microarray, non-experts were asked to identify the presence of cancer cells, provide an estimate of their amount, the proportion of cells stained positive, and the intensity of their stain. This information was then combined and utilized to approximate the Allred scoring system, commonly used in practice for identifying tumors [20]. Both the crowd-based Allred score and the expert scores were then utilized to detect an association between estrogen receptor expression and disease prognosis. Candido *et al* (2015) [20] concluded that the crowd-based scoring is sufficiently accurate to detect this association.

In the detection of colorectal polyps, the precursor to malignant colorectal cancer, from computed tomographic (CT) images, there were no significant difference between aggregated crowd detection and automated techniques [60], indicating that "minimally trained ... workers could perform expert-level task[s] rapidly and with high quality," [26]. Such rapid, high quality work has also been demonstrated in the categorization of optic fundus photos, with early detection being important for the prevention of vision loss [10]. In both cases, crowd workers were given images of CT and optic fundus images, respectively, and were required to classify each as normal or abnormal. Finally, in Warby *et al* (2014)[89], non-expert consensus outperformed some automated detection algorithms in the identification of sleep spindles in electroencephalography (EEG) recordings, an important feature in the diagnosis of several neurological diseases. In this task, crowd workers not only had to identify the presence/absence of spindles, but in cases where they did exist, define a bounding box indicating its onset and offset.

In addition to platforms created for specific diagnostic purposes, like the ones mentioned above, there are other systems devoted to medical crowdsourcing on a case-by-case basis.

Such platforms include DocCHIRP [77], CrowdMed [58], or the mainstream Figure 1[2] application. In these systems, people can post medical cases and receive feedback from the crowd (including both non-experts and experts) on diagnostic possibilities. With Figure1, there is even the ability to page an expert in the field, which sends an alert to a verified specialist [2].

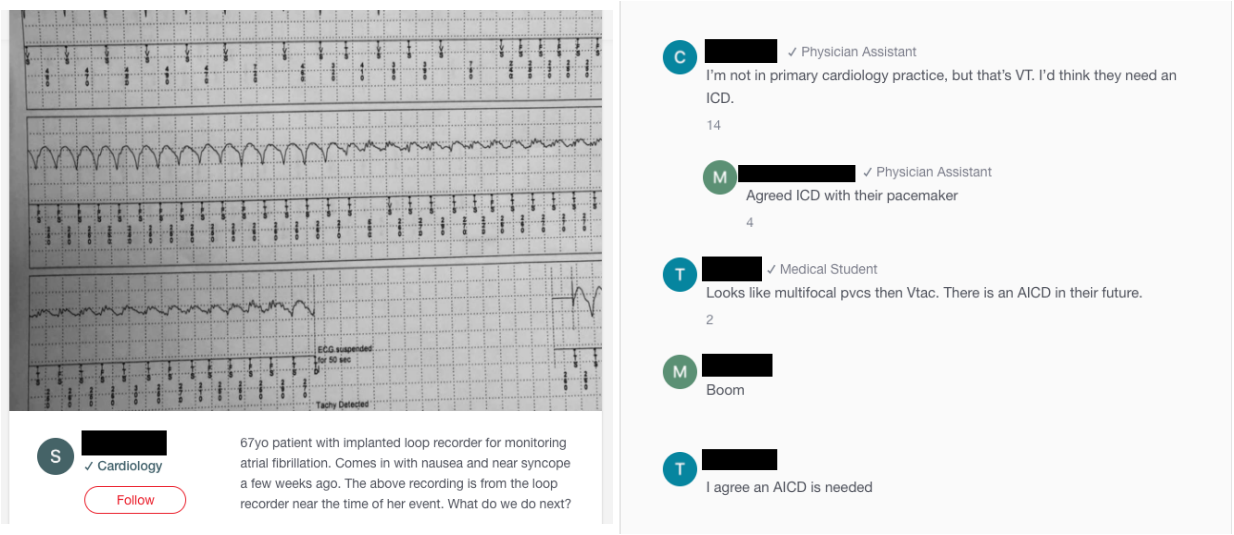


Figure 2.5: Figure1 application where people can post medical cases and receive feedback from users.

Although there has been success in medical crowdsourcing, there is a valid concern behind having non-medical professionals provide medical analysis. However, Mavandadi *et al* (2012)[56] argues that crowdsourcing can still be used to relay the data to a medical professional, who can then make the final diagnosis [69]. For example, a pathologist must look at more than 1000 red blood cells (RBC) to determine whether a given sample is negative, but if the infected cells can be identified via crowdsourcing, all a pathologist has to do is confirm the diagnosis with a single image [69]. As a result, crowdsourcing has not only shown to produce quality analysis at scale, but has the potential to increase the volume of such analysis without affecting accuracy, motivating its use for heart sound classification.

[2]http://figure1.com

2.5.2 Crowdsourcing Audio Analysis

Another relevant domain in which crowdsourcing has been applied to is the analysis of audio data. In music, information such as genre, mood, or instrumentation is important to musical information retrieval (MIR) researchers in solving music classification and recommendation problems [40]. Utilizing crowdsourcing for tag generation of audio has shown to be a valid approach for collecting accurate and meaningful labels which can then be used to train predictive models [85, 44]. In MoodSwings [37], players are paired together to collaboratively provide labels on how the mood of a music clip changes over time. Given a game board with a continuum of mood ratings, the system captures each player's mouse movements over the board as the audio plays. Scoring is based upon how close each player's rating is to the other, providing an incentive for producing "high-quality labels that others can agree upon," [37]. MajorMiner [55], The Listen Game [85] and Tag-A-Tune [44] also utilized players and their level of agreement to provide high quality descriptive tags for music. Both Turnbull *et al* (2007)[85] and Law *et al* (2007,2010)[44, 42] then used tags from their work in predictive models for music annotation.

In MajorMiner, player's were required to listen to a music clip and provide a list of descriptive tags [55]. The goal of the game was to collect original, yet relevant tags for audio [55]. In the case of collecting relevant tags, users only scored points when other users agreed with them. To foster originality, users were also given points for being the first to provide a particular tag. They found that the most popular/stable tags were relevant in describing the characteristics of music [55].

The Listen Game differs from MajorMiner, in that instead of getting users to provide labels, the system provides a set of random vocabulary words [85]. Players must then choose the best and worst word to describe the given audio clip. Song-label associations and player scores were then determined to be weighted values based on percent agreement. A supervised multiclass labelling model was then trained to predict song-label associations in addition to the strength of such association [85].

TagATune [44] also utilizes agreement to provide tags for music, however in this case, players must collaborate (and thus agree upon labels) in real-time in order to progress. Law *et al* (2010) [42] then created a topic model based on these labels and showed how they can be used to generate labels for other music clips.

With a lot of work focusing on the generation of tags by use of social or crowdsourcing methodologies, Dulacka *et al* (2012)[22] focused specifically on validation of this music

metadata produced by such methods. They created a game in which a player was presented with a music track and several sets of annotations produced from various methods. The player must guess the most relevant set and provide a level of confidence in their answer (in the form of betting their own points) [22]. Correct guesses are rewarded and incorrect guesses lead to point deduction. Tag sets are continually altered for each music clip so that relevant tags can be recognized [22]. The game illustrated that the crowd was able to filter out music metadata that was not usable, on a global scale [22].

Similar work in crowdsourcing audio analysis has been done with acoustic scene classification, where crowd workers have located and/or classified sound events in recordings. Zhang *et al* (2017)[95] were able to derive the classification of animal sounds, such as distinguishing between different types of birds and mammals, by asking workers whether pairs of audio recordings sounded similar to one another. They utilized a triplet comparison approach, by asking questions of the nature "Is a more similar to b than to c," [95], stating it is easier for humans to compare two objects rather than determining their exact identity. A plurality voting rule is used to combine decisions and come to a crowd-based classification [95]. Utilizing this method, the crowd received a high classification accuracy for the given dataset [95].

In Shamir *et al* (2013)[76], citizen scientists were asked to match whale calls with other similar sounding whale calls, which were then used to create a phylogeny of whales. A user was given a spectrogram and corresponding audio recording, along with thirty six other randomly generated examples. They were then required to find a similar sounding recording within this set. The similarity of whale calls can then be estimated by comparing the ratio of matches to anti-matches [76]. The results provided an informative analysis of whale phylogeny, even though the crowd was not asked to identify specific calls or classify individual whales and species [76]. The positive outcomes produced by the crowd in these expert-related tasks provides additional motivation for using the crowd as a tool to assist in heart sound classification.

Designing For Audio Analysis

It is important to note that the design of audio annotation tools can effect the quality of their output [11]. Cartwright *et al* (2017)[11] emphasizes the limited amount of research that has been conducted on the design and evaluation of audio annotation tools in comparison to other domains. Their study investigated the tradeoffs between reliable and

redundant annotations as well as the effects of sound visualization and sound complexity on annotation quality by crowdsourcing the annotation of sound events in synthetic city soundscapes.

Their research is especially important because the task of defining and classifying such events in city soundscapes is similar to that in murmur detection (ie. the workers must locate and define the boundaries of various audio regions). Of relevance is the finding that certain sound classes led to discrepancies between "the perceived onset and offset times when a sound is in a mixture and when ... in isolation," [11] which may present itself in cases of murmurs overlapping with fundamental heart sounds.

In addition, our work extends their open-sourced annotator to provide additional functionality for audio analysis (see Section 3.2.1).

2.6 Human-Machine Classification Frameworks

One of the use cases for crowdsourcing in machine learning is to see if the crowd can be used as a tool to accurately collect annotations and/or labels for unlabeled data (to be used in training a learning algorithm), or give feedback about instances in which a learning algorithm is uncertain. Such examples include Flock, which uses the crowd to generate informative features in cases where machine-extracted features are not predictive, or to improve algorithm performance in subregions of the input space [13]. The system Chimera utilizes the crowd to evaluate classification models of product labels and descriptions [81]. Cases deemed incorrect or ambiguous are forwarded to in-house analysts, who develop rules and update models to address these issues [81]. Other frameworks exist that directly embed an oracle into the learning process, and are termed active learning frameworks [13].

2.6.1 Active Learning

Active learning is a type of machine learning where the learning algorithm is allowed to "choose the data from which it learns," [74]. An active learning algorithm often starts with a small number of labelled instances and then requests labels for unlabelled instances based on a number of querying strategies [74]. It then learns from these results and uses them to determine which instances to query next [74].

When it comes to querying strategies, the most common method is uncertainty sampling [47, 74]. This strategy queries the instances that the learner is the least certain about, or in terms of binary classification problems, "those whose posterior probability of being positive is nearest 0.5," [74, 47, 48]. Another popular strategy is query by committee, where multiple models are trained on the same set, and the instance to query next is the one in which there is maximal disagreement between the models [75]. Other strategies involve selecting instances which would have the greatest influence on the model or those that would reduce the expected generalization error [74].

Active learning has been applied to a number of different problems. In the domain of biosignal classification, Wiens *et al* (2010) [90] used active learning to create a patient-adaptive model for heart beat classification of electrocardiograms (ECG). Research has shown that patient-adaptive classifiers provide increased classification accuracy, however, are not often used in practice because they require an unrealistic amount of labor and data [90]. By applying an active learning strategy to beat classification, Wiens *et al* (2010) [90] outperformed some of the current state-of-the-art algorithms and rule-based methods using less data.

Similarly, Lawhern *et al* (2015) [45] utilized active learning in the classification of electroencephalograms (EEG) to improve existing EEG artifact detection classifiers. They utilized a query by committee approach to select which instances to query next. Their results showed that a classification accuracy similar to models trained on a full data set can be achieved with less than 25% of the data when using an active learning strategy [45].

In the application of automatic speech recognition, Hakkani-Tür *et al* (2002) [29] explored a query strategy inspired by the classical certainty-based method in active learning, to select instances that their speech recognizer may misrecognize. They first computed a confusion network from the output of the speech recognizer and then used this network to compute a confidence score for a given utterance (set of words). The utterances with the lowest confidence score are transcribed by a human and fed back into the model [29]. Utilizing an active learning strategy, they achieved the same accuracy as random sampling utilizing 27 % less data [29]. Work by Kuo *et al* (2005) [38] also explores an active learning based model to speech recognition, but uses the minimum expected error approach to selecting instances. In this case, instances are selected that are likely to correct the most training errors [38]. Their results show that at small sample sizes, the approach has significant gains in accuracy in comparison to certainty-based query strategies. However, at larger sample sizes, there was no significant effect [38].

Active learning has also been used in text classification. Tong and Koller (2001) [83] as well as Yang *et al* (2009) [92] both employ active learning for text classification utilizing support vector machines (SVM). Both utilize a query strategy that selects instances based on those that would minimize the model loss, however Yang *et al* (2009) [92] extends this to the multi-label space. Similar results are shown to that in other domains, where gains in classification performance can be achieved with less training data [92, 83].

Our framework, introduced in Section 3.4, utilizes a modified uncertainty sampling with a pre-trained classifier, to determine when to accept a classifier's output or forward the instance to a human (crowd or expert). Work by Nguyen *et al* (2015) [59] also focuses on choosing labels from the crowd and experts, however the focus of this thesis differs. In Nguyen *et al* (2015) [59], they select an instance to query, and then decide whether to forward it to the crowd or expert for classification. The classifier is then re-trained and the process repeats until some budget is exhausted. Our focus is not on iteratively training a classifier, but on achieving the highest accuracy possible for a given batch of instances by using a query strategy from active learning to route these instances to the proper resources. Such resources could include accepting a classifier's output, or forwarding the instance to the crowd or expert (Section 3.4). Although the loop can be closed and the classifier re-trained, it would not be on a per-instance basis, and is out of scope for this thesis.

2.7 Co-Training for Human Collaboration

Co-training, proposed by Blum and Mitchell (1998) [8], is a concept in machine learning where two learning algorithms are trained on separate views of data, and then each algorithm's "predictions on new unlabeled examples are used to enlarge the training set of the other," [8]. In their example, the phrase "research interests" on a web page could be a weak indicator that a page is a faculty page, and the hyperlink text "advisor" could be an indicator that the page being linked to is a faculty page [8]. Both the text on the web page and the hyperlink text represent two different views of data. Pages with the hyperlink text "advisor" could then be used to further train an algorithm based on the text in the document, and vice-versa [8].

Of particular relevance, is the work by Zhu *et Al* (2011) [97], where they create a human collaboration policy for a categorical learning task based on the initial co-training algorithm. In this task, Alice and Bob label s unlabeled items that they are most confident

about. However, Alice sees the data from one view, whereas Bob sees the same data from a different view [97]. Alice can then see Bob's labels (from her own view) and decides whether to accept/believe Bob's labels and vice-versa [97]. Data labeled by either individual is removed from the set of unlabeled instances, and the process continues until the unlabeled data is exhuasted [97]. One of the strategies evaluated in our framework, called Crowd Ensemble (Section 3.2.5) is inspired by the work from Zhu *et Al* (2011) [97] and Blum and Mitchell (1998) [8], giving crowd workers different views of data in order to come to a decision of the classification of a given instance.

Chapter 3

Data and Methods

The following chapter describes the dataset and methods used to facilitate and evaluate binary heart sound classification ("Normal" or "Abnormal") in both humans and algorithms individually, as well as in a combined framework.

3.1 Heart Sound Dataset

A total of thirty audio files were sampled from the CinC dataset, published by [50]. These recordings covered four different heart conditions: Normal, Aortic Stenosis (AS), Mitral Regurgitation (MR), and Mitral Valve Prolapse (MVP), although more conditions were present in the full dataset. These three abnormal conditions were chosen as they are among the class of heart abnormalities known as murmurs.

In total, fifteen normal heart sound recordings and fifteen abnormal recordings (five from each abnormal heart condition) were used. Although more normal cases are presented in the population than abnormal, a balanced design between normal and abnormal was chosen in order to better understand the effects of different variables on the response variables studied.

Approximately ten consecutive beats were then sampled from each recording, making each audio recording around ten seconds in length. The CinC dataset provided the ground

truth classification for each heart sound recording, however did not include any information regarding the locations of the murmurs in the recordings. A caridologist was recruited to provide this information for each abnormal recording in the subset.

3.2 Crowd Annotation Framework

To study the ability of crowd workers to accurately classify heart sound recordings, we conducted a study on Amazon's Mechanical Turk (MTurk).

Figure 3.1: Crowd annotation interface for classifying and annotating heart sounds.

3.2.1 Audio Annotator

To run this study, we needed a web-based audio-annotation tool that crowd workers could use to listen to, analyze, and compare audio files. We extended the Audio Annotator tool initially developed by Cartwright *et al* (2017)[11] to be appropriate for bio-acoustic signal analysis. The initial audio annotator, as described in Cartwright *et al* (2017) [11], incorporated the following features:

- *Visualization*: A large, rectangular visualization of the waveform.

- *Seeking*: A user can seek to a specific time in the audio by clicking on the desired location in the x-dimension of the visualization. The user can then play/pause from this location.
- *Creating an Annotation*: Click and drag within the visualization to identify a region. Multiple annotations can be made with overlapping time intervals.
- *Deleting an Annotation*: Click the small "x" button within the annotation label's top right corner.
- *Moving/Resizing an Annotation*: Click and drag the annotation to move it to a desired temporal location or resize the boundaries by dragging the left/right edges of the annotation.
- *Playing an Annotation*: Click on the play button within the annotation's label.
- *Selecting an Annotation*: A user can select/deselect an annotation by double clicking on the annotation label. This activates the annotation (ie. displays a bounding box) and allows for the annotation to be edited.
- *Assign a Class*: Once an annotation is active, a class within each defined category can be assigned to the annotation (eg. event:jackhammer, proximity:near).

The extensions made to the annotator (Figure 3.1) allowed for the following additional functionality:

Zoom and Pan/Scroll

Heart sound analysis occurs at a much finer time resolution than soundscape analysis (the initial purpose of the annotator used in Cartwright *et al* (2017)[11]). As a result, a zooming functionality was needed so that crowd workers could work at this finer granularity. As the size of the visualization window stayed the same, the ability to pan/scroll was also implemented so crowd workers could still work with the whole waveform.

Volume

The audio files in the CinC dataset [50] were collected from a variety of sources with varying degrees of noise. Other than normalizing the audio, no other pre-processing was done (in order to preserve the original quality of the audio), making it important for crowd workers to be able to adjust the volume for proper hearing.

Global Classification

In the original version of the annotator [11], labels could only be attached to individual annotations in a recording. We therefore added a global classification functionality, that allowed a label to be applied to an entire audio recording. This gave crowd workers the ability to classify entire recordings as normal or abnormal.

Support for Contextual Information

Discerning the stages of the cardiac cycle (S1, Systole, S2, Diastole) is an essential steps in the analysis of heart sound recordings [79]. Heart sound segmentation allows for the subsequent detection and classification of pathological events [50, 79, 17]. Therefore, it was important that the segmentation information was provided to crowd workers to aid in their analysis. The reference segmentation data was provided by the PhysioNet/CinC dataset and was also available to the algorithms for training [50]. The segmentation needed to be displayed in a way that did not interfere with the audio waveform and any annotations a worker created. We represented the segmentation as a set of labels (just like how the annotation labels would appear) that were aligned at the bottom of the label stack (Figure 3.1). Taking advantage of the existing framework, this allowed users to see the segmentation in a familiar way, listen to individual segments, and easily distinguish the location of their annotations with respect to the segmentation.

Support for Rules to Guide Work

By utilizing existing knowledge about the location of murmurs, we can define rules to help guide workers in the murmur detection task. We implemented two rules for the murmur

(a) Defined murmur crosses beat boundary

(b) Multiple murmurs in one beat

Figure 3.2: Rules to guide work in murmur detection task.

detection task. As murmurs often do not cross beat boundaries, we limited a worker's annotation to be within a single heart beat. Secondly, we limited one annotation per heart beat (ie. one murmur is defined within a single heart beat). A warning symbol would appear above the annotation (Figure 3.2) and would not allow the worker to proceed to the next clip if either of these two rules were broken.

Example Viewer

The example viewer (Figure 3.3) is a modified version of the annotator interface that allows workers to reference various heart sound examples. It contains all the same functionality as the main annotation interface with the exception of allowing workers to define regions. This allows workers to compare various heart sounds against one another and with the current clip being analyzed.

3.2.2 Pre-Study Questionnaire and Screening

Before completing the classification task, crowd workers had to complete a pre-questionnaire and hearing screening test. The pre-questionnaire (see Appendix A) included questions of demographic nature (eg. age, gender, education, career) in addition to medical affiliation, musical ability and headphone type. This information was collected in order to identify whether a relationship exists between a worker's background and their performance in the study, which could be used to filter or weight the contributions of future workers.

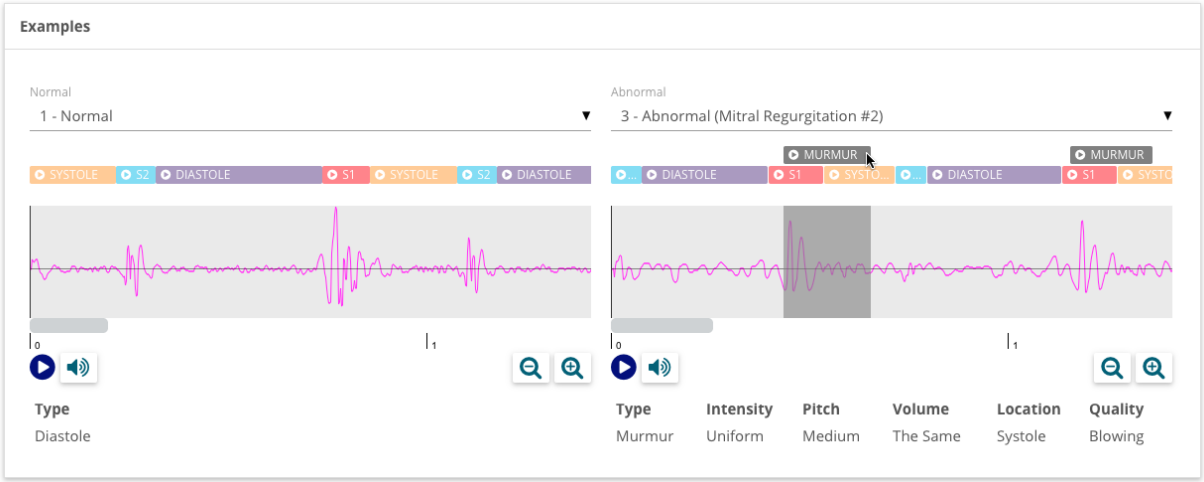


Figure 3.3: Example viewer which allows workers to compare different types of heart sounds.

The hearing screening test[1] ensured that workers were listening over adequate headphones or speakers, and were not hard of hearing. In the test, the workers had to listen to two audio recordings and count the number of tones that they heard in the recording. Workers were only allowed to continue if they were successful with counting the tones in both recordings. The tones ranged from a variety of frequencies, with some that could not be heard if the worker was hard of hearing or listening through inadequate speakers.

3.2.3 Human Intelligence Tasks

For both the binary heart sound classification task (ie. "Normal" or "Abnormal") and the murmur detection task, we created separate HITs (Human Intelligence Task) which contained ten recordings out of the total possible thirty recordings. Five recordings were randomly selected from each condition (normal or abnormal) and the order of recordings presented to a given worker was randomized. Workers were paid $4.00 to analyze all ten recordings. Restrictions were in place to ensure that each worker only completes the study once. Workers were also required to watch a training video and complete a training round in order to familiarize themselves with the interface and task.

[1]https://github.com/mcartwright/hearing-screening.js

In the training task, once a worker submitted their work for a given audio clip, they were shown the correct answer and had a chance to review their work before progressing to the next training example. The training clips shown (two normal and three abnormal), were also sampled from the CinC dataset [50] but were separate from the thirty recordings collected for evaluation purposes.

Workers who did not analyze all ten clips or those who did not play the audio recording at all during the task were filtered out from the analysis.

Classification Task

In the binary heart sound classification task (Figure 3.1), workers were required to listen to each audio clip and classify the recording as "Normal" or "Abnormal".

Murmur Detection Task

In the murmur detection task, workers were asked to detect the presence or absence of murmurs in the ten heart sound recordings. Workers were told to define the boundaries of all murmurs within a heart sound recording, if any existed. If they thought no murmurs existed, they had to click a box indicating the absence of murmurs. The interface would not let them continue to the next recording until they either defined a murmur or indicated the absence of them.

3.2.4 Classification by Proxy

In addition to the classification task as a method for heart sound classification, a decision on the normality of a given heart sound can be determined by utilizing the murmur detection task as a proxy. If a worker defines the presence of murmur(s) on a given recording, the recording is subsequently classified as abnormal. Similarly, selecting the checkbox indicating the absence of murmurs indicates a normal recording. The majority vote determines the final crowd classification of the recording.

3.2.5 Crowd Ensemble

A third crowd-based classification method utilizes the information from the first two methods, and is inspired by the work of Zhu et al (2011) [97]. In this method, we look at instances where there is disagreement between the Normal/Abnormal voting and Classification by Proxy methods, and define the final classification for a given instance to be the output from the method that is the most confident in its answer (see Equation 3.3).

3.3 Machine Classifiers

The machine classifiers used were open-sourced entries from the CinC Challenge [14]. We selected four entries to use in the evaluation of our framework, with the restriction that these models produced some probabilistic output. The models were selected based on the top scoring entries from the challenge, as listed on the PhysioNet website[2]. The following section gives an overview of each classifier used and how it was incorporated into the hybrid framework.

As the testing set for the classifiers in the CinC challenge was hidden from the public, our subset of thirty records were sampled from the public dataset (ie. the challenge training set)[50]. Therefore, of the 3000+ records in the training set, these methods may have been trained using some of the records in our subset.

3.3.1 Potes *et al* (2016)

The classifier developed developed by Potes *et al* (2016)[67] was the top performing entry in the challenge. Their method utilized an ensemble of the AdaBoost classifier and a Convolutional Neural Network (CNN). The AdaBoost classifier was trained using 124 time-frequency based features and the CNN utilized segmented cardiac cycles decomposed into four frequency bands as input. A final decision rule was used to determine the overall heart sound classification, where an instance was considered abnormal if either method computed the probability of such instance being abnormal to be greater than 40%.

[2]https://physionet.org/challenge/2016/sources/

3.3.2 Kay and Agarwal (2016)

Kay and Agarwal (2016)[35] utilized a DropConnected Neural Network trained on time-frequency and inter-beat features for heart sound classification, ranking third overall in the challenge [15]. They trained a number of different networks on a range of hyper parameters and training sets [35]. The final classification is given by the majority vote of the ensemble, with the probability of a given instance being abnormal being equal to the average probability of abnormality.

3.3.3 Bobillo (2016)

The method developed by Bobillo (2016)[9] ranked fourth in the challenge and utilized a tensor-based approach to heart sound classification. Time-frequency based features were calculated on full heart beats as well as each stage in the beat, creating a 3-way tensor for each recording. These tensors were then concatenated into a 4-way tensor and reduced using Tucker discrimination to get a tensor of higher discriminatory power. This was then fed into a K-Nearest Neighbor classifier. The decision boundary was adjusted to 0.225 to account for the imbalanced data set.

3.3.4 Maknickas and Maknickas (2017)

Maknickas and Maknickas (2017)[54] utilized a Deep Convoluted Neural Network (CNN) trained on mel-frequency spectral coefficients (MFSC), ranking sixth overall in the challenge [15, 54]. The MFSC were computed on each recording and divided into frames of 64ms in length. An equal number of normal and abnormal frames were used for training. The trained CNN then predicted a normal/abnormal label for each frame in the testing set, with the final label being equal to the majority label.

3.4 Hybrid Human-Machine Framework

The hybrid framework combines both machine and human classifiers to come to a final decision about the classification of a given heart sound recording. However, the system

does not query the crowd on every instance, but only on those where the classifier is uncertain. In binary classification problems, uncertainty sampling queries "the instance whose posterior probability of being positive is closest to 0.5" [74, 47, 48]. In our framework, we define uncertain instances i as:

$$i = \{x \in D \mid \ |P(x = \text{Abnormal}) - t| \leq w\} \text{ for a given } w \geq 0, t \leq 1 \tag{3.1}$$

where x are all the instances in the dataset D whose probability of being abnormal is within the window size, w, from the classifier's decision margin t. For example, given a classifier whose decision margin between Normal and Abnormal is $t = 0.5$, a $w = 0.1$ would send all instances x to the crowd whose probability of being abnormal is between 0.4 and 0.6. The framework also supports full machine classification ($w = 0$ or baseline classifier accuracy) and the ability to send all instances to the crowd ($w = 1 - t$).

When an instance is sent to the crowd, the classification is determined by majority voting, and the probability that the crowd believes a given instance is abnormal is defined by percent agreement:

$$\%Agreement_{Abnormal} = \frac{\#\text{ Abnormal Votes}}{\#\text{ Normal Votes} + \#\text{ Abnormal Votes}} \tag{3.2}$$

In cases where a classifier and the crowd disagree on the classification of a given instance, the final decision is made by using the method (crowd or classifier) that is most certain about its given classification:

$$FinalClass = \underset{c \in \{\text{Normal,Abnormal}\}}{\arg\max} (\max(|P_{Classifier}(x = \text{c}) - t_{Classifier}|, |P_{Crowd}(x = \text{c}) - t_{Crowd}|)) \tag{3.3}$$

where P is the probability that a method has classified a given instance x as c and t is the decision margin for that given method. For example, given a decision margin of $t = 0.5$ for both methods, if $P_{Crowd}(x = \text{Abnormal}) = 0.2$ and $P_{Classifier}(x = \text{Abnormal}) = 0.6$, the crowd method would be used as the final decision. This is because $|0.2 - 0.5| = 0.3 > |0.6 - 0.5| = 0.1$ indicating the crowd is more confident in its classification than the machine classifier. Note that we refer to this difference (eg. $|0.2 - 0.5|$) as the decision difference. A pseudocode representation of the framework is presented in Figure 3.4.

3.4.1 Expert Querying

Just as we impose a certainty threshold on the machine classifier, we can do the same for the instances classified by the crowd. Similarly, if the decision difference of the crowd is less than the threshold, w, we send the instance to an expert for classification. For the purposes of simulation, we assume that the expert returns the correct (ground truth) answer. We can then calculate the precision, recall and F1-scores (Section 3.5.1) over increasing values of w to assess how the performance changes when adding an expert to the process. In addition, we can also compare the percent of instances sent to an expert for classification when the crowd is present or absent in the process to get an understanding of the crowd's benefit on heart sound analysis.

```
/**
w is the window value
clf is the trained classifier
D is the list of recordings
**/
procedure HybridFramework(w, clf, D)
    allClassifications ← []
    decisionMargin ← clf.decisionMargin
    crowdDecisionMargin ← 0.5
    for record in D do
        label, probAbnormal ← clf.predict(record)
        finalClass ← label
        decisionDiff ← |probAbnormal − decisionMargin|
        //Is the classifier's level of certainty below the threshold?
        if decisionDiff ≤ w then
            crowdLabel, crowdProbAbnormal ← sendToCrowd(record)
            crowdDecisionDiff ← |crowdProbAbnormal − crowdDecisionMargin|
            if label ≠ crowdLabel then
                //Is the crowd more confident than the classifier?
                if crowdDecisionDiff > decisionDiff then
                    finalClass ← crowdLabel
                end if
            end if
        end if
        allClassifications.append((record, finalClass))
    end for
end procedure
```

Figure 3.4: Pseudocode representation of hybrid framework with no expert involvement

3.5 Analysis Methods

The following section presents the methods used to evaluate the machine and crowd classifiers separately and from within the hybrid framework. Specifically, we give a description of the metrics used for both evaluating these classifiers (machine, crowd, hybrid) and

for measuring the effect of crowd contribution on overall classification performance. We also present the models developed to investigate whether there are any pre-experiment attributes (ie. pre-questionnaire responses, training performance) that are indicative of performance in both tasks, and if so, how they can be used to increase performance and minimize cost.

3.5.1 Crowd and Machine Performance

To evaluate the performance of the crowd (normal or abnormal voting, classification by proxy, ensemble), machine classifiers and the hybrid framework in binary heart sound classification, we compute precision (P), recall (R) and F1-score (F1) of each method by comparing the output with the ground truth. These measures are defined as:

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}, \quad F1 = \frac{2PR}{P + R} \tag{3.4}$$

where TP, TN, and FN are the number of true positives, true negatives, and false negatives respectively.

Window Size Changes on Framework Performance

If we vary the window size, w, used in the hybrid framework, we can evaluate how the final F1 score changes across machine classifiers. We can then use this evaluation to suggest a reasonably efficient windowing value for use with machine classifiers. By increasing the value of w, we impose a greater restriction on the initial acceptance of a classifier's output for a given recording. That is, as we increase w, a classifier must be increasingly more confident about its label for a given instance, or else that instance will be sent to the crowd for classification. However, as mentioned in Section 3.4, this does not mean the crowd generated classification will be used as the final answer. The final classification is still based on which method (crowd or machine) is more confident in its output.

Evaluating Classifier Bias

Among the different crowd-based strategies and machine classifiers we can evaluate whether any of these techniques are biased towards certain outcomes. We aim to understand whether a given condition (Normal, MR, MVP, AS) is more likely to be classified as abnormal in addition to which method is more likely to categorize a clip as abnormal.

To study this, we first create a boolean variable representing whether the probability of a given instance being classified as abnormal by a given method is greater than that method's specified decision margin. We then fit a logistic regression model with this boolean being the dependent variable and the condition and method being the independent variables.

Crowd Query Frequency and Accuracy

In evaluating the hybrid framework, it is important to analyze, of the instances queried (sent to the crowd), what proportion:

1. Did the crowd classify correctly?
2. Was selected as the final answer?
3. Was correctly classified in the final answer?

These percentages are calculated for each combination of machine classifier and crowd method, averaged across all window sizes. By answering these questions, we can gain an understanding of the effect that crowd contribution has on the final classification and whether including the crowd is useful.

Murmur Detection Performance

In addition to evaluating binary classification performance by proxy of murmur detection, we can also investigate whether the murmurs detected by the crowd are actually correct. We compute aggregate annotations for a given recording by first dividing each recording into non-overlapping, fixed-length (eg. 100 ms) time frames. This is similar to how Cartwright

et al (2017) [11] analyzed their collection of crowd-based annotations of sound events in city soundscapes. We then take the majority vote of the presence or absence of an annotation in each time frame where the population is the total number of people that defined at least one annotation in the recording. A murmur (or part there of) is considered to exist if at least half of all votes fall within the majority and the population is greater than one person. Once we have an aggregate annotation for a given recording, we can compute the above classification measures on a frame-level basis as implemented in sed_eval, a python library for sound event detection and evaluation [57].

Once the F1 score for each recording is calculated, we can perform a Wilcoxon One-Sided Signed-Rank test [91] to test if the murmur detection F1 scores are significantly greater than 0.5 (random).

3.5.2 Indicators of Crowd Performance

Given the data submitted from each crowd worker in the pre-questionnaire (Section 3.2.2) as well as their performance in the training round, we can examine whether there exists a relationship between these pre-experiment attributes and their ability to perform binary heart sound classification and/or murmur detection. If such relationship(s) exists and are significant, they may be able to function as a mechanism to filter out poor performers.

In the case of binary heart sound classification (both Normal/Abnormal voting and Classification by Proxy), we fit a logistic regression model to the data. In order to select the variables for the model, we use the stepwise selection technique to select from the list of independent variables that include both main and two factor interaction effects. The Akaike Information Criterion (AIC) was used by the stepwise selection technique to decide whether a variable is included or not in the model.

We can then perform an analysis of variance (ANOVA) to compare the effects of the pre-questionnaire data and F1-Score in the training round to see if any of these attributes, or their interactions, have an effect on an individual's final F1-Score.

In further analyzing the crowd's murmur detection performance, we can fit a linear model to investigate the relationship between a user's mean F1-Score on the abnormal training instances, and their F1-Scores for abnormal instances in the experiment.

3.5.3 Summary of Variables

The following gives a summary of variables used, including their names, descriptions and type.

Variable	Description	Type
userID	A unique identifier for each crowd worker.	Label
$\text{Var}(\text{F1}_{training})$	The variance in a user's F1-Score in the training round.	Continuous
$\overline{F1}_{training}$	A user's mean F1-Score in the training round.	Continuous
agreement	Whether the majority vote classification is the same as the ground truth.	Binary
numParticipants	The number of participants who submitted an answer for a given clip.	Discrete
w	Otherwise known as the window size or windowing parameter.	Continuous

Table 3.1: Summary of variables

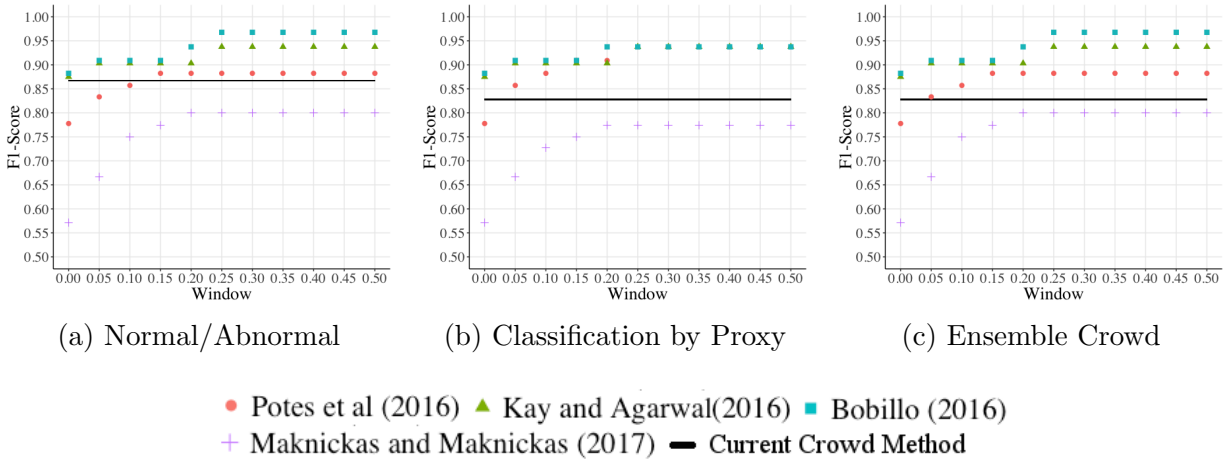
Chapter 4

Results and Discussion

The following chapter presents and discusses the results from the evaluation of the crowd, machine and hybrid framework on binary heart sound classification. In total, 89 crowd workers completed the Normal/Abnormal classification task and 67 crowd workers completed the murmur detection task.

4.1 Hybrid Framework Performance

The evaluation of the hybrid framework, utilizing different crowd and machine methods, is summarized in Figure 4.1. The results indicate that an increase in F1-Score for binary heart sound classification is achieved in all combinations of crowd and machine methods, with the normal/abnormal voting and ensemble crowd strategies producing the same final classification results. The top performing combination of human and machine methods is the normal/abnormal voting (or ensemble) with the classifier developed by Bobillo (2016) [9]. Even with this classifier having the greatest initial F1-Score ($F1 = 0.882$) when used independent of our framework, the use of our hybrid approach still leads to a increase in performance, with a F1-Score of 0.968 at $w = 0.25$ (in the Normal/Abnormal voting case). In addition, classifiers with lower initial F1-Scores achieve considerable gains in performance, as seen with the classifier by Maknickas and Maknickas (2017) [54] having a baseline F1-Score of 0.571 and a F1-Score of 0.800 at $w = 0.25$ when used with Normal/Abnormal voting. Such results motivate the use of machine classifiers in a hybrid framework for increased classification performance.



(a) Normal/Abnormal (b) Classification by Proxy (c) Ensemble Crowd

Figure 4.1: Hybrid framework performance using different crowd classification strategies

We can see that as we increase the windowing parameter, w, with more instances being sent to the crowd for analysis, the F1-Scores increase and plateau at around $w = 0.25$ for all four models. As a result, this value may be an efficient value for higher overall classification performance.

When looking at the subsets of instances where the crowd is queried, we can see from Table 4.5 that the crowd performs well in classifying most of these instances correctly (#Correct/#Query). The percentage of instances that are then used (#Used/#Query) as the final answer varies (with the average being 70.9%), but is a result of the framework picking the classification from the method that is most confident in its decision for that particular instance. What is of importance is the very high number of crowd-classified instances that are correct among the crowd-classified instances that are used in the final answer (#Correct/#Used), with the lowest and highest accuracy being 86% and 100% respectively. This illustrates that when the crowd is more confident than the machine in the classification of a given instance, they are most often correct.

4.2 Crowd and Machine Performance

The results for the base crowd (normal/abnormal voting, classification by proxy, ensemble) and machine classifier performance on our subset are presented in Table 4.3. The crowd

performs well at binary heart sound classification, with the simple majority voting strategy producing the best results among the crowd based classification strategies. The F1-Score from both the classification by proxy and ensemble crowd methods are the same, however the instances in which each method classifies correctly slightly differ.

	Model Parameters			
Variable	$\hat{\beta}$	*Std. Error*	z	*p-value*
MR	-0.54	0.61	-0.89	
MVP	16.97	1066.37	0.02	
Normal	-2.66	0.52	-5.07	***
Kay and Agarwal (2016)	-0.43	0.66	-0.66	
Maknickas and Maknickas (2017)	0.21	0.65	0.33	
Normal/Abnormal Voting	-0.88	0.67	-1.31	
Classification by Proxy	-1.04	0.69	-1.50	
Ensemble Crowd	-1.12	0.68	-1.64	
Potes *et al* (2016)	0.43	0.66	0.65	

Table 4.1: Logistic model to model the effect of condition and method on classifying a given clip as abnormal.

	Model Parameters			
Variable	$\hat{\beta}$	*Std. Error*	t	*p-value*
userID	3.1×10^{-6}	9.8×10^{-6}	0.32	
Var($\text{F1}_{training}$)	0.35	0.84	0.42	
$\overline{F1}_{training}$	1.03	0.32	3.19	**
userID:Var($\text{F1}_{training}$)	1.3×10^{-4}	9.5×10^{-5}	1.40	
userID:$\overline{F1}_{training}$	-5.1×10^{-5}	4.2×10^{-5}	-1.23	
Var($\text{F1}_{training}$) : $\overline{F1}_{training}$	-1.72	1.83	-0.94	

Table 4.2: Linear model to model effect of user's training F1-score for murmur detection on experimental F1-scores.

The logistic model exploring the effect of condition and method on the ability to correctly classify abnormal instances is presented in Table 4.1. There were no significant two factor interaction effects, indicating that the condition and method do not have any dependency between each other. As a result, we chose a simpler model that did not have any interaction terms. Goodness-of-fit was validated by the Hosmer-Lemeshow [31] ($\chi^2(3, N = 208) = 1.49, p = 0.68$), Osius-Rojek [65] ($z = -0.0007, p = 1.00$) and Stukel [80] tests ($\chi^2(2, N = 208) = 3.38, p = 0.18$). The model shows that all methods perform just as well as the baseline method (Bobillo (2016) [9]) when it comes to classifying a

given clip as abnormal. Similarly, compared to the baseline AS condition, all methods are equally likely to categorize conditions MR and MVP as Abnormal, but are significantly[1] less likely to categorize a normal condition as abnormal. These results indicate consistency for both the crowd and machine methods which is especially important when it comes to classifying new data.

In analyzing the effect of the number of crowd workers on the ability to classify heart sounds, the logistic models (see Table 4.4) show evidence that as the number of crowd workers participating increase, the probability of correctly classifying the heart sound increases. This is the case in both the normal/abnormal voting and classification by proxy methods. Goodness-of-fit was validated by the Osius-Rojek [65] ($z = 0.84, p = 0.40$) and Stukel [80] tests ($\chi^2(2, N = 30) = 2.65, p = 0.27$) in the normal/abnormal model and the Hosmer-Lemeshow [31] ($\chi^2(3, N = 28) = 0.86, p = 0.84$), Osius-Rojek [65] ($z = -0.20, p = 0.84$) and Stukel [80] tests ($\chi^2(2, N = 28) = 3.49, p = 0.17$) in the classification by proxy model. The QQ plots were checked to ensure there was no clear violation of the model's assumptions (ie. the points fall on a straight line in the QQ plots, indicating the residuals are normally distributed).

When investigating the effect of user's pre-questionnaire data and training F1-Score on their final F1-Score in the normal/abnormal classification task, the stepwise selection method selected a model containing only the training F1-Score. The ANOVA showed that the effect of the user's training F1-Score on their final F1-Score was significant ($F(1, 86) = 4.96, p = 0.03$). As a result, a user's training F1-Score could be used to filter out potential poor performers in this task. We explore this possibility in Section 4.4. In the case of the classification by proxy method, the stepwise selection method produced a null model, indicating the absence of any significant factors effecting a user's final F1-Score. Therefore, any pre-filtering of workers by pre-questionnaire or training F1-Score would not prove useful with this method. As a result, the use of this method may not be ideal in cases where financial resources are finite.

4.3 Classification by Proxy and Murmur Detection

As the classification by proxy method is derived from the murmur detection task, we can also evaluate how well the crowd performed in detecting and defining the murmurs that

[1]Statistically significant results are reported as follows: $p < 0.001$(***), $p < 0.01$(**), $p < 0.05$(*), $p < 0.1$(.)

	Methods	Precision	Recall	F1-Score
Crowd	Normal/Abnormal Voting	0.867	0.867	0.867
	Classification by Proxy†	0.857	0.800	0.828
	Ensemble	0.857	0.800	0.828
Machine	Kay and Agarwal (2016)	0.824	0.933	0.875
	Bobillo (2016)	0.789	1.000	0.882
	Potes *et al* (2016)	0.667	0.933	0.778
	Maknickas and Maknickas (2017)	0.500	0.667	0.571

† Two instances resulted in ties. These are considered as inconclusive and as a result were not included in the calculation.

Table 4.3: Base crowd and machine classifier performance

	Model Parameters			
Variable	$\hat{\beta}$	*Std. Error*	*z*	*p-value*
agreement	-0.32	1.27	-0.25	
numParticipants	0.26	0.11	2.34	*

(a) Normal/Abnormal Voting

	Model Parameters			
Variable	$\hat{\beta}$	*Std. Error*	*z*	*p-value*
agreement	0.98	1.74	0.56	
numParticipants	0.72	0.27	2.69	**

(b) Classification by Proxy

Table 4.4: Logistic model to model the effect of the number of crowd workers on the ability to classify heart sounds.

lead to the abnormal classification decision for a given instance. This method could then be used to not only classify a given recording as normal or abnormal, but in the case where the instance is abnormal, provide an expert with the evidence behind the crowd's decision. The results of the aggregation method presented in Section 3.5.1 on the crowd's annotations of abnormal recordings are presented in Table 4.6. The Wilcoxon One-Sided Signed-Rank test [91] indicates that the aggregate murmur detection scores are significantly greater than random ($p < 0.001$) illustrating a degree of competency in the combined effort of the crowd to detect and define the boundaries of murmurs in abnormal recordings. To understand the effects of the disease state on the ability for the aggregate to capture the murmurs present in the recordings, we performed an ANOVA with the dependent variable being the aggregate F1-Score for a given recording. The results indicate a statistically significant ($F(2, 12) = 4.20, p = 0.04$) effect between the disease condition and the aggregate F1-Score. A post-hoc Tukey test showed that the MR and MVP condition differed significantly

Classifier	Crowd Strategy	#Correct/#Query	#Used/#Query	#Correct/#Used
Bobillo (2016)	Normal/Abnormal	0.948	0.722	1.00
	Classification by Proxy	0.772	0.506	1.00
	Ensemble	0.948	0.722	1.00
Kay and Agarwal (2016)	Normal/Abnormal	0.837	0.672	1.00
	Classification by Proxy	0.800	0.617	1.00
	Ensemble	0.834	0.672	1.00
Maknickas and Maknickas (2017)	Normal/Abnormal	0.899	0.802	0.948
	Classification by Proxy	0.815	0.765	0.945
	Ensemble	0.878	0.802	0.948
Potes *et Al* (2016)	Normal/Abnormal	0.852	0.799	0.861
	Classification by Proxy	0.748	0.634	0.930
	Ensemble	0.836	0.799	0.861

Table 4.5: Summary averages of crowd query frequency and accuracy over all windows

($p = 0.03$) in aggregate F1-Scores, however the other condition pairs, MR-AS ($p = 0.22$) and MVP-AS ($p = 0.53$) did not.

In addition, the linear model (Table 4.2) showed a statistically significant positive relationship between a user's mean F1-Score on the abnormal training instances, and their F1-Scores for abnormal instances in the experiment, indicating we may be able to weight individual user's annotations to achieve better results in defining the murmur boundaries of recordings we know are abnormal. We leave this for future work. In addition, the variance in F1-Scores for training were not significant, suggesting worker's performance may be an effect of their innate ability and not their ability to learn during the training round. It is also worth noting that recordings c0014 and c0018 have the two lowest F1-scores for aggregate murmur detection but were also classified incorrectly by the Normal/Abnormal voting method.

4.4 Filtering Workers on Training Performance

As mentioned in Section 4.2, the user's training F1 score in the Normal/Abnormal voting classification task could be used to filter out potential poor performers. To explore this, we evaluate two different filtering techniques. The first strategy removes all users who

		Aggregate Murmur Detection			Classification by Proxy		
Data Name	Type	Precision	Recall	F1-Score	Total Votes	$\frac{\#\text{ Abnormal Votes}}{\text{Total Votes}}$	Classification Outcome
c0025	AS	1.000	0.939	0.969	21	0.905	Abnormal
c0026	AS	0.903	0.875	0.889	22	0.864	Abnormal
c0019	AS	0.676	0.926	0.781	19	0.947	Abnormal
c0021	AS	0.667	0.743	0.703	22	0.955	Abnormal
c0014	AS	0.393	0.478	0.431	17	0.412	Normal
c0001	MR	0.852	0.793	0.821	18	0.944	Abnormal
c0023	MR	0.645	0.667	0.656	17	0.824	Abnormal
c0013	MR	0.875	0.452	0.596	20	0.650	Abnormal
c0004	MR	0.625	0.345	0.444	21	0.381	Normal
c0018	MR	0.389	0.269	0.318	15	0.467	Normal
a0002	MVP	0.875	1.000	0.933	19	0.842	Abnormal
a0320	MVP	0.977	0.857	0.913	23	0.826	Abnormal
a0024	MVP	1.000	0.805	0.892	16	1.000	Abnormal
a0045	MVP	0.882	0.750	0.811	14	0.786	Abnormal
a0220	MVP	0.871	0.750	0.806	18	0.778	Abnormal

Table 4.6: Aggregate crowd murmur performance and classification by proxy outcome for abnormal recordings of different disease types.

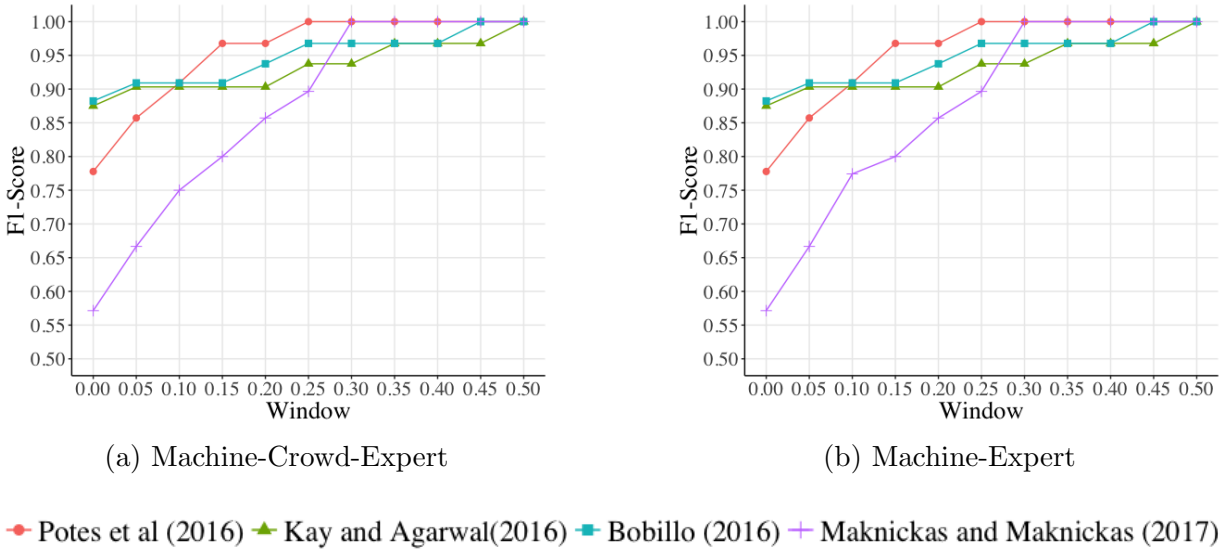
have a training F1-Score less than or equal to 0.5 (random). The second method removes a bottom percentage (5% and 10%) of users who performed poorly in the training round. We evaluate both the modified crowd performance and its effect on the hybrid framework.

Filtering by the training F1-Score being less than random and filtering by the bottom 10% produce the same result, indicating that less than 10% of users obtained a training F1-Score less than random. Filtering by the bottom 5% of performers changes one instance's classification (b0277) from Abnormal to a tie vote. The other two filters correctly alter the classification of two instances (c0018 and c0014) from Normal to Abnormal.

The top filtered crowd method produce no change when used with two of the classifiers (Kay and Agarwal (2016) [35] and Bobillo (2016) [9]) and the change of a single instance (to the correct classification) when used with the Potes *et al* (2017) [67] and Maknickas and Maknickas (2017)[54] classifiers. Although the increase in performance isn't substantial, what is important is that there was no decrease in performance. As a result, such a filtering

technique can still produce the same results, but using less crowd workers and thus less financial resources.

4.5 Expert Querying



(a) Machine-Crowd-Expert

(b) Machine-Expert

Figure 4.2: Hybrid framework performance using a combination of machines, crowd and experts.

The results from adding experts into the workflow are presented in Figure 4.2. We evaluated two types of expert involvement. The first is the original hybrid framework, with the addition of sending instances from the crowd to an expert when their certainty was below the threshold w. The second is a framework with only the machine and the expert. In this case, if the machine classifier's certainty in a given instance is below the threshold w, the instance is sent directly to the expert instead of going through the crowd.

The change in F1-scores across different machine classifiers and thresholds may be similar (only differing by one data point when used with the Maknickas and Maknickas (2017) [54] classifier at $w = 0.10$), however the level of expert involvement in the classification process is very different. Table 4.7 shows the average percent of automated classification (ie. classification without experts) across the two types of expert involvement. In all cases,

utilizing the crowd in a hybrid machine-crowd-expert framework utilizes less expert resources than a strict machine-expert framework. For example, in the case of utilizing the Normal/Abnormal crowd strategy with the Bobillo (2016) [9] classifier, 76% of the expert's work is automated when it comes to classifying heart sound instances. This means that the expert is only given 24% of instances to classify, as opposed to utilizing a framework without the crowd, when the expert must classify 33% of all instances. This result motivates the use of a framework that includes the crowd, as it reduces expert time (and by extension cost, which is often much greater) while maintaining similar classification performance.

Classifier	Crowd Strategy	Average % Automated (without experts)		Average Instances Automated (out of 30)
		Machine-Crowd-Expert	Machine-Expert	
Bobillo (2016)	Normal/Abnormal	75.8	66.4	2.82
	Classification by Proxy	71.5		1.53
	Ensemble	75.8		2.82
Kay and Agarwal (2016)	Normal/Abnormal	77.0	69.4	2.28
	Classification by Proxy	74.8		1.62
	Ensemble	77.0		2.28
Maknickas and Maknickas (2017)	Normal/Abnormal	59.7	31.8	8.37
	Classification by Proxy	54.2		6.72
	Ensemble	59.7		8.37
Potes *et Al* (2016)	Normal/Abnormal	61.2	46.1	4.53
	Classification by Proxy	58.8		3.81
	Ensemble	61.2		4.53

Table 4.7: Summary averages of automated expert work over all windows

4.6 Discussion

There are a few main takeaways from the evaluation of our hybrid framework. Firstly, the framework achieves greater performance than a baseline classifier alone. In addition, any probabilistic classifier can be used within the framework, as shown with the various machine classifiers tested. Secondly, the framework utilizes less expert resources while achieving similar performance, when compared to a framework that does not use the crowd. We have also illustrated how users can be filtered out based on their training F1-Score (in the case of the Normal/Abnormal voting method), further minimizing resources while maintaining overall classification performance.

When it comes to the two crowd-based heart sound analysis tasks, the Classification by Proxy method resulted in a slightly lower F1-Score than the Normal/Abnormal voting method, however this may be indicative of the potential difficulty of the task. Regardless, our analysis did show that the crowd has an overall competency when it came to defining and detecting murmurs in recordings they think to be abnormal. Such ability is important in evidence-based medicine, strengthing the initial argument made by Mavandadi *et al* (2012) [56] that crowdsourcing can be used to relay information to a medical professional, who can then make a final diagnosis. Although the final decision would be made by an expert, the initial analysis is made by the crowd, which can still lead to a reduction in expert time.

In both the Normal/Abnormal voting and Classification by Proxy tasks, the crowd was consistent in their performance regardless of the abnormality. Although the overall F1-Score was higher in the Normal/Abnormal voting method than the Classification by Proxy method, the latter method at least provides a reason behind the given diagnosis. One way to benefit from the mutual information of both methods is to utilize the Normal/Abnormal voting as a measure of normality and the corresponding Classification by Proxy analysis as the evidence behind such decision. However, this is only feasible in cases where there is agreement between the two methods.

For instances routed to the crowd and accepted as the final classification, the aforementioned methodology provides evidence to the final decision maker of the reasons behind the classification of a given instance. However, the question arises of how the instances classified only by the machine (or those in which the machine is more confident than the crowd) are interpreted. Many machine learning models currently exist where humans do not understand (and may be hesitant to trust) the information they contain and the rationale behind the model's decision making [39]. In addition, what about instances where a machine learning algorithm or the crowd is correct, but unconfident? Should we still trust their output or is a second opinion warranted?

Our hybrid framework alleviates these issues in two ways, the first being the use of the windowing parameter (w). Remember that by increasing the value of w, we impose a greater restriction on the initial acceptance of a classifier's output. That is, as we increase w, a classifier must be increasingly more confident about it's label for a given instance, or this instance is escalated to a human. Although this method still does not provide interpretability to the decision maker, it at least ensures a threshold of acceptable certainty. Secondly, in the use of our hybrid framework that includes expert querying, if neither the machine nor crowd reach the acceptable level of certainty, the instance is

forwarded to the expert, and as such, interpretbality from the machine or crowd is not needed.

Based on these results and discussion, the hybrid human-machine framework shows promise in the area of binary heart sound classification.

4.7 Real-World Applications

Online communities such as Figure 1[2] contain a user base of medical personnel as well as those who do not have such background, but are interested in diagnostic medicine. When a medical case is posted, containing anything from an X-ray to an ECG, users have the opportunity to weigh in on the case, regardless of their credentials [2]. The platform even has a paging feature, which sends an alert to verified specialist(s) when expert input is needed on a given case [2].

One hybrid framework could be integrated into applications like Figure1, or similarly CrowdMed [58], where patient heart data could be uploaded for analysis. Just as in our framework, a machine learning algorithm would take a first pass over the data, and then decide whether to route given instances to the users. Based on the users' analysis, we could then accept their output, the output from the learning algorithm, or page an expert for further input. Although Figure1 is volunteer-based, platforms like CrowdMed [58] do compensate their "medical detectives" for their work on medical cases, which is more in-line with our existing framework. The use of such a platform, and by extension our hybrid framework, is particularly important for medical data analysis in regions where sufficient medical resources are not available to support the population.

[2]https://figure1.com/

Chapter 5

Conclusion and Future Work

The following chapter presents a summary of the work completed in this thesis: the design and evaluation of a hybrid human-machine framework for heart sound classification. We then outline potential ways of extending this work in the future.

5.0.1 Conclusion

In this thesis, we first described the motivation behind developing a hybrid human-machine framework for the classification of heart sound recordings. This includes the past success in utilizing human computation to solve problems in medicine as well as the use of human-machine frameworks in areas such as biosignal classification, speech recognition, text and image processing. We surveyed the area of heart sound analysis, from a clinical and automated perspective, outlining past solutions and current challenges in the space. We also summarized work in the human computation and crowdsourcing space as it relates to medical data and audio analysis, as well as past work on human-machine frameworks, with a specific focus on active learning and co-training for human collaboration.

Informed by this previous work, we introduced our hybrid human-machine framework and crowd-based annotation platform for heart sound classification. The framework decides how to escalate the analysis of heart sound recordings to different resources and incorporate their decision into a final classification. It comes to a decision based upon who has analyzed the heart sound (machine, crowd, expert), their level of uncertainty,

and a threshold of acceptable uncertainty. The results indicated our hybrid framework achieved greater performance than a baseline classifier alone, and utilized less expert resources while achieving similar performance, when compared to a framework that does not use the crowd. Based on these results, the framework shows promise for analysis of other bioacoustic signals.

We also studied how crowd-based heart sound analysis tasks can be used to classify heart sounds, and how the crowd performs in each of them. Our results showed that the crowd performed well in both binary heart sound classification and in the classification by proxy method. Although the performance on the murmur detection task varied by the heart abnormality, our analysis indicated competency in the aggregate crowd to detect the abnormalities in abnormal recordings. Such ability to not only classify the normality of a heart sound but to also provide the rationale behind such decision is important in evidence-based medicine. We also showed how two types of crowd-based heart sound analysis could be combined to come to a final heart sound classification. Finally, we identified potential filtering methods to remove poor performers in the training round of the binary heart sound classification task and illustrated how classification performance either stayed constant or increased when used in our hybrid framework. This result can be beneficial when dealing with finite resources.

In conclusion, this thesis has demonstrated that such a hybrid human-machine framework is a viable method for the accurate classification of heart sounds and motivates the continued research into such frameworks for bioacoustic signal analysis.

5.0.2 Future Work

Through the development and analysis of our framework, we have identified potential future areas to explore in the space of hybrid human-machine frameworks for signal analysis.

The CinC dataset [50] contains a large number of heart sound recordings, including normal and abnormal recordings from various heart sound conditions. However, there are still many other types of abnormal heart sounds (specifically in the class of murmurs) that are not included in this dataset (or are not classified specifically). Such murmurs include: aortic regurgitation, mitral stenosis and patent ductus arteriosus. This motivates the expansion of the dataset such that machine classifiers, crowdsourcing methodologies, and hybrid frameworks like the one we developed, can be continually evaluated and updated on a wider scope of heart sound abnormalities.

A similar challenge to the classification of heart sounds is the classification of lung sounds. As both heart sounds and lung sounds are bioacoustic signals, our framework can be directly applied to this problem. However, to our knowledge, there does not exist an open-source, standardized, rigorously evaluated database of lung sounds, similar to the CinC dataset [50] of heart sounds. The R.A.L.E. repository [32] is a small dataset (relative to the CinC dataset) that exists for lung sounds, however it is only available through purchase. The lack of an accessible dataset for research motivates its curation, in addition to a challenge to develop open-source classification algorithms, in a similar fashion to PhysioNet/CinC [50, 15].

In addition, we can also consider how this framework extends to multi-class classification problems. There are other certainty-based query strategies in active learning (margin sampling, Shannon entropy), that apply to multi-class problems, but reduce to uncertainty sampling in the binary case [74]. Such strategies could easily be substituted into the framework in the case of a multi-class classification problem.

We also ask how this framework performs in other signal classification problems, especially to those that contain features indicative of different stages. This would allow us to evaluate how our classification by proxy method generalizes to other scenarios. An example of such a classification problem includes sleep staging, where a feature like the sleep spindle is indicative of a given stage of sleep [89].

Finally, when it comes to studying human performance, a future direction could look at how we improve the training of non-experts (crowd workers or even medical students) to become better at heart sound analysis? Such studies could be centered around designing different training protocols to look at how the length of training or the information provided effects performance and the ability to learn.

References

[1] Cell-slider. https://2nznub4x5d61ra4q12fyu67t-wpengine.netdna-ssl.com/wp-content/uploads/2012/10/Cell-Slider.jpg, 10 2012.

[2] Figure 1 is empowering healthcare professionals to connect, cooperate and collaborate via social media. https://www.investinontario.com/spotlights/figure-1-empowering-healthcare-professionals-connect-cooperate-and-collaborate-social, 1 2017.

[3] Y. M. Akay, M. Akay, W. Welkowitz, and J. Kostis. Noninvasive detection of coronary artery disease. *IEEE Engineering in Medicine and Biology Magazine*, 13(5):761–764, Nov 1994.

[4] Samit Ari, Koushik Hembram, and Goutam Saha. Detection of cardiac abnormality from pcg signal using lms based least square svm classifier. *Expert Systems with Applications*, 37(12):8019 – 8026, 2010.

[5] L. D. Avendaño-Valencia, J. I. Godino-Llorente, M. Blanco-Velasco, and G. Castellanos-Dominguez. Feature extraction from parametric time–frequency representations for heart murmur detection. *Annals of Biomedical Engineering*, 38(8):2716–2732, Aug 2010.

[6] P. M. Bentley, P. M. Grant, and J. T. E. McDonnell. Time-frequency and time-scale techniques for the classification of native and bioprosthetic heart valve sounds. *IEEE Transactions on Biomedical Engineering*, 45(1):125–128, Jan 1998.

[7] Sanjay R. Bhatikar, Curt DeGroff, and Roop L. Mahajan. A classifier based on the artificial neural network approach for cardiologic auscultation in pediatrics. *Artificial Intelligence in Medicine*, 33(3):251 – 260, 2005.

[8] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998.

[9] Ignacio J Diaz Bobillo. A tensor approach to heart sound classification. In *Computing in Cardiology Conference (CinC), 2016*, pages 629–632. IEEE, 2016.

[10] J. Christopher Brady, C. Andrea Villanti, L. Jennifer Pearson, R. Thomas Kirchner, P. Omesh Gupta, and P. Chirag Shah. Rapid grading of fundus photographs for diabetic retinopathy using crowdsourcing. *J Med Internet Res*, 16(10):e233, Oct 2014.

[11] Mark Cartwright, Ayanna Seals, Justin Salamon, Alex Williams, Stefanie Mikloska, Duncan MacConnell, Edith Law, Juan P. Bello, and Oded Nov. Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW):29:1–29:21, December 2017.

[12] Sunita Chauhan, Ping Wang, Chu Sing Lim, and V. Anantharaman. A computer-aided mfcc-based hmm system for automatic auscultation. *Computers in Biology and Medicine*, 38(2):221 – 233, 2008.

[13] Justin Cheng and Michael S Bernstein. Flock: Hybrid crowd-machine learning classifiers. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 600–611. ACM, 2015.

[14] G. D. Clifford, C. Liu, B. Moody, D. Springer, I. Silva, Q. Li, and R. G. Mark. Classification of normal/abnormal heart sound recordings: The physionet/computing in cardiology challenge 2016. In *2016 Computing in Cardiology Conference (CinC)*, pages 609–612, Sept 2016.

[15] Gari D Clifford, Chengyu Liu, Benjamin Moody, Jose Millet, Samuel Schmidt, Qiao Li, Ikaro Silva, and Roger G Mark. Recent advances in heart sound analysis. *Physiological Measurement*, 38(8):E10, 2017.

[16] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, and Foldit Players. Predicting protein structures with a multiplayer online game. *Nature*, 466:756, aug 2010.

[17] J.S. Coviello. *Auscultation Skills: Breath & Heart Sounds*. Wolters Kluwer Health, 2013.

[18] J. P. de Vos and M. M. Blanckenberg. Automated pediatric cardiac auscultation. *IEEE Transactions on Biomedical Engineering*, 54(2):244–252, Feb 2007.

[19] Abdelghani Djebbari and Fethi Reguig. Detection of the valvular split within the second heart sound using the reassigned smoothed pseudo wignerville distribution. 12:37, 04 2013.

[20] Francisco J. Candido dos Reis, Stuart Lynn, H. Raza Ali, Diana Eccles, Andrew Hanby, Elena Provenzano, Carlos Caldas, William J. Howat, Leigh-Anne McDuffus, Bin Liu, Frances Daley, Penny Coulson, Rupesh J. Vyas, Leslie M. Harris, Joanna M. Owens, Amy F.M. Carton, Janette P. McQuillan, Andy M. Paterson, Zohra Hirji, Sarah K. Christie, Amber R. Holmes, Marjanka K. Schmidt, Montserrat Garcia-Closas, Douglas F. Easton, Manjeet K. Bolla, Qin Wang, Javier Benitez, Roger L. Milne, Arto Mannermaa, Fergus Couch, Peter Devilee, Robert A.E.M. Tollenaar, Caroline Seynaeve, Angela Cox, Simon S. Cross, Fiona M. Blows, Joyce Sanders, Renate de Groot, Jonine Figueroa, Mark Sherman, Maartje Hooning, Hermann Brenner, Bernd Holleczek, Christa Stegmaier, Chris Lintott, and Paul D.P. Pharoah. Crowdsourcing the general public for large scale molecular pathology studies in cancer. *EBioMedicine*, 2(7):681 – 689, 2015.

[21] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 1013–1022, New York, NY, USA, 2012. ACM.

[22] Peter Dulačka, Mária Bieliková, et al. Validation of music metadata via game with a purpose. In *Proceedings of the 8th International Conference on Semantic Systems*, pages 177–180. ACM, 2012.

[23] David S. Gerbarg, Angelo Taranta, Mario Spagnuolo, and John J. Hofler. Computer analysis of phonocardiograms. *Progress in Cardiovascular Diseases*, 5(4):393 – 405, 1963.

[24] Arash Gharehbaghi, Inger Ekman, Per Ask, Eva Nylander, and Birgitta Janerot-Sjoberg. Assessment of aortic valve stenosis severity using intelligent phonocardiography. *International Journal of Cardiology*, 198:58 – 60, 2015.

[25] M. . Goda and P. Hajas. Morphological determination of pathological pcg signals by time and frequency domain analysis. In *2016 Computing in Cardiology Conference (CinC)*, pages 1133–1136, Sept 2016.

[26] Benjamin M. Good and Andrew I. Su. Crowdsourcing for bioinformatics. *Bioinformatics*, 29(16):1925–1933, 2013.

[27] I. Grzegorczyk, M. Soliski, M. epek, A. Perka, J. Rosiski, J. Rymko, K. Stpie, and J. Gieratowski. Pcg classification using a neural network approach. In *2016 Computing in Cardiology Conference (CinC)*, pages 1129–1132, Sept 2016.

[28] D. Hakkani-Tur, G. Tur, M. Rahim, and G. Riccardi. Unsupervised and active learning in automatic speech recognition for call classification. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–429–32 vol.1, May 2004.

[29] Dilek Hakkani-Tür, Giuseppe Riccardi, and Allen Gorin. Active learning for automatic speech recognition. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 4, pages IV–3904. IEEE, 2002.

[30] M. N. Homsi, N. Medina, M. Hernandez, N. Quintero, G. Perpian, A. Quintana, and P. Warrick. Automatic heart sound recording classification using a nested set of ensemble algorithms. In *2016 Computing in Cardiology Conference (CinC)*, pages 817–820, Sept 2016.

[31] David W Hosmer and Stanley Lemeshow. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics: Theory and Methods*, 9(10):1043–1069, 1980.

[32] PixSoft Inc. R.A.L.E.® repository. http://www.rale.ca/.

[33] A. J. Joshi, F. Porikli, and N. P. Papanikolopoulos. Scalable active learning for multiclass image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2259–2273, Nov 2012.

[34] Alexander Kawrykow, Gary Roumanis, Alfred Kam, Daniel Kwak, Clarence Leung, Chu Wu, Eleyine Zarour, Phylo players, Luis Sarmenta, Mathieu Blanchette, and Jrme Waldisphl. Phylo: A citizen science approach for improving multiple sequence alignment. *PLOS ONE*, 7(3):1–9, 03 2012.

[35] Edmund Kay and Anurag Agarwal. Dropconnected neural network trained with diverse features for classifying heart sounds. In *Computing in Cardiology Conference (CinC), 2016*, pages 617–620. IEEE, 2016.

[36] Firas Khatib, Frank DiMaio, Foldit Contenders Group, Foldit Void Crushers Group, Seth Cooper, Maciej Kazmierczyk, Miroslaw Gilski, Szymon Krzywda, Helena Zabranska, Iva Pichova, James Thompson, Zoran Popović, Mariusz Jaskolski, and David Baker. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature Structural &Amp; Molecular Biology*, 18:1175, sep 2011.

[37] Youngmoo E Kim, Erik M Schmidt, and Lloyd Emelle. Moodswings: A collaborative game for music mood label collection. In *Ismir*, volume 2008, pages 231–236, 2008.

[38] Hong-Kwang Jeff Kuo and Vaibhava Goel. Active learning with minimum expected error for spoken language understanding. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[39] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1675–1684. ACM, 2016.

[40] Paul Lamere. Social tagging and music information retrieval. *Journal of new music research*, 37(2):101–114, 2008.

[41] P. Langley and A. Murray. Abnormal heart sounds detected from short duration unsegmented phonocardiograms by wavelet entropy. In *2016 Computing in Cardiology Conference (CinC)*, pages 545–548, Sept 2016.

[42] Edith Law, Burr Settles, and Tom Mitchell. Learning to tag from open vocabulary labels. In José Luis Balcázar, Francesco Bonchi, Aristides Gionis, and Michèle Sebag, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 211–226, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[43] Edith Law and Luis von Ahn. Human computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 5(3):1–121, 2011.

[44] Edith LM Law, Luis Von Ahn, Roger B Dannenberg, and Mike Crawford. Tagatune: A game for music and sound annotation. In *ISMIR*, volume 3, page 2, 2007.

[45] Vernon Lawhern, David Slayback, Dongrui Wu, and Brent J Lance. Efficient labeling of eeg signal artifacts using active learning. In *Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on*, pages 3217–3222. IEEE, 2015.

[46] Daniel TL Lee and Akio Yamamoto. Wavelet analysis: theory and applications. 1994.

[47] David D Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of the eleventh international conference on machine learning*, pages 148–156, 1994.

[48] David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12. Springer-Verlag New York, Inc., 1994.

[49] Huiying Liang and I Nartimo. A feature extraction algorithm based on wavelet packet decomposition for heart sound signals. In *Time-Frequency and Time-Scale Analysis, 1998. Proceedings of the IEEE-SP International Symposium on*, pages 93–96. IEEE, 1998.

[50] Chengyu Liu, David Springer, Qiao Li, Benjamin Moody, Ricardo Abad Juan, Francisco J Chorro, Francisco Castells, Jos Millet Roig, Ikaro Silva, Alistair E W Johnson, Zeeshan Syed, Samuel E Schmidt, Chrysa D Papadaniil, Leontios Hadjileontiadis, Hosein Naseri, Ali Moukadem, Alain Dieterlen, Christian Brandt, Hong Tang, Maryam Samieinasab, Mohammad Reza Samieinasab, Reza Sameni, Roger G Mark, and Gari D Clifford. An open access database for the evaluation of heart sound algorithms. *Physiological Measurement*, 37(12):2181, 2016.

[51] Angel Miguel Luengo-Oroz, Asier Arranz, and John Frean. Crowdsourcing malaria parasite quantification: An online game for analyzing images of infected thick blood smears. *J Med Internet Res*, 14(6):e167, Nov 2012.

[52] Madhero. Phonocardiograms from normal and abnormal heart sounds. https://upload.wikimedia.org/wikipedia/commons/4/4a/Phonocardiograms_from_normal_and_abnormal_heart_sounds.png, 3 2010.

[53] Ilias Maglogiannis, Euripidis Loukis, Elias Zafiropoulos, and Antonis Stasis. Support vectors machine-based identification of heart valve diseases using heart sounds. *Computer Methods and Programs in Biomedicine*, 95(1):47 – 61, 2009.

[54] Vykintas Maknickas and Algirdas Maknickas. Recognition of normalabnormal phonocardiographic signals using deep convolutional neural networks and mel-frequency spectral coefficients. *Physiological Measurement*, 38(8):1671, 2017.

[55] Michael I Mandel and Daniel PW Ellis. A web-based game for collecting music metadata. *Journal of New Music Research*, 37(2):151–165, 2008.

[56] Sam Mavandadi, Stoyan Dimitrov, Steve Feng, Frank Yu, Uzair Sikora, Oguzhan Yaglidere, Swati Padmanabhan, Karin Nielsen, and Aydogan Ozcan. Distributed medical image analysis and diagnosis through crowd-sourced games: A malaria case study. *PLOS ONE*, 7(5):1–8, 05 2012.

[57] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Metrics for polyphonic sound event detection. *Applied Sciences*, 6(6), 2016.

[58] N.D Ashley Meyer, A. Christopher Longhurst, and Hardeep Singh. Crowdsourcing diagnosis for patients with undiagnosed illnesses: An evaluation of crowdmed. *J Med Internet Res*, 18(1):e12, Jan 2016.

[59] An Thanh Nguyen, Byron C Wallace, and Matthew Lease. Combining crowd and expert labels using decision theoretic active learning. In *Third AAAI Conference on Human Computation and Crowdsourcing*, 2015.

[60] Tan B. Nguyen, Shijun Wang, Vishal Anugu, Natalie Rose, Matthew McKenna, Nicholas Petrick, Joseph E. Burns, and Ronald M. Summers. Distributed human intelligence for colonic polyp classification in computer-aided detection for ct colonography. *Radiology*, 262(3):824–833, 2012. PMID: 22274839.

[61] T. Nilanon, J. Yao, J. Hao, S. Purushotham, and Y. Liu. Normal / abnormal heart sound recordings classification using convolutional neural network. In *2016 Computing in Cardiology Conference (CinC)*, pages 585–588, Sept 2016.

[62] Chin Kim On, Paulraj M Pandiyan, Sazali Yaacob, and Azali Saudi. Mel-frequency cepstral coefficient analysis in speech recognition. In *Computing & Informatics, 2006. ICOCI'06. International Conference on*, pages 1–5. IEEE, 2006.

[63] World Health Organization. Cardiovascular diseases (cvds). http://www.who.int/mediacentre/factsheets/fs317/en/, 5 2017.

[64] J. J. G. Ortiz, C. P. Phoo, and J. Wiens. Heart sound classification based on temporal alignment techniques. In *2016 Computing in Cardiology Conference (CinC)*, pages 589–592, Sept 2016.

[65] Gerhard Osius and Dieter Rojek. Normal Goodness-of-Fit Tests for Multinomial Models with Large Degrees of Freedom. *Journal of the American Statistical Association*, 87(140):1145–1152, 1992.

[66] Shivnarayan Patidar, Ram Bilas Pachori, and Niranjan Garg. Automatic diagnosis of septal defects based on tunable-q wavelet transform of cardiac sound signals. *Expert Systems with Applications*, 42(7):3315 – 3326, 2015.

[67] C. Potes, S. Parvaneh, A. Rahman, and B. Conroy. Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds. In *2016 Computing in Cardiology Conference (CinC)*, pages 621–624, Sept 2016.

[68] A. F. Quiceno-Manrique, J. I. Godino-Llorente, M. Blanco-Velasco, and G. Castellanos-Dominguez. Selection of dynamic features based on time–frequency representations for heart murmur detection from phonocardiographic signals. *Annals of Biomedical Engineering*, 38(1):118–137, Jan 2010.

[69] Benjamin L. Ranard, Yoonhee P. Ha, Zachary F. Meisel, David A. Asch, Shawndra S. Hill, Lance B. Becker, Anne K. Seymour, and Raina M. Merchant. Crowdsourcing—harnessing the masses to advance health and medicine, a systematic review. *Journal of General Internal Medicine*, 29(1):187–203, Jan 2014.

[70] J. Rogstadius, V. Kostakos, A. Kittur, B. Smus, J. Laredo, and M. Vukovic. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In *International AAAI Conference on Web and Social Media*, ICWSM, pages 321–328, 2011. This work is supported in part by an IBM Open Collabora- tion Award; by the Portuguese Foundation for Science and Technology (FCT) grant CMU-PT/SE/0028/2008 (Web Security and Privacy); and by NSF grants OCI-0943148 and IIS-0968484.

[71] J. Rubin, R. Abreu, A. Ganguli, S. Nelaturi, I. Matei, and K. Sricharan. Classifying heart sound recordings using deep convolutional neural networks and mel-frequency cepstral coefficients. In *2016 Computing in Cardiology Conference (CinC)*, pages 813–816, Sept 2016.

[72] Rdvan Saraolu. Hidden markov model-based classification of heart valve disease with pca for dimension reduction. *Engineering Applications of Artificial Intelligence*, 25(7):1523 – 1528, 2012. Advanced issues in Artificial Intelligence and Pattern Recognition for Intelligent Surveillance System in Smart Home Environment.

[73] Amir A. Sepehri, Joel Hancq, Thierry Dutoit, Arash Gharehbaghi, Armen Kocharian, and A. Kiani. Computerized screening of children congenital heart diseases. *Computer Methods and Programs in Biomedicine*, 92(2):186 – 192, 2008.

[74] Burr Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11, 2010.

[75] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294. ACM, 1992.

[76] Lior Shamir, Carol Yerby, Robert Simpson, Alexander M von Benda-Beckmann, Peter Tyack, Filipa Samarra, Patrick Miller, and John Wallin. Classification of large acoustic datasets using machine learning and crowdsourcing: Application to whale calls. *The Journal of the Acoustical Society of America*, 135(2):953–962, 2014.

[77] Max H. Sims, Jeffrey Bigham, Henry Kautz, and Marc W. Halterman. Crowdsourcing medical expertise in near real time. *Journal of Hospital Medicine*, 9(7):451–456, 2014.

[78] N. E. Singh-Miller and N. Singh-Miller. Using spectral acoustic features to identify abnormal heart sounds. In *2016 Computing in Cardiology Conference (CinC)*, pages 557–560, Sept 2016.

[79] D. B. Springer, L. Tarassenko, and G. D. Clifford. Logistic regression-hsmm-based heart sound segmentation. *IEEE Transactions on Biomedical Engineering*, 63(4):822–832, April 2016.

[80] Thrse A Stukel. Generalized Logistic Models. *Journal of the American Statistical Association*, 83(402):426–431, 1988.

[81] Chong Sun, Narasimhan Rampalli, Frank Yang, and AnHai Doan. Chimera: Large-scale classification using machine learning, rules, and crowdsourcing. *Proceedings of the VLDB Endowment*, 7(13):1529–1540, 2014.

[82] A.J. Taylor. *Learning Cardiac Auscultation: From Essentials to Expert Clinical Interpretation*. Springer London, 2015.

[83] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

[84] Gokhan Tur, Dilek Hakkani-Tür, and Robert E Schapire. Combining active and semi-supervised learning for spoken language understanding. *Speech Communication*, 45(2):171–186, 2005.

[85] Douglas Turnbull, Ruoran Liu, Luke Barrington, and Gert RG Lanckriet. A game-based approach for collecting semantic annotations of music. In *ISMIR*, volume 7, pages 535–538, 2007.

[86] Harun Uğuz. Adaptive neuro-fuzzy inference system for diagnosis of the heart valve diseases using wavelet transform with entropy. *Neural Computing and Applications*, 21(7):1617–1628, Oct 2012.

[87] S. Vernekar, S. Nair, D. Vijaysenan, and R. Ranjan. A novel approach for classification of normal/abnormal phonocardiogram recordings using temporal signal analysis and machine learning. In *2016 Computing in Cardiology Conference (CinC)*, pages 1141–1144, Sept 2016.

[88] Ping Wang, Chu Sing Lim, Sunita Chauhan, Jong Yong A. Foo, and Venkataraman Anantharaman. Phonocardiographic signal analysis method using a modified hidden markov model. *Annals of Biomedical Engineering*, 35(3):367–374, Mar 2007.

[89] Simon C Warby, Sabrina L Wendt, Peter Welinder, Emil G S Munk, Oscar Carrillo, Helge B D Sorensen, Poul Jennum, Paul E Peppard, Pietro Perona, and Emmanuel Mignot. Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods. *Nature Methods*, 11:385, feb 2014.

[90] Jenna Wiens and John V Guttag. Active learning applied to patient-adaptive heartbeat classification. In *Advances in neural information processing systems*, pages 2442–2450, 2010.

[91] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

[92] Bishan Yang, Jian-Tao Sun, Tengjiao Wang, and Zheng Chen. Effective multi-label active learning for text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 917–926. ACM, 2009.

[93] Dong Yu, Balakrishnan Varadarajan, Li Deng, and Alex Acero. Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion. *Computer Speech & Language*, 24(3):433–444, 2010.

[94] M. Zabihi, A. B. Rad, S. Kiranyaz, M. Gabbouj, and A. K. Katsaggelos. Heart sound anomaly and quality detection using ensemble of neural networks without segmentation. In *2016 Computing in Cardiology Conference (CinC)*, pages 613–616, Sept 2016.

[95] Shan Zhang, Aditya Vempaty, Susan E Parks, and Pramod K Varshney. On classification of environmental acoustic data using crowds. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 5880–5884. IEEE, 2017.

[96] Yineng Zheng, Xingming Guo, and Xiaorong Ding. A novel hybrid energy fraction and entropy-based approach for systolic heart murmurs identification. *Expert Systems with Applications*, 42(5):2710 – 2721, 2015.

[97] Xiaojin Zhu, Bryan R Gibson, and Timothy T Rogers. Co-training as a human collaboration policy. In *AAAI*, 2011.

APPENDICES

Appendix A

Pre-Study Questionnaire

The following questionnaire was given to crowd workers before completing the heart sound classification task. This information was collected in order to identify whether a relationship exists between a worker's background and their performance in the study, which could be used to filter or weight the contributions of future workers.

1. Do you have a medical background?
    - Yes
    - No
2. Are you affiliated with anyone involved in medicine?
    - Yes
    - No
3. What is your occupation/job/career?
4. What is your highest level of education?
    - High School
    - Undergraduate Education
    - Masters Degree

- Doctorate Degree
- Other

5. What is your field of education (eg. Computer Science, English, Geography, etc.)?
6. What is your gender?
    - Male
    - Female
    - Prefer not to say
7. What is your age group?
    - 18-24
    - 25-34
    - 35-44
    - 45-54
    - 55-64
    - 65-74
    - 75+
8. Do you play a musical instrument?
    - Yes
    - No
9. What type of headphones are you wearing?
    - Earbuds
    - Over the ear
    - On the ear
    - Not sure
    - Other
10. Are you wearing noise cancelling headphones?

- Yes
- No
- Not Sure

Appendix B

Feature Space

The following section contains a description of common features used in the modelling of phonocardiogram signals, as presented in Section 2.4.

B.1 Time

Common features calculated in the time domain are often based upon the length of certain phases within a signal. In the context of heart sounds, we have the following phases from which we can collect this information: S1, Systole, S2, Diastole, and the total length of a given heartbeat. Utilizing this information, common features calculated include the average length and standard deviation of each phase, or the ratio of given phase lengths within a cardiac cycle [27, 25, 64]. Another common feature includes calculating the mean and standard deviation of the absolute amplitude of a signal within a given phase (or the ratio between phases) [27].

B.2 Frequency

The Fourier transform is a widely used tool in signal processing, which converts a signal from the time domain to the frequency domain. [25]. With this frequency information,

summary metrics can be calculated on different frequency ranges across the spectrum. Similar to time-based features, these summary metrics can also be collected on a phase-level basis [27, 67, 25].

The use of Mel-Frequency Cepstral Coefficients (MFCC) is a method of describing the audio spectrum of a recording based on how the human ear resolves frequency information [62]. It is based upon the mel-scale, which is a non-linear scale that groups together certain frequencies which aim to "mimic the human ear in terms of the manner with which frequencies are sensed and resolved" [62].

B.3 Time-Frequency

When a Fourier transform is applied to a signal, all time localization information is removed and only the frequency information remains [86]. This is adequate when dealing with stationary signals, whose frequencies do not change over time, but results in a loss of information when dealing with non-stationary signals (such as PCG) [86]. Although frequency-based methods are still used in the processing of non-stationary signals, time-frequency based methods are important as they preserve both types of information.

One example of a time-frequency based method is the short-time Fourier transform, which provides time and frequency information over the course of short window periods [86]. However, the granularity of which time information is preserved in this technique is dependent on the window size [86].

As a result, wavelet analysis was developed to provide better time and frequency localization. Similar to Fourier analysis, wavelet analysis is used to represent a signal in terms of a set of basis functions (in this case, wavelets) [46]. Unlike Fourier analysis, wavelets are localized in time and frequency (whereas the functions used in Fourier analysis are continuous). As a result, wavelet analysis allows for the preservation of time and frequency information [46]. A signal can be decomposed into a set of wavelet functions, from which the coefficients can be used as features or other methods can be used to further process the decomposed signal [25].