MechanicalHeart: A Human-Machine Framework for the Classification of Phonocardiograms

WILLIAM CALLAGHAN, University of Waterloo, Canada
JOSLIN GOH, University of Waterloo, Canada
DR. MICHAEL MOHAREB, William Osler Health System, Canada
DR. ANDREW LIM, Sunnybrook Health Sciences Centre, Canada
EDITH LAW, University of Waterloo, Canada

Listening to heart sounds is an important first step in evaluating the cardiovascular system and is important in the early detection of cardiovascular disease. 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 is motivated by the past success in utilizing human computation to solve problems in medicine and the use of human-machine frameworks in other domains. We describe the methods that decide when and how to escalate the analysis of heart sounds to different resources and incorporate their decision into a final classification. Our framework was tested with a combination of machine classifiers and crowd workers from Amazon's Mechanical Turk. The results indicate a hybrid approach achieves greater performance than a baseline classifier alone, utilizing less expert resources while achieving similar performance, compared to a framework without the crowd.

CCS Concepts: • **Information systems** → **Crowdsourcing**; • **Human-centered computing**; • **Applied computing** → *Health informatics*;

Additional Key Words and Phrases: Phonocardiogram; Heart Sounds; Classification; Annotation

ACM Reference Format:
William Callaghan, Joslin Goh, Dr. Michael Mohareb, Dr. Andrew Lim, and Edith Law. 2018. MechanicalHeart: A Human-Machine Framework for the Classification of Phonocardiograms. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 28 (November 2018), 17 pages. https://doi.org/10.1145/3274297

1 INTRODUCTION

The analysis of heart sounds is an important step in the evaluation of the cardiovascular system and may reveal pathological cardiac conditions such as arrhythmias and heart failure [25, 44]. 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 [25]. Automated methods have been developed to analyze heart sounds, although they have often been trained and/or evaluated on unrealistic clean data [25].

Authors' addresses: William Callaghan, University of Waterloo, Waterloo, Ontario, Canada, wrcallag@uwaterloo.ca; Joslin Goh, University of Waterloo, Waterloo, Ontario, Canada, jtcgoh@uwaterloo.ca; Dr. Michael Mohareb, William Osler Health System, Etobicoke, Ontario, Canada, michael.mohareb@williamoslerhs.ca; Dr. Andrew Lim, Sunnybrook Health Sciences Centre, Toronto, Ontario, Canada, andrew.lim@utoronto.ca; Edith Law, University of Waterloo, Canada, edith.law@uwaterloo.ca.

2573-0142/2018/11-ART28 $15.00
https://doi.org/10.1145/3274297

Crowdsourcing is an approach that enlists the help of humans to solve challenging problems that are currently unsolved or difficult for automated approaches to perform effectively [20, 37]. On crowdsourcing platforms (e.g. Amazon Mechanical Turk [1]), people, henceforth referred to as workers, often perform short microtasks such as image labelling and classification, audio transcription, or surveys, in exchange for small amounts of compensation [11, 37]. In medicine, crowdsourcing analysis of medical data is in its infancy, however there are a number of studies that have already shown its promise [36]. Such previous works have focused on using the crowd to detect abnormalities important in the early detection of disease, including parasites in red blood cell smears [26, 29] and colorectal polyps in computed tomographic (CT) images [33].

The crowd has also been used in conjunction with machine learning algorithms, as a tool to collect annotations and/or labels for data, or give feedback about instances in which a learning algorithm is uncertain. In this paper, we present a hybrid human-machine framework for binary heart sound classification in addition to exploring how crowd workers perform in heart sound analysis tasks. The framework decides how to escalate the analysis of heart sound recordings to different resources and incorporate their analyses into a final classification. It comes to a decision based on who has analyzed the heart sound (machine, crowd, expert), their level of uncertainty, and a threshold of acceptable uncertainty. Our results show that the hybrid framework achieves greater performance than a baseline classifier alone, utilizing less expert resources, while achieving similar performance when compared to a framework that does not use the crowd.

2 RELATED WORK

2.1 Crowdsourcing Medical Data Analysis

In diagnostic medicine, Mavandadi *et al* (2012) [29] and Luengo-Oroz *et al* (2012) [26] crowdsourced the analysis of red blood cell smears to assist in the identification of malarial infection, and achieved expert-level performance [13]. Diagnostic decisions made by non-expert participants in Mavandadi *et al* (2012) [29] were within 1.25% of those made by a medical professional. Similarly, in Luengo-Oroz *et al* (2012) [26], non-experts achieved a parasite counting accuracy of over 99%. In CellSlider [10], 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. 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 [33], indicating that "minimally trained ... workers could perform expert-level task[s] rapidly and with high quality," [13]. Such rapid, high quality work has also been demonstrated in the categorization of diabetic optic fundus photos as normal or abnormal, with early detection being important for the prevention of vision loss [4]. Finally, in Warby *et al* (2014) [48], 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 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 [40], CrowdMed [31], 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 Figure 1, there is even the ability to page an expert in the field, which sends an alert to a verified specialist [1].

[1] https://www.mturk.com
[2] http://figure1.com

Although there has been success in medical crowdsourcing, there is a valid concern behind having non-medical professionals provide medical analysis [29, 36]. However, Mavandadi *et al* (2012) [29] argues that crowdsourcing can still be used to relay the data to a medical professional, who can then make the final diagnosis [36]. 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 [36]. 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.

2.2 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 music information retrieval (MIR) researchers in solving music classification and recommendation problems [19]. Utilizing crowdsourcing for tag generation of audio has shown to be a valid approach for collecting accurate and meaningful labels [21, 47]. Such work includes MajorMiner [28], The Listen Game [47] and Tag-A-Tune [21], which utilized players and their level of agreement to provide high quality, descriptive tags for music. Similar work has been done in the area of acoustic scene classification, where crowdsourcing has been used to derive the classification of animals by comparing their calls [39, 53].

2.3 Heart Sound Analysis

In clinical practice, the physical examination of a patient is one of the first steps in evaluating their cardiovascular system [25]. 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 and heart failure [25, 44]. 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 [25].

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 [8, 25]. Listening for specific heart sounds can give an indication of the heart's health [25]. An audio or graphical recording of these vibrations is referred to as a heart sound recording or phonocardiogram (PCG) [25].

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 [8]. Immediately following S1 and lasting until S2 is Systole, and from S2 until the following S1 is known as Diastole [8]. These four stages make up the cardiac cycle [8]. Other sounds may be present such as the third (S3) and fourth (S4) heart sounds, clicks, snaps, or heart murmurs [8, 25]. A heart murmur refers to an abnormal heart sound with "an underlying physiologic pathology,"[8] often caused by turbulent blood flow due to abnormal valves.

Automated heart sound classification has been widely studied since the original work by Gerbarg *et al* (1963) [12] and have been historically grouped into four categories: artificial neural networks (ANN), support vector machines (SVM), Hidden Markov Models (HMM), or clustering-based classification [25]. However, Liu *et al* (2016) [25] 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, Liu *et al* (2016) [25] 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 [7].

2.4 Human-Machine 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 [6]. The system Chimera utilizes the crowd to evaluate classification models of product labels and descriptions [43]. Cases deemed incorrect or ambiguous are forwarded to in-house analysts, who develop rules and update models to address these issues [43]. Other frameworks exist that directly embed an oracle into the learning process, and are termed active learning frameworks [6]. These frameworks allow the learning algorithm to "choose the data from which it learns," [38]. 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 [38]. It then learns from these results and uses them to determine which instances to query next [38]. Active learning has been applied to problems in areas such as biosignal classification [22, 49], speech recognition [14, 15, 46, 52], image classification [17] and text classification [45, 51].

Nguyen *et al* (2015) [32] also explores choosing labels from the crowd or expert, but the focus of this work differs. They use a greatest expected loss reduction strategy to querying a single instance in each iteration, which they then use to update their classification model. Our focus is on iteratively choosing the instance and oracle based on the uncertainty of the pre-trained classifier. We apply this hybrid approach to the domain of heart sound classification which, to our knowledge, has not been explored.

2.4.1 Co-Training for Human Collaboration. One of the crowd-based classification strategies evaluated in our framework, called Crowd Ensemble (Section 3.5), is inspired by the work of Zhu *et al* (2011) [54]. They created a human collaboration policy for a categorical learning task based on the initial co-training algorithm proposed by Blum and Mitchell (1998) [2]. Co-training 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," [2]. In the categorical learning 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 [54]. Alice can then see Bob's labels (from her own view) and decides whether to accept/believe Bob's labels and vice-versa [54]. Data labeled by either individual is removed from the set of unlabeled instances, and the process continues until the unlabeled data is exhuasted [54].

3 STUDY DESIGN

In this section, we describe the methods used to facilitate binary heart sound classification ("Normal" or "Abnormal") in both humans and algorithms individually, as well as in a combined framework.

3.1 Research Questions

Our study aims to answer the following research questions:

Q1 Can the crowd reliably classify heart sounds as Normal or Abnormal?
Q2 How do we combine crowdsourcing with automated methods to analyze heart sounds?
Q3 How do we determine when to involve an expert?

3.2 Heart Sound Dataset

A total of thirty-five recordings were sampled from the PhysioNet/Computing in Cardiology (CinC) Challenge 2016 public heart sound database, published by Liu *et al* (2016) [25]. These recordings covered four different heart conditions: Normal, Aortic Stenosis (AS), Mitral Regurgitation (MR), and Mitral Valve Prolapse (MVP). Thirty recordings were used for evaluation while five recordings were saved for training of workers. In the evaluation set, 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 [25] 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 cardiologist was recruited to provide this information.

3.3 Crowdsourcing Heart Sound Analysis

To study the ability of workers to analyze heart sounds, we conducted a study on Amazon's Mechanical Turk.

We created two separate Human Intelligence Tasks (HITs) which contained the following tasks:

Normal/Abnormal Task Workers were required to classify the overall heart sound as normal or abnormal.

Murmur Detection Task Workers were required to outline murmurs within a recording (if they exist) or indicate that no murmurs exist in the recording.

Each task contained ten recordings for evaluation 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 ten recordings. Restrictions were in place to ensure that each worker only completes the study once. We also required that workers complete a hearing test, watch a training video and participate in a training round. Workers could complete both tasks if desired, but must complete the Normal/Abnormal classification task first.

The hearing test ensured that workers were listening over adequate headphones or speakers, and were not hard of hearing [3]. 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. The training video and exercise allowed workers to familiarize themselves with the interface and task(s), and provided them with ground truth feedback on their performance.

3.4 Task Interface

We extended the web-based audio annotator tool (Figure 1) initially developed by Cartwright *et al* (2017) [5] to be more appropriate for bio-acoustic signal analysis. This included the following additions:

(1) *Zoom and Pan/Scroll*: Heart sound analysis occurs at a much finer time resolution, so workers needed the ability to work at this level of granularity.

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

Fig. 1. Crowd annotation interface for classifying and annotating heart sounds.

(2) *Support for Contextual Information*: Heart sound segmentation into the stages of the cardiac cycle aids in the subsequent detection and classification of pathological events [8, 25, 41]. Therefore, it was important that this 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 [25]. We represented the segmentation as a set of labels aligned at the bottom of the label stack as to not interfere with the audio waveform and any annotations a worker created.
(3) *Rules to Guide Work*: By utilizing existing knowledge about the location of murmurs, we defined rules to help guide workers in the murmur detection task. This included limiting workers to one annotation per heart beat and not allowing annotations to cross beat boundaries.
(4) Example Viewer: The example viewer is a modified version of the annotator interface that allows workers to reference various heart sound examples and compare them against one another.

3.5 Crowd Classification Methods

From the two heart sound analysis tasks, we can derive three methods of heart sound classification. The first, called Classification, is a simple majority voting method based on the classifications given for each instance in the Normal/Abnormal task. The second method is called Detection, and utilizes the information from the murmur detection task to come to a decision about the normality of a given heart sound. If a worker defines the presence of murmur(s) in 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.

The third method utilizes information from the first two methods and is inspired by the work of Zhu *et al* (2011) [54]. In this method, called Crowd Ensemble, we look at which of the two previous methods are more confident in its answer, by using the principles of uncertainty sampling (see Section 3.7), and use this label as the classification for the instance.

3.6 Machine Classifiers

The machine classifiers used were pre-trained, open-sourced entries from the PhysioNet/CinC Challenge [7]. We selected four entries to use in the evaluation of our framework, with the restriction

that these models produced some probabilistic output for the predicted target label. The models were selected based on the top scoring entries from the challenge, as listed on the PhysioNet website[4].

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 (i.e. the challenge training set) [25]. Therefore, of the 3000+ records in the training set, these methods may have been trained using some of the records in our subset.

3.7 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," [23, 24, 38]. In our framework, we define uncertain instances i as:

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

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. In binary classification, t is most often 0.5, but can be adjusted to other values, as was done in Potes *et al* (2016) [35] and Bobillo (2016) [3] to 0.4 and 0.225 respectively.

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:

$$\%Agreement_{Abnormal} = \frac{\text{\# Abnormal Votes}}{\text{\# Normal Votes} + \text{\# Abnormal Votes}} \tag{2}$$

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:

$$FinalClass = \underset{c \in \{\text{Normal,Abnormal}\}}{\arg\max} \ (\max(|P_{Classifier}(x = \text{c}) - t_{Classifier}|, |P_{Crowd}(x = \text{c}) - t_{Crowd}|)) \tag{3}$$

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 (e.g. $|0.2 - 0.5|$) as the decision difference.

3.7.1 Expert Involvement. Just as we impose a certainty threshold on the machine classifier, we can also impose one on the instances classified by the crowd. In this case, 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 the expert returns the correct (ground truth) answer and that this represents the *FinalClass*.

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

4 ANALYSIS METHODS

To ensure the quality of the data, data cleaning was performed by removing spammers from the database. We consider spammers to be workers that did not play the audio recording at all while completing the task. We also filtered out participants that did not complete the whole experiment (10 audio clips).

4.1 Classification Performance

To evaluate the overall classification performance of the crowd-based methods, machine classifiers, and 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:

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}, \quad F1 = \frac{2PR}{P + R} \tag{4}$$

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

4.2 Murmur Detection Performance

The information from the murmur detection task was evaluated by how well the crowd performs at detecting murmurs in abnormal heart sound recordings. We define an aggregation strategy similar to that used in Cartwright *et al* (2017) [5]. We first divide each recording into non-overlapping, fixed-length (e.g. 100 ms) time frames. 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 [30]. Due to the small sample size and the lack of knowledge regarding the distribution of the F1 scores, we performed a Wilcoxon One-Sided Signed-Rank test [50] to test if the aggregate murmur detection F1 scores are significantly greater than 0.5 (random).

4.3 Evaluating Heart Sound Classification

To understand what type of heart sounds are difficult for machines and humans to classify, a logistic regression model was used. In this model, the dependent variable is a binary variable indicating whether the probability of a given instance being classified as abnormal by a given method (crowd or machine) is greater than that method's specified decision margin (t). The independent variables are the condition (AS, MR, MVP, Normal) and the method (crowd or machine-based).

4.4 Hybrid Framework Performance

It is inevitable that the window size, w, used in the hybrid framework affects the final F1-score. To understand the effect of window size, we evaluated the final F1-Score with varying window size for different combinations of machine classifiers and levels of human involvement. Such human involvement includes using one of the crowd-based classification methods and either the presence or absence of an expert. By comparing the final F1-score, a window size will be proposed.

4.4.1 Measuring Crowd Usefulness. To evaluate whether the crowd provides a useful contribution in the context of the hybrid framework (without experts), we can calculate the following metrics. Specifically, of the instances 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 proportions are calculated for each combination of machine classification and crowd classification method, averaged across all window sizes.

4.4.2 Crowd Effect on Expert Resources. The effect that the crowd has on alleviating expert involvement was evaluated by comparing two variations of the hybrid framework. The first is the original hybrid framework, with the addition of expert involvement as outlined in Section 3.7.1. The second is a framework with only the machine and the expert. In this case, if the machine classifier is uncertain in the classification of a given instance, the instance is sent directly to the expert instead of going through the crowd.

Analysis was done to compare the difference in the number of instances sent to the expert in the two variations to obtain a measure of the crowd impact on expert resources.

5 RESULTS

5.1 Crowd Performance

A total of 89 crowd workers completed the Normal/Abnormal classification task, and 67 completed the murmur detection task. The performance of the crowd classification and machine methods are presented in Table 1. The crowd performed well at binary heart sound classification, with the Classification method producing the best results among the crowd-based methods. The F1-Score from both the Detection and Crowd Ensemble classification methods are the same, however the instances in which each method classifies correctly slightly differ.

When evaluating murmur detection performance, the Wilcoxon One-Sided Signed-Rank test [50] 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 murmurs in abnormal recordings. An Analysis of Variance (ANOVA) was used to understand the effects of the disease condition on the ability for the aggregate to capture the murmurs present in these recordings, where the dependent variable is 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 ($p = 0.03$) in aggregate F1-Scores, however the other condition pairs, MR-AS ($p = 0.22$) and MVP-AS ($p = 0.53$) did not.

5.2 Difficulty of Heart Sound Classifications

Table 2 summarizes the result of the logistic model used to understand the effect of disease condition, crowd classification and machine methods on the ability to correctly classify abnormal instances. All two-factor interaction terms between the variables listed in the table were considered but none appeared to be significant, indicating lack of dependency between those variables. As a result, we chose the simpler model shown in Table 2, that can model the ability to correctly classify abnormal instances just as well as the model with two-factor interaction terms. Goodness-of-fit was validated by the Hosmer-Lemeshow [16] ($\chi^2(3, N = 208) = 1.49, p = 0.68$), Osius-Rojek [34] ($z = -0.0007, p = 1.00$) and Stukel [42] 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) [3]) 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[5] less

[5]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	Classification	0.87	0.87	0.87
	Detection†	0.86	0.80	0.83
	Crowd Ensemble	0.86	0.80	0.83
Machine	Bobillo (2016)	0.79	1.00	0.88
	Kay and Agarwal (2016)	0.82	0.93	0.88
	Maknickas and Maknickas (2017)	0.50	0.67	0.57
	Potes *et al* (2016)	0.67	0.93	0.78

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

Table 1. Base crowd and machine classifier performance

	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	***
Classification	-0.88	0.67	-1.31	
Detection	-1.04	0.69	-1.50	
Crowd Ensemble	-1.12	0.68	-1.64	
Kay and Agarwal (2016)	-0.43	0.66	-0.66	
Maknickas and Maknickas (2017)	0.21	0.65	0.33	
Potes *et al* (2016)	0.43	0.66	0.65	

Table 2. Logistic model to model the effect of condition and method on classifying a given clip as abnormal.

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.

5.3 Hybrid Framework Evaluation without Experts

The evaluation of the hybrid framework (Q2), utilizing different crowd and machine methods, is summarized in Figure 2. 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 Classification and Crowd Ensemble methods producing the same final classification results. The top performing combination of human and machine methods is the Classification (or Crowd Ensemble) with the classifier developed by Bobillo (2016) [3]. Even with this classifier having the greatest initial F1-Score ($F1 = 0.88$) when used independent of our framework, the use of our hybrid approach still leads to an increase in performance, with a F1-Score of 0.97 at $w = 0.25$ (in the Classification method). In addition, classifiers with lower initial F1-Scores achieve considerable gains in performance, as seen with the classifier by Maknickas and Maknickas (2017) [27] having a baseline F1-Score of 0.57 and a F1-Score of 0.80 at $w = 0.25$ when used with the Classification method.

As the the windowing parameter, w, changes, and more instances are sent to the crowd for analysis, the F1-Scores increase and plateau at around $w = 0.25$ for all four models (see Figure 2). As a result, this value may be appropriate for higher overall classification performance. To validate the change in F1-Scores, a three-way ANOVA was used. Condition on the effects of the different

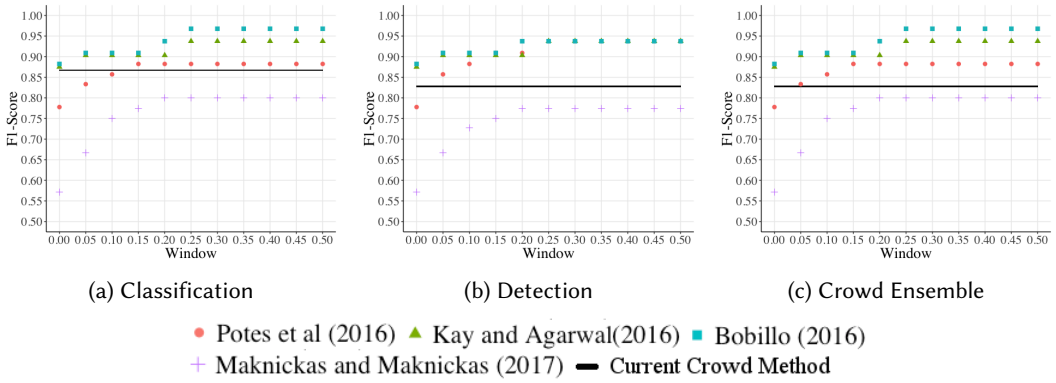

Fig. 2. Hybrid framework performance (without experts) using different crowd classification strategies

Classifier	Crowd Strategy	#Correct/#Query	#Used/#Query	#Correct/#Used
Bobillo (2016)	Classification	0.95	0.72	1.00
	Detection	0.77	0.51	1.00
	Crowd Ensemble	0.95	0.72	1.00
Kay and Agarwal (2016)	Classification	0.84	0.67	1.00
	Detection	0.80	0.62	1.00
	Crowd Ensemble	0.83	0.67	1.00
Maknickas and Maknickas (2017)	Classification	0.90	0.80	0.95
	Detection	0.82	0.76	0.94
	Crowd Ensemble	0.88	0.80	0.95
Potes *et Al* (2016)	Classification	0.85	0.80	0.86
	Detection	0.75	0.63	0.93
	Crowd Ensemble	0.84	0.80	0.86

Table 3. Summary averages of crowd query frequency and accuracy over all windows

crowd and machine classifiers, we found that the change in windowing parameter causes significant change in the F1-Scores ($F(1, 123) = 134.60, p < 0.001$), regardless of crowd and machine strategies.

When looking at the subsets of instances where the crowd is queried, we can see from Table 3 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, 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.

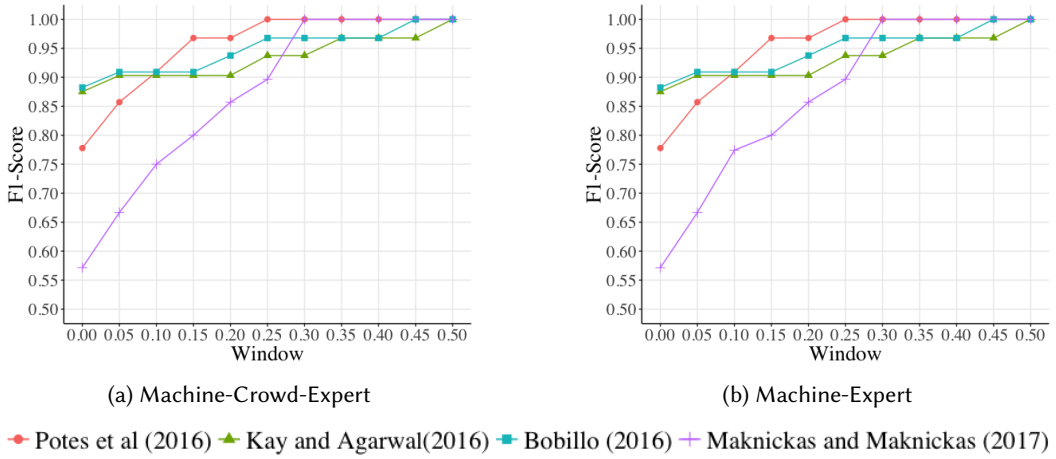

Fig. 3. Hybrid framework performance using a combination of machines, crowd and experts.

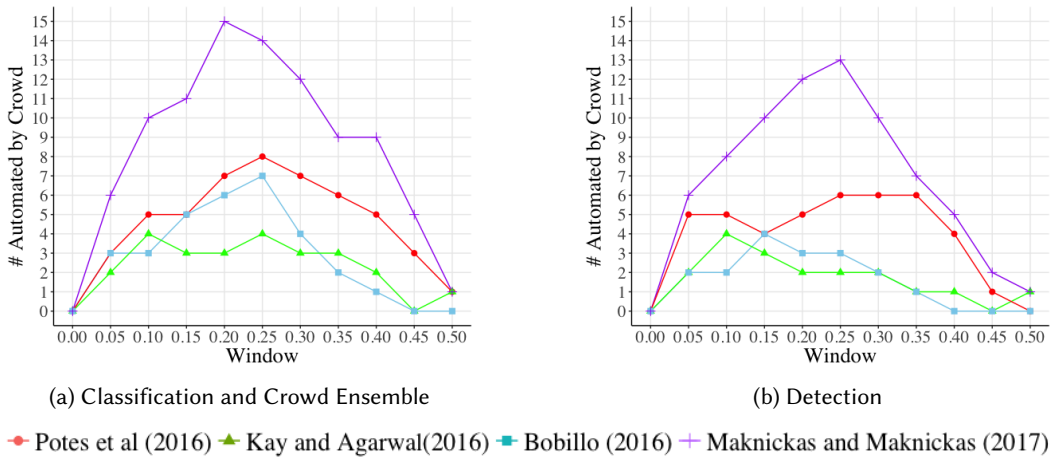

Fig. 4. # Instances (out of 30) automated by crowd at different windowing values.

5.4 Hybrid Framework Evaluation with Experts

The results from adding experts into the workflow are presented in Figure 3. Note the same F1-Scores are achieved regardless of the crowd classification strategy used. 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) [27] classifier at $w = 0.10$), however the level of expert involvement in the classification process is different.

Figure 4 shows the number of instances (out of 30) that by-pass expert resources. Specifically, it shows the difference in the number of instances sent to the expert in the two variations of the hybrid framework. A $w = 0$ indicates that no instances are sent to the crowd or expert (only the machine classifier is used). Alternatively, as we increase the value of w, we put more trust only in the expert. As previously mentioned, a windowing value of 0.25 provided the best classification

results in our framework. These results also show that it may be an appropriate value for saving expert resources.

In addition, an ANOVA was performed to understand how expert querying changes with varying windowing values and the difference in expert querying between the Machine-Crowd-Expert condition and the Machine-Expert condition. The proportion of queries sent to the expert differ significantly as w changes ($F(1,164) = 1333.81$, $p < 0.001$). In addition, the interaction between the type of machine classifiers and windowing parameter is statistically significant, $F(3,164) = 16.10$, $p < 0.001$, i.e. the effect of the windowing parameter on the proportion of queries sent to the expert changes with the type of machine classifiers. The proportion of queries sent to the expert also differ significantly with the different type of machine classifiers ($F(3, 164) = 88.87$, $p < 0.001$) and between Machine-Expert and Machine-Crowd-Expert conditions ($F(1, 164) = 84.3748$, $p < 0.001$). In fact, the model, which considered the variables and their corresponding interaction effects, suggests that 7% more queries are estimated to be sent to the expert in the Machine-Expert condition, $\beta = 0.07, t(164) = 2.53, p = 0.01$. Under the Machine-Expert condition, significantly more queries were sent to the expert, $F(3, 164) = 8.9803, p < 0.001$.

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.

When it comes to comparing the Classification method with Detection, the latter method resulted in a slightly lower F1-Score than the former 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 detecting murmurs in recordings they think to be abnormal. Such ability is important in evidence-based medicine, strengthening the initial argument made by Mavandadi *et al* (2012) [29] 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 Classification and Detection methods, the crowd was consistent in their performance regardless of the abnormality. Although the overall F1-Score was higher in the Classification method than the Detection 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 Classification method as a measure of normality and the corresponding Detection 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 [18]. 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. 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, interpretability from the machine or crowd is not necessarily needed.

When it comes to choosing a value for w, 0.25 is a proposed value based on the data used. Our dataset was costly to curate as the heart sounds were additionally annotated by a cardiologist. In order to find an optimal value, more data is required. However, this value can serve as a starting point for other applications. The takeaway is that different w values may produce different results, and we recommend others to evaluate different values within their respective domains.

6.0.1 Applications. Online communities such as Figure 1[6] 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 [1].

Our hybrid framework could be integrated into applications like Figure 1, or similarly CrowdMed [31], 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 Figure 1 is volunteer-based, platforms like CrowdMed [31] do compensate their "medical detectives" for their work on medical cases. 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.

6.0.2 Future Work. Some areas for future work include looking at how people (non-expert crowd workers or medical students) learn to analyze heart sounds and whether such interfaces can be used to better train non-experts in heart sound analysis. In addition, a study into how people utilize the information presented to them could be conducted. For example, do people rely more on the visual or audio information in a phonocardiogram? How does the information provided effect overall performance? Other questions include how does such a framework apply to other types of bioacoustic signal analysis, such as lung sound classification.

The goal of our paper is not to find the best aggregation method, but to demonstrate how humans and machines can work together to accomplish this complex medical annotation task. As such, we chose to use a simple majority voting scheme to aggregate crowd-generated classifications. This allows us to describe the average performance of human classifiers in heart sound analysis without any artificial alteration, e.g., dynamic weighing of worker contributions or other algorithmic augmentation.

Given more sophisticated aggregation methods (e.g. Dawid *et al* (1979) [9]), we expect two outcomes. First, the crowd performance may exceed algorithm performance; however, the cost of crowd labour is still substantial, warranting the use of a hybrid human-machine framework. Second, the hybrid performance of the crowd and machines may exceed that of the current results. A promising direction for future work is to incorporate more sophisticated aggregation algorithms into our hybrid framework to further boost performance.

[6] https://figure1.com/

7 CONCLUSION

In this work, we presented and evaluated a hybrid human-machine framework for binary heart sound classification in addition to exploring how crowd workers perform in heart sound analysis tasks. Our results indicated that the crowd performs well at heart sound analysis and that our hybrid framework achieved greater performance than a baseline classifier alone, utilizing less expert resources while achieving similar performance, when compared to a framework that does not use the crowd.

ACKNOWLEDGMENTS

This work is supported by NSERC CHRP under Grant No.: CHRP 478468-15 and CIHR CHRP under Grant No.: CPG-140200.

REFERENCES

[1] 2017. 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).

[2] Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*. ACM, 92–100.

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

[4] J. Christopher Brady, C. Andrea Villanti, L. Jennifer Pearson, R. Thomas Kirchner, P. Omesh Gupta, and P. Chirag Shah. 2014. Rapid Grading of Fundus Photographs for Diabetic Retinopathy Using Crowdsourcing. *J Med Internet Res* 16, 10 (30 Oct 2014), e233. DOI: http://dx.doi.org/10.2196/jmir.3807

[5] Mark Cartwright, Ayanna Seals, Justin Salamon, Alex Williams, Stefanie Mikloska, Duncan MacConnell, Edith Law, Juan P. Bello, and Oded Nov. 2017. Seeing Sound: Investigating the Effects of Visualizations and Complexity on Crowdsourced Audio Annotations. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 29 (Dec. 2017), 21 pages. DOI: http://dx.doi.org/10.1145/3134664

[6] Justin Cheng and Michael S Bernstein. 2015. Flock: Hybrid crowd-machine learning classifiers. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 600–611.

[7] G. D. Clifford, C. Liu, B. Moody, D. Springer, I. Silva, Q. Li, and R. G. Mark. 2016. Classification of normal/abnormal heart sound recordings: The PhysioNet/Computing in Cardiology Challenge 2016. In *2016 Computing in Cardiology Conference (CinC)*. 609–612. DOI: http://dx.doi.org/10.23919/CIC.2016.7868816

[8] J.S. Coviello. 2013. *Auscultation Skills: Breath & Heart Sounds* (5 ed.). Wolters Kluwer Health.

[9] P. Dawid, A. M. Skene, A. P. Dawidt, and A. M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics* (1979), 20–28.

[10] 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. 2015. Crowdsourcing the General Public for Large Scale Molecular Pathology Studies in Cancer. *EBioMedicine* 2, 7 (2015), 681 – 689. DOI: http://dx.doi.org/10.1016/j.ebiom.2015.05.009

[11] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the Crowd Yields Better Work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 1013–1022. DOI: http://dx.doi.org/10.1145/2145204.2145355

[12] David S. Gerbarg, Angelo Taranta, Mario Spagnuolo, and John J. Hofler. 1963. Computer analysis of phonocardiograms. *Progress in Cardiovascular Diseases* 5, 4 (1963), 393 – 405. DOI: http://dx.doi.org/10.1016/S0033-0620(63)80007-9

[13] Benjamin M. Good and Andrew I. Su. 2013. Crowdsourcing for bioinformatics. *Bioinformatics* 29, 16 (2013), 1925–1933. DOI: http://dx.doi.org/10.1093/bioinformatics/btt333

[14] Dilek Hakkani-Tür, Giuseppe Riccardi, and Allen Gorin. 2002. Active learning for automatic speech recognition. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 4. IEEE, IV–3904–IV–3907.

[15] D. Hakkani-Tur, G. Tur, M. Rahim, and G. Riccardi. 2004. Unsupervised and active learning in automatic speech recognition for call classification. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1. I–429. DOI: http://dx.doi.org/10.1109/ICASSP.2004.1326014

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

[17] A. J. Joshi, F. Porikli, and N. P. Papanikolopoulos. 2012. Scalable Active Learning for Multiclass Image Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 11 (Nov 2012), 2259–2273. DOI: http://dx.doi.org/10.1109/TPAMI.2012.21

[18] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. 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*. ACM, 1675–1684.

[19] Paul Lamere. 2008. Social tagging and music information retrieval. *Journal of New Music Research* 37, 2 (2008), 101–114.

[20] Edith Law and Luis von Ahn. 2011. Human Computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 5, 3 (2011), 1–121. DOI: http://dx.doi.org/10.2200/S00371ED1V01Y201107AIM013 arXiv:https://doi.org/10.2200/S00371ED1V01Y201107AIM013

[21] Edith LM Law, Luis Von Ahn, Roger B Dannenberg, and Mike Crawford. 2007. TagATune: A Game for Music and Sound Annotation.. In *ISMIR*, Vol. 3. 2.

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

[23] David D Lewis and Jason Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of the Eleventh International Conference on Machine Learning*. 148–156.

[24] David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3–12.

[25] Chengyu Liu, David Springer, Qiao Li, Benjamin Moody, Ricardo Abad Juan, Francisco J Chorro, Francisco Castells, José 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. 2016. An open access database for the evaluation of heart sound algorithms. *Physiological Measurement* 37, 12 (2016), 2181. http://stacks.iop.org/0967-3334/37/i=12/a=2181

[26] Angel Miguel Luengo-Oroz, Asier Arranz, and John Frean. 2012. Crowdsourcing Malaria Parasite Quantification: An Online Game for Analyzing Images of Infected Thick Blood Smears. *J Med Internet Res* 14, 6 (29 Nov 2012), e167. DOI: http://dx.doi.org/10.2196/jmir.2338

[27] Vykintas Maknickas and Algirdas Maknickas. 2017. Recognition of normal-abnormal phonocardiographic signals using deep convolutional neural networks and mel-frequency spectral coefficients. *Physiological Measurement* 38, 8 (2017), 1671. http://stacks.iop.org/0967-3334/38/i=8/a=1671

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

[29] Sam Mavandadi, Stoyan Dimitrov, Steve Feng, Frank Yu, Uzair Sikora, Oguzhan Yaglidere, Swati Padmanabhan, Karin Nielsen, and Aydogan Ozcan. 2012. Distributed Medical Image Analysis and Diagnosis through Crowd-Sourced Games: A Malaria Case Study. *PLOS ONE* 7, 5 (05 2012), 1–8. DOI: http://dx.doi.org/10.1371/journal.pone.0037245

[30] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2016. Metrics for Polyphonic Sound Event Detection. *Applied Sciences* 6, 6 (2016).

[31] N.D Ashley Meyer, A. Christopher Longhurst, and Hardeep Singh. 2016. Crowdsourcing Diagnosis for Patients With Undiagnosed Illnesses: An Evaluation of CrowdMed. *J Med Internet Res* 18, 1 (14 Jan 2016), e12. DOI: http://dx.doi.org/10.2196/jmir.4887

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

[33] Tan B. Nguyen, Shijun Wang, Vishal Anugu, Natalie Rose, Matthew McKenna, Nicholas Petrick, Joseph E. Burns, and Ronald M. Summers. 2012. Distributed Human Intelligence for Colonic Polyp Classification in Computer-aided Detection for CT Colonography. *Radiology* 262, 3 (2012), 824–833. DOI: http://dx.doi.org/10.1148/radiol.11110938 arXiv:https://doi.org/10.1148/radiol.11110938 PMID: 22274839.

[34] Gerhard Osius and Dieter Rojek. 1992. Normal Goodness-of-Fit Tests for Multinomial Models with Large Degrees of Freedom. *J. Amer. Statist. Assoc.* 87, 140 (1992), 1145–1152.

[35] C. Potes, S. Parvaneh, A. Rahman, and B. Conroy. 2016. Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds. In *2016 Computing in Cardiology Conference (CinC)*. 621–624. DOI: http://dx.doi.org/10.23919/CIC.2016.7868819

[36] 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. 2014. Crowdsourcing—Harnessing the Masses to Advance Health and Medicine, a Systematic Review. *Journal of General Internal Medicine* 29, 1 (01 Jan 2014), 187–203. DOI: http://dx.doi.org/10.1007/s11606-013-2536-8

[37] J. Rogstadius, V. Kostakos, A. Kittur, B. Smus, J. Laredo, and M. Vukovic. 2011. An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets. In *International AAAI Conference on Web and Social Media (ICWSM)*. 321–328. DOI: http://dx.doi.org/10.13140/RG.2.2.19170.94401

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

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

[40] Max H. Sims, Jeffrey Bigham, Henry Kautz, and Marc W. Halterman. 2014. Crowdsourcing medical expertise in near real time. *Journal of Hospital Medicine* 9, 7 (2014), 451–456. DOI: http://dx.doi.org/10.1002/jhm.2204

[41] D. B. Springer, L. Tarassenko, and G. D. Clifford. 2016. Logistic Regression-HSMM-Based Heart Sound Segmentation. *IEEE Transactions on Biomedical Engineering* 63, 4 (April 2016), 822–832. DOI: http://dx.doi.org/10.1109/TBME.2015.2475278

[42] Thérèse A Stukel. 1988. Generalized Logistic Models. *J. Amer. Statist. Assoc.* 83, 402 (1988), 426–431.

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

[44] A.J. Taylor. *Learning Cardiac Auscultation: From Essentials to Expert Clinical Interpretation.*

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

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

[47] Douglas Turnbull, Ruoran Liu, Luke Barrington, and Gert RG Lanckriet. 2007. A Game-Based Approach for Collecting Semantic Annotations of Music. In *ISMIR*, Vol. 7. 535–538.

[48] 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. 2014. Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods. *Nature Methods* 11 (Feb 2014), 385. DOI: http://dx.doi.org/10.1038/nmeth.2855

[49] Jenna Wiens and John V Guttag. 2010. Active learning applied to patient-adaptive heartbeat classification. In *Advances in Neural Information Processing Systems*. 2442–2450.

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

[51] Bishan Yang, Jian-Tao Sun, Tengjiao Wang, and Zheng Chen. 2009. Effective multi-label active learning for text classification. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 917–926.

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

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

[54] Xiaojin Zhu, Bryan Gibson, and Timothy Rogers. 2011. Co-Training as a Human Collaboration Policy. In *AAAI Conference on Artificial Intelligence*.

Received April 2018; revised July 2018; accepted September 2018