PLOS DIGITAL HEALTH

RESEARCH ARTICLE

A recurrent neural network and parallel hidden Markov model algorithm to segment and detect heart murmurs in phonocardiograms

Andrew McDonald *, **Mark J. F. Gales, Anurag Agarwal**

Department of Engineering, University of Cambridge, Cambridge, United Kingdom

* andrewmcdonald@cantab.net

Abstract

The detection of heart disease using a stethoscope requires significant skill and time, making it expensive and impractical for widespread screening in low-resource environments. Machine learning analysis of heart sound recordings can improve upon the accessibility and accuracy of diagnoses, but existing approaches require further validation on larger and more representative clinical datasets. For many previous algorithms, segmenting the signal into its individual sound components is a key first step. However, segmentation algorithms often struggle to find S1 or S2 sounds in the presence of strong murmurs or noise that significantly alter or mask the expected sound. Segmentation errors then propagate to the subsequent disease classifier steps. We propose a novel recurrent neural network and hidden semi-Markov model (HSMM) algorithm that can both segment the signal and detect a heart murmur, removing the need for a two-stage algorithm. This algorithm formed the 'CUED_Acoustics' entry to the 2022 George B. Moody PhysioNet challenge, where it won the first prize in both the challenge tasks. The algorithm's performance exceeded that of many end-to-end deep learning approaches that struggled to generalise to new test data. As our approach both segments the heart sound and detects a murmur, it can provide interpretable predictions for a clinician. The model also estimates the signal quality of the recording, which may be useful for a screening environment where non-experts are using a stethoscope. These properties make the algorithm a promising tool for screening of abnormal heart murmurs.

Author summary

The use of machine learning algorithms to detect heart disease from sound recordings has great potential to enable widespread and low-skill screening, improving early detection and treatment. The area has seen increasing interest in recent years, with many novel algorithms inspired by deep learning advancements in other fields. However, the size of heart sound datasets remains small, making deep learning models particularly susceptible to overfitting. In addition, the performance of these algorithms has rarely been directly

compared on unseen data. We describe a novel lightweight algorithm to detect and classify murmurs in heart sound recordings. This algorithm was the winning entry into the George B. Moody PhysioNet 2022 challenge, beating many complex deep-learning approaches. Our approach both detects and localises the murmur, providing an interpretable result for a clinician.

Introduction

Cardiovascular disease is the leading cause of mortality worldwide, with 18 million deaths every year [1]. However, compared to other serious diseases, awareness of many types of heart disease remains low and many are critically underdiagnosed [2]. In young people, rheumatic valvular heart disease is the most common cardiovascular disease, with an estimated global burden of 41 million cases, primarily in developing countries [3]. More than a quarter of patients with rheumatic disease present to a clinician at a late stage with heart failure [4]. There is a clear need for widespread early detection, to improve early treatment of the disease and prevent long-lasting morbidity [1].

Many heart diseases, such as rheumatic valve disease, cause structural changes in the heart that lead to abnormal sounds, such as heart murmurs [5]. The only tool currently available to detect these sounds in primary care is a stethoscope. Listening to the chest with a stethoscope (auscultation) is a quick and non-invasive test. However, auscultation proficiency varies widely amongst clinicians. The sensitivity of an experienced general practitioner in detecting valvular heart disease can be as low as 44% [6]. Enabling low-cost screening of the disease, especially in resource-constrained areas, will require a test that can be quickly and accurately performed by non-specialist clinicians.

Automated analysis of heart sound recordings (phonocardiograms) is a promising solution to improve the accuracy and accessibility of auscultation. A number of novel methods have been proposed in recent years, driven by an increased amount of open-access datasets and renewed interest in machine learning and AI [7]. However, many of these algorithms suffer from issues such as overfitting and poor generalisation to new data. The George B. Moody PhysioNet 2022 challenge [8] tasked participants to design novel algorithms to detect and classify heart murmurs in a new paediatric dataset, enabling an independent comparison of approaches that is more representative of a real-world clinical environment.

In this paper, we describe a novel algorithm that won the First Prize in both tasks in the challenge [8, 9]. The algorithm is inspired by takeaways from the earlier 2016 PhysioNet challenge on heart sound classification [7, 10] and, in contrast to many other approaches, does not use an end-to-end deep learning model. We also explore the results of the challenge and compare the algorithm's efficacy on the two distinct tasks.

Materials and methods

Dataset

Limitations of previous datasets.   An ongoing limitation of research into automated analysis of heart sound recordings (also known as phonocardiograms, PCGs) is the availability of high-quality labelled data. Whilst electronic stethoscopes have received regulatory approval and are available to purchase, the vast majority of stethoscopes in clinical practice are analogue. Therefore, heart sounds are rarely recorded and stored with patient records, unlike other cardiac tools such as electrocardiography (ECG) and echocardiography. The creation of

heart sound datasets therefore requires bespoke clinical research studies that are expensive and time-consuming, especially in resource-strained cardiology units.

Due to these constraints, open-access data has proved a valuable resource for advancing algorithm design. These datasets include the 2016 PhysioNet/Computing in Cardiology dataset [11] and the PASCAL Challenge dataset [12]. However, these datasets are limited by a lack of detailed information on the recording environment, murmur assessments, and patient outcomes. The 2016 PhysioNet challenge included a withheld test set to assess the generalisation of algorithms [7]. However, this test set was never made public after the challenge. Research using this dataset since 2016 has created new test sets by applying varying cross-validation and splitting strategies to the training set. These varied strategies have made direct comparisons between approaches challenging. The 2016 PhysioNet dataset also used recordings from multiple different devices, with its largest dataset using different devices to record abnormal and normal patients [11]. This unintentionally prompted machine learning algorithms to overfit to the dataset by learning to distinguish device characteristics rather than diagnostic sounds [10].

The CirCor DigiScope dataset.   In this study, we train and evaluate algorithms using the open-access paediatric CirCor DigiScope dataset [13–15], which was used in the George B. Moody PhysioNet Challenge 2022 [8]. To date, this dataset forms the largest open-access dataset for murmur detection and classification [13].

As described by Oliveira et al. [13], the dataset was gathered as a part of two screening programs in Brazil in 2014 and 2015. Approval for the study protocol was granted by the 5192-Complexo Hospitalar HUOC/PROCAPE Institutional Review Board under the request of the Real Hospital Português de Beneficiência em Pernambuco. Written consent was obtained for all participants, with parental consent where appropriate.

A total of 5268 phonocardiogram recordings were collected from 1452 patients. Some patients were recruited more than once, giving 1568 unique patient encounters. Recordings were made using the Littmann 3200 electronic stethoscope at up to four of the standard auscultation sites on the chest (aortic, pulmonic, tricuspid, mitral). All patients underwent a detailed cardiac investigation (radiography, electrocardiogram, echocardiogram). However, the released public dataset gives limited per-patient information on the individual diseases present, simply labelling each patient as 'normal' or 'abnormal'.

Each patient was labelled by an operator as 'murmur absent', 'murmur present', or 'unknown'. The unknown class was assigned when the signal quality of the patient's recordings was too poor for confident labelling. The chest locations where a murmur could be heard were labelled, allowing for per-recording labels to be assigned. The murmur timing, grade, and characteristics were also labelled. However, the dataset does not differentiate between pathological and physiological (innocent) murmurs [8]. Patients with an abnormal clinical outcome may therefore have an innocent murmur that is unrelated to their cardiac disease. Additionally, whilst a total of 305 patients have a murmur, with 304 systolic murmurs, only 10 have diastolic murmurs. This means the dataset is not suitable for designing or evaluating algorithms for diastolic murmur detection.

A subset (n = 110) of patients self-reported pregnancy during recruitment [13]. However, the age of these patients was not specified, making it unclear if they belong to the paediatric population [8].

George B. Moody PhysioNet challenge 2022.   Reyna et al. [8] provide a detailed overview of the 2022 PhysioNet challenge. 60% of the CirCor dataset was released publicly by the organisers as a training set (942 encounters), with the remaining 40% split between a validation (10%) and a final test set (30%) [8]. Participants were able to submit their algorithms to the validation set throughout the challenge to assess their performance, whilst the challenge organisers ran each entry only once on the test set to determine the final ranking [8]. The structure

and hyperparameters of the algorithms described in this work were developed using the challenge training set.

The challenge was split into two distinct tasks [8]. The first was murmur detection, where teams were asked to label each patient encounter as 'murmur absent', 'murmur present', or 'unknown', matching the labelling assigned by the expert annotator. A weighted accuracy metric was used to judge entries, assigning a weight of five to 'murmur present', three to 'unknown', and one to 'murmur absent'. This was designed to model the potential high cost of a false negative, and the importance of judging when a signal could be not be confidently diagnosed [8].

The second task was to predict the overall clinical outcome for a patient encounter as 'normal' or 'abnormal'. The challenge organisers designed a custom cost function for this task, modelling the potential costs of using the algorithm in a screening pathway [8]. A lower cost was better. Similar to the murmur task, the cost function placed a high cost on false negatives because this could lead to delayed or missed treatments.

Previous work

Whilst the CirCor dataset is a significant addition to open-access PCG data, it is still small compared to many machine learning domains where deep learning algorithms dominate performance. Feature extraction is a beneficial step to reduce the complexity of the audio data and hence the complexity required of subsequent classifiers, making it easier to train them to identify diagnostic features. A key conclusion of the 2016 PhysioNet challenge was that feature extraction was the most 'crucial and important' part of algorithms [7].

Segmentation of heart sound signals. The nature of heart sounds is well-understood from a clinical perspective [5]. A phonocardiogram is a non-stationary signal consisting of a generally periodic set of S1 (lub) and S2 (dub) sounds corresponding to closures of the atrio-ventricular and semilunar valves respectively [5]. Abnormal murmurs can appear in the systolic and diastolic regions of the signal, depending on the particular structural pathology of the patient's heart. Other abnormal sounds, such as S3 and S4, may also appear, giving distinctive rhythms in the signal. Due to this periodic structure, a common first step in classifiers is segmentation, where the individual heart sound states (S1, S2, systole, diastole) are labelled in time. This allows subsequent algorithm stages to focus on diagnostically relevant areas of the recording and apply ensemble averaging to segments to reduce noise.

In the 2016 challenge, a segmentation algorithm designed by Springer et al. [16] was provided to participants to aid their design [7]. This model used a hybrid structure, where a logistic regression provided observations for a hidden semi-Markov model (HSMM). It was considered state-of-the-art at the time [7]. However, the algorithm assumes a healthy heart cycle which makes it susceptible to errors when loud murmurs or noise overwhelm weaker S1 or S2 sounds [17].

Kay [17] observed this limitation and designed a segmentation algorithm that directly models the heart sound state. He calculates a series of band-pass-filtered homomorphic envelopes and power spectral densities (PSDs) to give a feature set that can distinguish murmurs from healthy sounds. He then replaces the logistic regression of Springer with a fully connected neural network, giving greater non-linear discrimination between states. Noting that the algorithm of Springer struggles to distinguish S1 from S2 and systole from diastole, Kay's algorithm predicts only three states: murmur, major heart sound, and silence. These neural network predictions are used as observations for two hidden semi-Markov models, one of which assumes the murmur state appears in systole and one that assumes a healthy heart sound with silence in systole. The most confident segmentation produced by the HSMM gives

a murmur diagnosis. We provide more detail on this HSMM approach in the Methods section of this paper.

One limitation of the approaches of Springer [16] and Kay [17] is that the discriminative classifiers providing observation probabilities predict on a per-frame basis and do not model dependencies between adjacent time frames. When inspecting a phonocardiogram, a skilled clinician will identify the S1 and S2 sounds by assessing the timing of the sounds. The models of Springer and Kay cannot do this as they view each frame independently, and so rely on the HSMM to provide this timing information. In previous work, recurrent neural networks (RNNs) have been applied to model this time dependency [18], offering improved predictive power compared to logistic regressions. However, the models deployed in previous work neglect to predict a murmur state and assume that only a healthy heart sound cycle is present.

Method

The murmur detection and segmentation algorithm we describe in the following sections is an improvement over Kay's work [17] and has been optimised for the 2022 PhysioNet challenge. The algorithm, as shown in Fig 1, consists of four distinct stages: (i) feature extraction, (ii) neural network prediction, (iii) segmentation, and (iv) final classification through a comparison of segmentation confidences. We then describe an algorithm to combine the output of murmur predictions from multiple auscultation positions to predict the patient's final clinical outcome.

We define a time-series heart sound recording with N samples as $r_{1:N}$. All the recordings in the challenge dataset were made using the Littmann 3200 electronic stethoscope, which has a fixed sampling frequency of 4000 Hz. In feature extraction, the time-series recording is converted into a time-frequency series with T time windows, $x_{1:T}$. Four parallel HSMM models ($\omega_1, \ldots, \omega_4$), are then applied to produce four distinct segmentations $q_{1:T}^{(\omega)}$. The confidence in these segmentations is then compared to produce a final segmentation and classification.

The following subsections describe each step of the algorithm in more detail. All computation was conducted in Python 3 using PyTorch 1.11, NumPy 1.21, and SciPy 1.7. The code to train these models and generate results is available on a public repository [19].

Feature extraction.   The amplitude of the PCG is first normalised by removing its mean and dividing by the resulting peak amplitude. The absolute amplitude of a PCG is an unreliable feature as it varies significantly depending on the application pressure of the stethoscope and the particular patient [20].
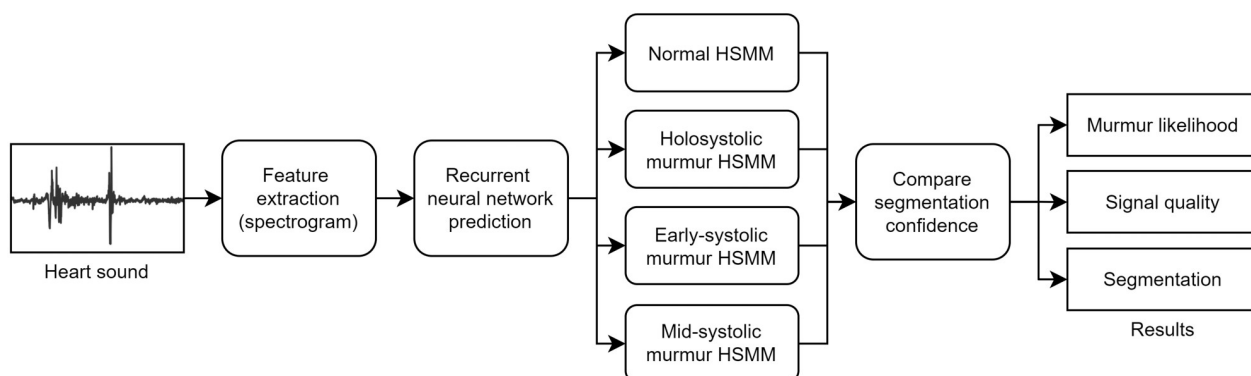

Fig 1. Parallel HSMM algorithm structure. A heart sound recording is transformed into log-spectrogram features, which are then input to an RNN. The RNN probabilities are then used to produce four segmentations using four parallel HSMMs. The confidence in these segmentations is then compared to give the most probable segmentation, murmur prediction, and a signal quality estimate.

The homomorphic envelope and PSD features of Kay [17] allow discrimination of energy in high and low frequencies but there is significant redundancy and extra computation in computing both the band-pass-filtered envelopes and the PSDs. Instead, in this work we compute a log-spectrogram using a Hann window of length 50 ms and step 20 ms. A larger window length enables a higher frequency resolution at the cost of a lower time resolution. The time duration of S1 and S2 sounds is approximately 100 ms, so a 50 ms window was chosen to enable precise identification of the major heart sounds whilst maintaining an effective frequency resolution of 20 Hz. A secondary advantage of this approach is that spectrograms provide an interpretable 2D visualisation of the time-frequency energy in the recording, as shown in Fig 2.

Although the Nyquist frequency of the recordings is 2000 Hz, we further limit the spectrogram to 0–800 Hz to remove higher frequencies that contain no heart sound information. This reduces the risk of the subsequent neural network stages learning to overfit to irrelevant high-frequency noise such as speech and background sounds.

As a key final step, each frequency row in the spectrogram is z-score normalised by subtracting its mean and dividing by its variance. Murmurs commonly contain much less time-frequency energy than S1 and S2 sounds, and this normalisation reduces the dynamic range [10].

Recurrent neural network. A recurrent neural network with knowledge of the timing of heart sounds should be able to discriminate S1, S2 and murmur sounds without relying on a subsequent HSMM. We therefore define a ground-truth segmentation $q_{1:T}$ of five distinct
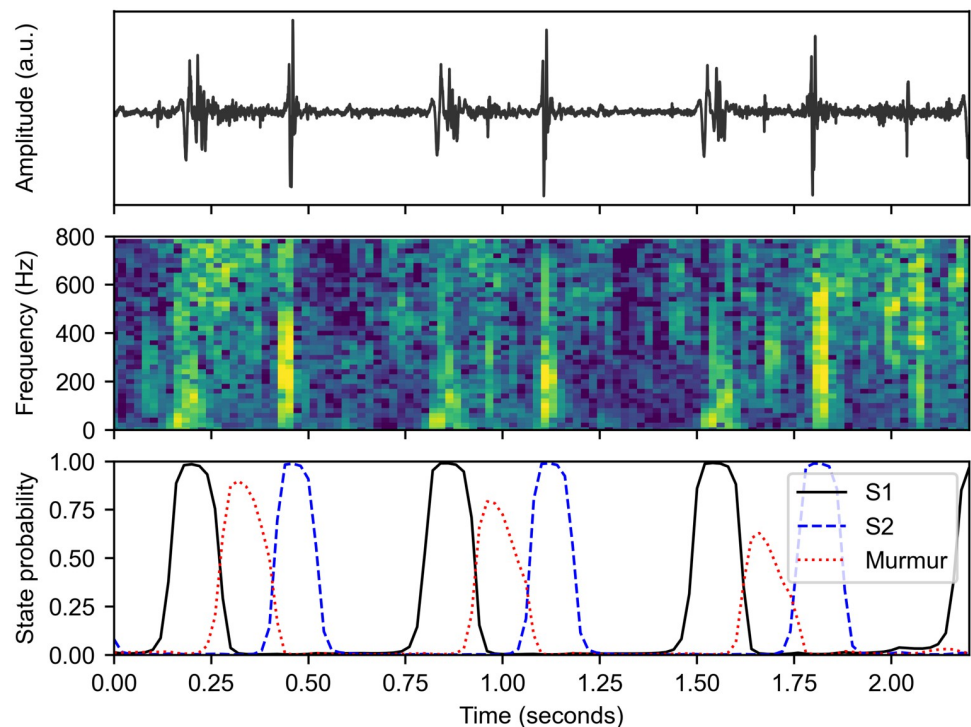
Fig 2. Neural network prediction of per-time heart sound categories. The heart sound recording (top) is transformed to a normalised spectrogram (middle) that is then passed to a RNN that predicts the state corresponding to each time window (bottom). The RNN is able to use the spectrogram to correctly distinguish between S1, S2, and the systolic murmur. The RNN also predicts systolic and diastolic states which are omitted from the bottom figure for clarity.

https://doi.org/10.1371/journal.pdig.0000436.g002

heart sound states $\xi_i \in$ {S1, S2, systole, diastole, murmur}. The challenge dataset includes segmented locations for the S1, S2, systole and diastole sounds but does not explicitly localise murmurs. However, the additional labels provided by the challenge include a prediction of the 'murmur timing' annotated by a clinician [13]. We use these labels to approximately annotate the location of the murmur in the ground-truth segmentation. If the recording is labelled to contain an early-systolic murmur, the first 50% of each systolic period is annotated as 'murmur'. If 'mid-systolic', the middle 50% is murmur. If 'holosystolic', the entire systole portion is annotated as a murmur.

We note that one limitation of this analysis is that diastolic murmurs cannot be detected, which may result in false negatives. Diastolic murmurs are much less prevalent than systolic murmurs in clinical practice, and only 5 patients in the training dataset have diastolic sounds. A future improvement, given more diastolic examples, would be to replicate this labelling approach for diastolic murmur signals.

Using this modified ground-truth segmentation, we train a bidirectional RNN with parameters θ to predict the state q_t at each time instance using the log-spectrogram features $x_{1:T}$, giving posterior probabilities $P(q_t = \xi_i|x_{1:T}, \theta)$. An example of the outputs of the RNN is shown in Fig 2. The RNN is confidently able to distinguish S1 and S2, verifying the five-state segmentation model.

The RNN structure and hyperparameters are optimised through cross-validation on the training dataset. The final model consists of a three-layer bidirectional RNN with Gated Recurrent Unit (GRU) [21] cells with a hidden layer size of 60. The concatenated forward and backward outputs of the RNN are passed to a two-layer fully connected neural network with Tanh activations and hidden sizes of 60 and 40. This fully connected network and a subsequent softmax layer reduce the RNN output to the five-dimensional output of the segmentation labels. Dropout with probability 0.1 is applied between the GRU and fully connected layers to reduce overfitting.

The model is trained using a cross-entropy loss with the Adam optimiser [22]. Some states in the segmentation (e.g. diastole) are much more prevalent than others (e.g. murmur), so to avoid models learning to favour one class the loss function is inversely weighted by the frequency of each class label in the overall dataset.

Parallel hidden semi-Markov models. Given posterior probabilities from the RNN, the simplest method to produce a segmentation would be to 'greedily' pick the state with the maximum probability at each time instance:

$$q_t = \max_{\xi_i} P(q_t = \xi_i|x_{1:T}, \theta) \tag{1}$$

A murmur could then be predicted if the greedy segmentation ever contains a murmur state. However, in practice, this approach is very prone to false positive murmur predictions because higher-frequency signal noise can cause occasional spurious predictions. Additionally, the heart is physiologically constrained to generate sounds in the order S1, systole, S2, diastole but the RNN is not similarly bound. This means that physically impossible state transitions are possible (e.g. S2 to systole) in the greedy segmentation.

To generate a globally optimum and physically valid segmentation we use the RNN probabilities as observations for hidden semi-Markov models, following the hybrid structure used in Springer [16] and Kay [17].

As described by Springer et al. [16], the HSMM is an extension to a traditional hidden Markov model that uses an explicit model for the duration of each state. This is particularly useful for physiological signals, such as phonocardiograms, where states have reasonably well-defined durations due to physical constraints.

The expected durations of states (particularly systole and diastole) significantly vary between patients due to the wide range of resting heart rates in the dataset. Springer et al. [16] therefore does not use one global model for the state durations, instead fine-tuning Gaussian distributions by scaling their means by an estimate of the heart rate. Their approach follows the work of Schmidt et al. [23] and estimates the heart rate by computing the autocorrelation of a smoothed homomorphic envelope of the heart sound signal. They then search for the highest peak in the autocorrelation in the 500 to 2000 ms range, corresponding to heart rates between 30 bpm and 120 bpm. This search range requires modification for paediatric use as approximately 20% of the dataset has a heart rate above 120 bpm.

In this work, we improve upon this estimate of the heart rate by noting that the posterior probabilities from the RNN are a filtered version of the original signal from which a period can be estimated. The homomorphic envelope Springer uses can therefore be replaced with the RNN posteriors to get an autocorrelation that is much smoother and less affected by noise spikes. See S1 Fig for more information.

Kay uses two HSMMs, one assuming a healthy sound and the other assuming a holosystolic murmur [17]. However, many of the murmurs in this dataset are early systolic [13].

In this work, we improve upon these assumptions by using four HSMMs that each make different assumptions about the underlying signal and hence generate a different segmentation. The state durations are shared between each HSMM but the observation probabilities and transition matrix differ:

ω_1 A normal signal with no murmur. A four-state segmentation model is used with the murmur posterior from the RNN discarded.

ω_2 A holosystolic murmur signal. A four-state segmentation model is used, where the murmur posterior replaces every systole posterior.

ω_3 An early-systolic murmur signal. A five-state segmentation model is used, with a transition matrix that forces the S1 state to transition to the murmur state and then to the systolic state.

ω_4 A mid-systolic murmur signal. A five-state segmentation model as above, but with a transition matrix that forces the S1 state to go to systole first.

A predicted segmentation $\hat{q}^{(\omega)}$ is calculated using each HSMM model above, giving four distinct interpretations of the signal. We then calculate a confidence measure of the segmentation $C^{(\omega)}$ by tracing the predicted segmentation path back through the RNN posterior probabilities:

$$C^{(\omega)} = \frac{1}{T} \sum_{t=1}^{T} P\big(q_t = \hat{q}_t^{(\omega)} | x_{1:T}, \theta\big) \tag{2}$$

The final model, $\hat{\omega}$, is chosen as the HSMM with the largest confidence:

$$\hat{\omega} = \arg \max_{\omega} \big(C^{(\omega)}\big) \tag{3}$$

This selects a final predicted segmentation of the signal, $\hat{q}_{1:T}^{(\hat{\omega})}$, and a final classification of the type and location of any systolic murmur. The confidence of the chosen model, $C^{(\hat{\omega})}$, can be used as a measure of the signal quality. The four-class model can easily be reduced to a binary murmur detector; if the chosen model is a murmur HSMM model ($\hat{\omega} \in \{\omega_2, \omega_3, \omega_4\}$) then a murmur is predicted. For the binary case, we can also calculate an overall confidence of the
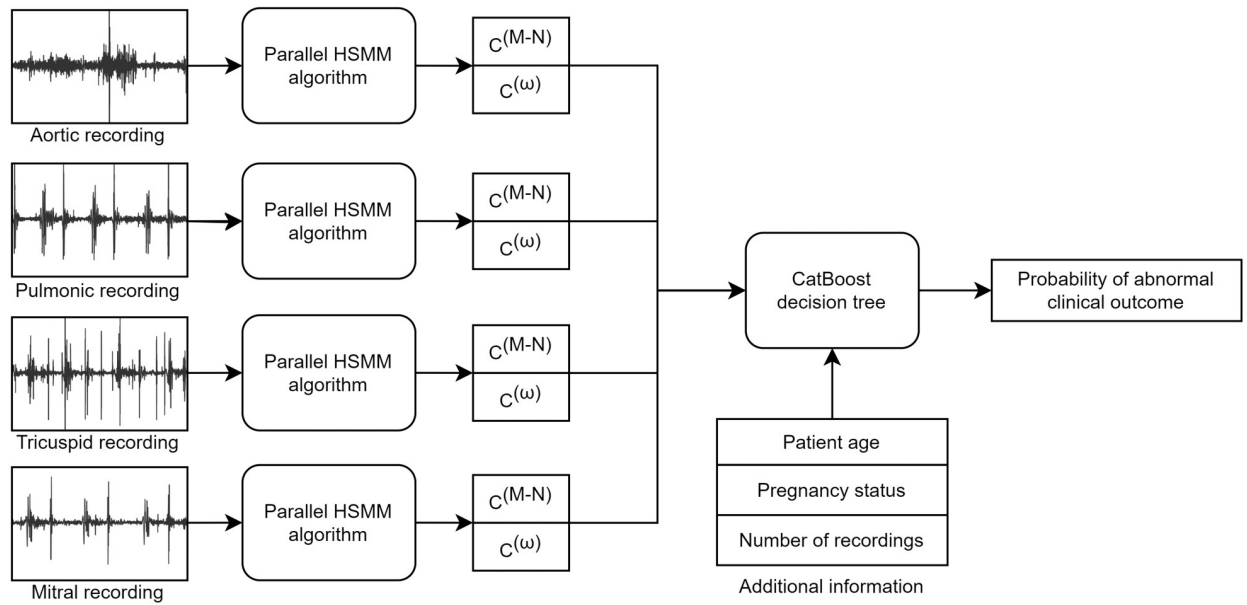
Fig 3. CatBoost algorithm structure to predict clinical outcome. The parallel HSMM algorithm described in Fig 1 is applied to each recording to generate murmur likelihood $C^{(M-N)}$ and signal quality $C^{(\hat{\omega})}$ metrics. These metrics are then combined with additional patient information and input to a CatBoost decision tree to predict the patient's final clinical outcome.

https://doi.org/10.1371/journal.pdig.0000436.g003

murmur, $C^{(M)}$, vs no murmur, $C^{(N)}$ decision:

$$C^{(M)} = \max(C^{(\omega_2)}, C^{(\omega_3)}, C^{(\omega_4)}) \tag{4}$$

$$C^{(N)} = C^{(\omega_1)} \tag{5}$$

The difference between these two confidences can then be used as a measure of the separability of the murmur and no murmur outcomes:

$$C^{(M-N)} = C^{(M)} - C^{(N)} \tag{6}$$

The above method produces a per-recording prediction of the presence of a murmur. To get a per-patient prediction in the format required by the challenge, we apply a simple common-sense criterion. If a murmur is detected in any of the recordings, the patient is predicted to have a murmur. If this is not true, the confidence of the chosen model $C^{(\hat{\omega})}$ is examined: if this is below a threshold (0.65), the patient is classified as 'unknown'. Otherwise, the patient is predicted to be 'no murmur'.

Prediction of clinical outcome. When predicting the overall clinical condition of a patient, a skilled clinician will auscultate each location of the chest and weigh the strength and characteristics of the sounds at each site [5]. They may also consider general patient biometrics such as age, and sex, along with the recorded patient history.

The CirCor dataset does not include detailed patient biomarkers or risk factors but does include basic biometrics such as age. To incorporate this information alongside the heart sound recordings, we apply a CatBoost gradient boosted decision tree [24] as shown in Fig 3. For each recording, we use the parallel HSMM method described above to calculate the murmur likelihood, $C^{(M-N)}$. We also compute the maximum confidence, $C^{(\hat{\omega})}$, so that the decision tree can reject poor-quality signals. In situations where a patient has multiple recordings from

the same chest location, the values are averaged before input to the CatBoost model. Added to the feature set are the patient's age, pregnancy status, and the overall number of recordings made. Age and pregnancy are expected to have some effect on the type and intensity of heart murmurs. The overall number of recordings is included to allow the algorithm to adjust the confidence of its predictions based on how many examples were taken.

The CatBoost model is trained with a cross-entropy loss and optimised through a five-fold cross-validation strategy alongside the RNN model. A class weight of 1.8 is used for the abnormal examples and 1 for the normal examples to prioritise sensitivity, because of the high potential cost of a false negative. The chosen decision tree has a depth of 9. The final threshold probability to decide an abnormal result (0.4738) is chosen to minimise the challenge cost function.

Results and discussion

We now explore the results of our algorithm on the CirCor dataset. A limitation of the dataset is that the test set has not yet been made publicly available, because of its use in the 2022 PhysioNet challenge. This means only a limited set of result metrics on the test set can be reported. For an in-depth exploration of the algorithm, we additionally report additional metrics evaluated through cross-validation on the public training set. As demonstrated later, the cross-validation results closely match those on the test set. We are therefore confident that the cross-validation results provide a representative example of the algorithm's performance.

Murmur detection

Cross-validated results. Fig 4 shows a plot of the HSMM confidence values for every recording in the training dataset. Using the HSMM confidences allows for a strong separation of murmur and normal signals, whilst producing an estimate of signal quality. A low threshold of $C^{(M-N)} = 0$ was chosen because the challenge weighted accuracy metric strongly penalised false negatives. However, a higher threshold could be picked for future applications, such as population screening, where a high specificity is essential to minimise false positive referrals. Fig 5 shows some examples of recordings analysed by the algorithm that would fall into different quadrants of the plot in Fig 4.

Table 1 shows a confusion matrix for the results of the algorithm on the training set. The corresponding micro-averaged accuracy is 0.771 and challenge weighted accuracy is 0.798. Table 2 shows the per-class performance of the algorithm in terms of sensitivity and positive predictive value (PPV). The macro-averaged F_1 score is 0.621.

Table 2 shows that the algorithm at its current operating point is 92.7% sensitive to murmurs. As expected, the sensitivity of the algorithm increases with the patient's reported murmur grade. For grade 1 (quiet) murmurs, 87.5% of cases are detected. This rises to 100% for grades 2 (moderate) and 3 (loud). The algorithm also has a high precision for the 'murmur absent' class, which would be important for use as a rule-out device where patients must be confidently rejected as having a murmur.

However, the performance at predicting the 'unknown' class is poor. Table 1 shows that 'unknown' patients (n = 68) were likely to be predicted as 'murmur' (n = 19) or 'no murmur' (n = 28). Only a single murmur patient was predicted to be 'unknown', but 39 'no murmur' patients were. The unknown class was used if the annotator was unable to confidently predict the presence of a murmur [13]. However, this extra label is subjective and highly dependent on how the recordings were listened to and annotated. The definition of a poor-quality signal from an algorithmic and human perspective is likely very different. It is possible that the algorithm can confidently predict cases that a human cannot, due to analysis of lower-frequency
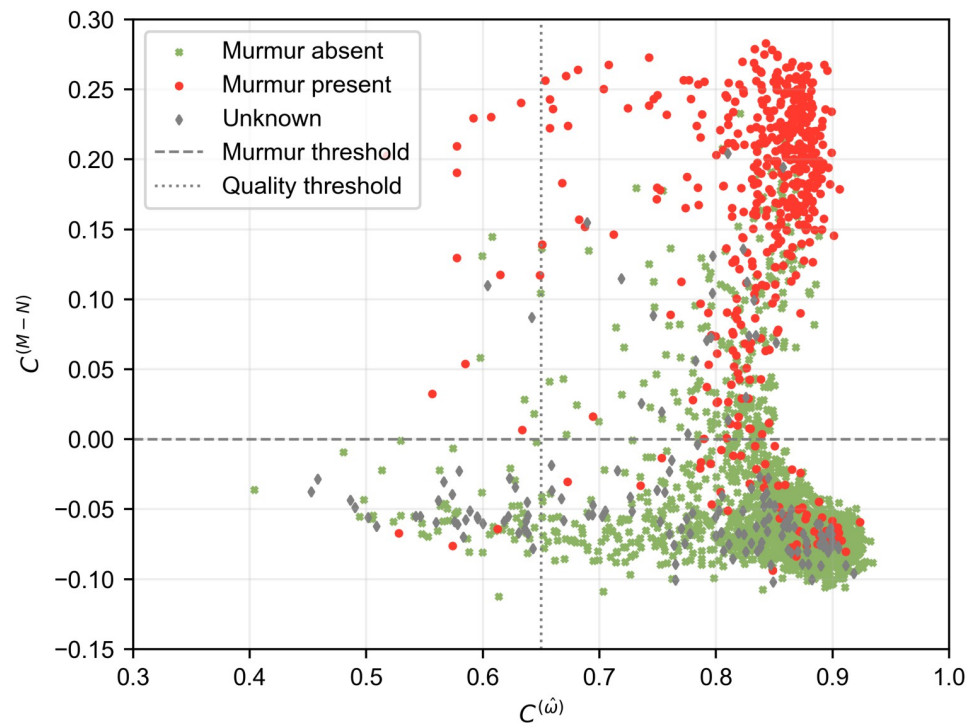
Fig 4. Separation of murmur, normal, and poor signal quality signals using HSMM confidences. The difference between the normal HSMM confidence, $C^{(N)}$, and the most confident murmur HSMM, $C^{(M)}$, is used to predict murmur likelihood $C^{(M-N)}$, whilst the most confident overall HSMM $C^{(\hat{\omega})}$ is used as an indication of signal quality. The horizontal and vertical lines show thresholds that have been chosen to separate the different classes, optimised for the challenge task. The marker type indicates the ground-truth labels.

https://doi.org/10.1371/journal.pdig.0000436.g004

inaudible energy or greater resilience to noise. Fig 5d shows an example of one of these recordings, which does contain some noise and is marked unknown by the clinician, but still has audible heart sounds and is confidently segmented by the algorithm and predicted as 'Murmur absent'.

If the 'unknown' patients are removed to create a new 'clean' dataset (n = 874), the murmur detection task reduces to binary classification and we can plot a receiver operating characteristic (ROC) curve, as shown in Fig 6. The murmur likelihood, $C^{(M-N)}$, can be used as the continuous test statistic to generate the curve. The model achieves an area under the ROC curve of 0.947, with an operating sensitivity of 92.7% and a specificity of 83.2%. However, the model performs well across a range of threshold and an operating point that prioritises specificity could be chosen if desired. Fig 7 shows a reliability diagram for this task, where the murmur likelihood is plotted against the relative frequency of murmurs. The approximately linear relationship shows that the murmur likelihood provides a calibrated estimate of the confidence in the decision.

Challenge test results. On the test set, the algorithm achieved a micro-averaged accuracy of 0.763, challenge weighted accuracy of 0.776, and a macro-averaged F_1 score of 0.623 [25]. These values closely match the results reported earlier on the cross-validated training set (0.771, 0.798, and 0.621 respectively), suggesting overfitting has not taken place. Per-class sensitivity and PPV metrics were not reported by the challenge organisers.

In the context of the PhysioNet challenge, the algorithm achieved the second-highest score, with an accuracy just 0.004 below the top score. The top-scoring team failed subsequent tests
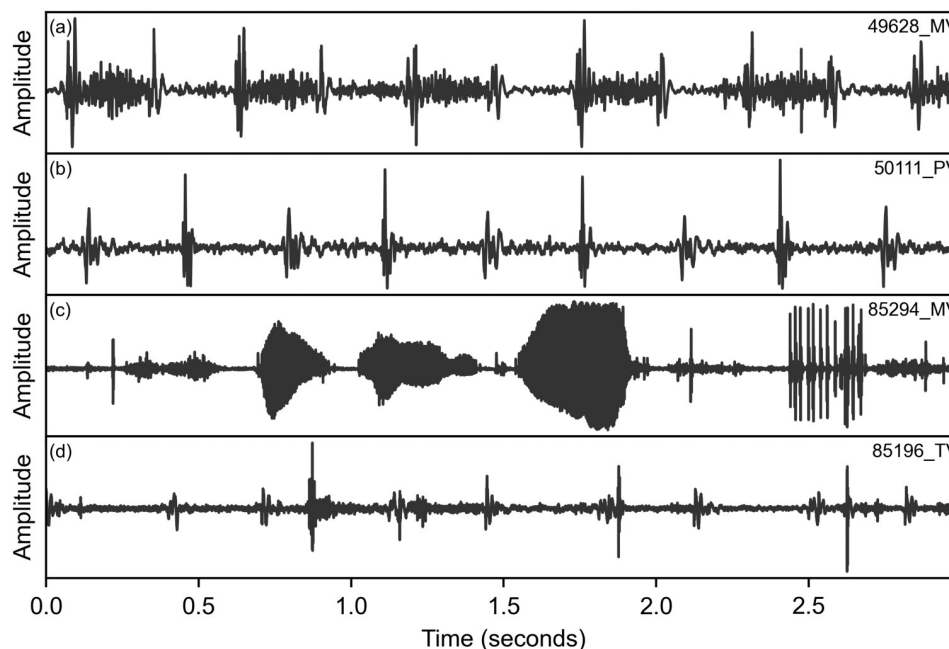
Fig 5. Four example recordings analysed by the algorithm. Three-second windows are shown for each recording, with their database ID shown in the top right. Recording (a) contains a strong systolic murmur that is confidently detected by the algorithm ($C^{(M-N)} = 0.25$). (b) is a healthy signal that is correctly identified ($C^{(M-N)} = -0.10$). (c) contains significant talking and other noise, and is marked as 'Unknown' by the clinician. The algorithm correspondingly rejects the signal with a low confidence of $C^{(\hat{u})} = 0.49$. (d) contains a lower amplitude signal with some noise that is marked as 'Unknown' by the clinician, but is segmented by the algorithm with a very high confidence of $C^{(\hat{u})} = 0.91$.

https://doi.org/10.1371/journal.pdig.0000436.g005

Table 1. Confusion matrix for the cross-validated murmur detection task.

Predicted Class	True Class		
	Murmur	**Unknown**	**No murmur**
Murmur	166	19	117
Unknown	1	21	39
No murmur	12	28	539

https://doi.org/10.1371/journal.pdig.0000436.t001

to their code which made them ineligible for the final prizes [8]. Our algorithm therefore won the First Prize for murmur detection. The chosen metric for the challenge was a 'weighted accuracy' that gave murmur examples five times the importance of normal examples, which meant the high sensitivity operating point shown in Fig 6 performed well.

Table 2. Per-class results for the murmur detection task.

Class	Cases	Sensitivity (%)	PPV (%)	F_1 score
Murmur present	179	92.7	55.0	0.690
Unknown	68	30.9	34.4	0.326
Murmur absent	695	77.6	93.1	0.846

Results are evaluated via 5-fold cross-validation of the training dataset. Shown are sensitivity (also known as recall), positive predictive value (PPV, also known as precision), and their combined F_1 mean.
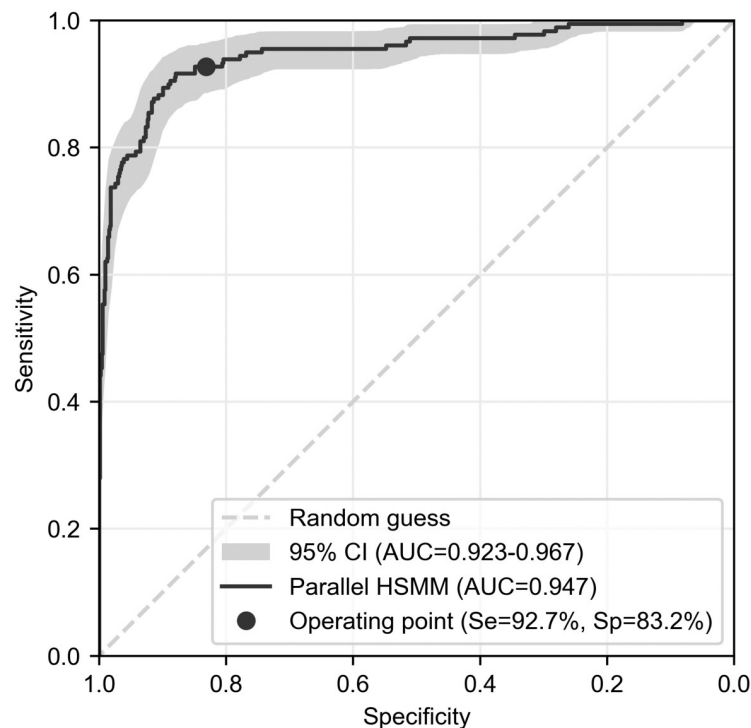
https://doi.org/10.1371/journal.pdig.0000436.t002

Fig 6. Receiver operating characteristic curve for murmur detection algorithm. The sensitivity and specificity of the parallel HSMM model is shown as the threshold confidence difference $C^{(M-N)}$ varies. The normal operating point of the algorithm (i.e. decide murmur if $C^{(M-N)} > 0$) is also shown. A 95% confidence interval (CI) in the curve has been calculated through bootstrap resampling with 1000 iterations. Patients with an 'unknown' murmur label have been removed.
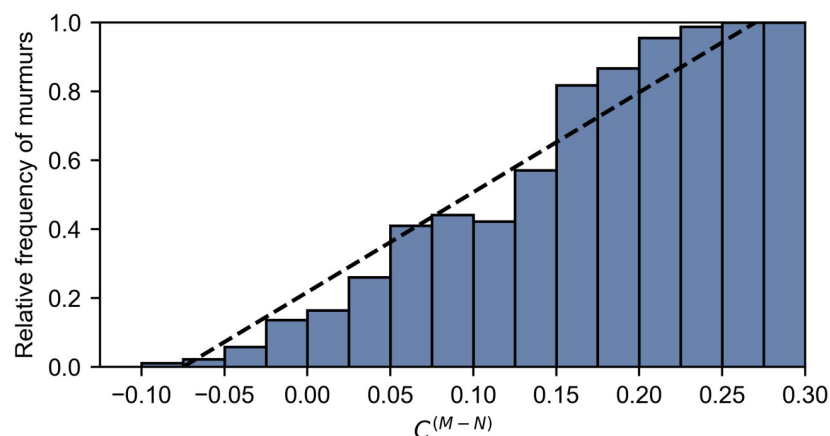
https://doi.org/10.1371/journal.pdig.0000436.g006

Fig 7. Reliability diagram for prediction of murmurs using HSMM confidence difference. As the confidence difference $C^{(M-N)}$ increases, so does the relative frequency of murmurs with an almost linear relationship (dashed line). This suggests good calibration of the murmur likelihood output.

https://doi.org/10.1371/journal.pdig.0000436.g007

One additional advantage of the algorithm compared to deeper methods is the computational resources required for training. The top-scoring algorithm took 18 hours and 51 minutes to train on a NVIDIA T4 Tensor Core GPU, whereas our algorithm took just 38 minutes [8].

Limitations of analysis. One limitation of the CirCor dataset is that the murmur labels were annotated by a single clinician rather than a board of reviewers [8]. The labelling of a heart murmur depends heavily on the annotator's skill, hearing acuity, and headphones. Therefore, it is not as repeatable a ground-truth as other cardiac tests such as electrocardiography and echocardiogram and some disagreements between the algorithm and the clinician label are to be expected. Future work should investigate recordings where a large majority of challenge teams disagree with the clinician label, as this may be indicative of mislabels.

An additional limitation in the dataset labelling, as mentioned earlier, is that the ground-truth does not distinguish between a pathological and physiological (innocent) murmur. This is an important aspect of clinical deployment that cannot be modelled using this dataset. Referring many patients with physiological murmurs for follow-up investigation may overwhelm secondary care departments. Future studies should use an echocardiographic gold standard to label pathological murmurs, and so assess the sensitivity and PPV of algorithms for the ultimate clinical aim of detecting significant structural heart disease.

Prediction of clinical outcome

Cross-validated results. Fig 8 plots an ROC curve for the performance of the CatBoost algorithm at predicting 'normal' or 'abnormal' clinical outcome. The CatBoost algorithm achieves an area under the ROC of 0.691. The challenge operating point of this algorithm gives a sensitivity of 84.2% and a specificity of 31.3%, which results in a final challenge score of
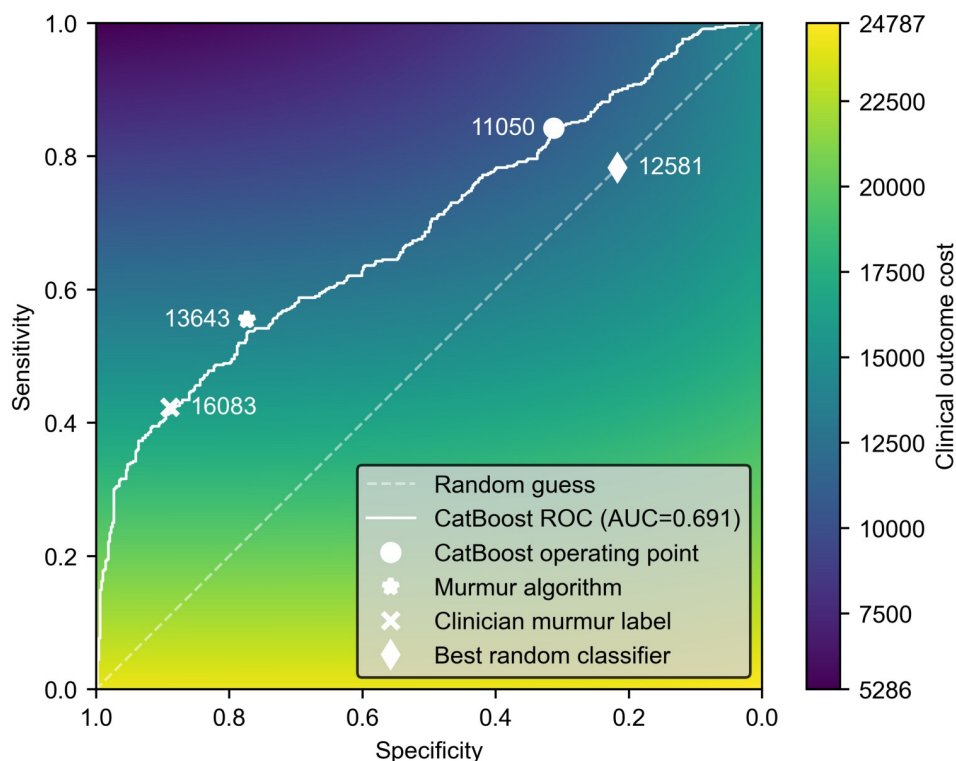

Fig 8. Performance of algorithms and clinician on clinical outcomes task. A receiver operating characteristic (ROC) curve for the CatBoost algorithm at predicting clinical outcome is shown. Also shown are operating points for the clinician (using the murmur label) and the parallel HSMM murmur detection algorithm. The colourmap shows how sensitivity and specificity relate to the PhysioNet challenge cost score. On the training set, an optimum random classifier can achieve a cost score of 12581.

https://doi.org/10.1371/journal.pdig.0000436.g008

Table 3. Performance of different predictors on the clinical outcome task.

Operating point	Sensitivity (%)	Specificity (%)	Youden's index	Challenge cost
CatBoost algorithm	84.2	31.3	0.155	11050
Murmur algorithm	55.5	77.4	0.328	13643
Clinician label	42.3	88.9	0.312	16083
Random classifier	78.3	21.7	0	12581

The operating points listed are shown graphically in Fig 8. Youden's index is calculated as 'sensitivity + specificity − 1'. The 'murmur algorithm' row denotes the performance if the murmur predictions from the parallel HSMM algorithm were directly used to predict clinical outcome. Similarly, the 'clinician label' row shows the performance if the annotated murmur labels were directly used to predict the clinical outcome, assuming unknown cases are referred on as positives.

https://doi.org/10.1371/journal.pdig.0000436.t003

11050. The corresponding micro-averaged accuracy is 0.569. For a fixed prevalence and dataset size, the challenge cost function can be plotted as a function of sensitivity and specificity, as shown in the colourmap of the figure.

Table 3 compares the CatBoost operating point to other possible predictors, which are also shown in Fig 8. Shown are operating points if the presence of a murmur (either annotated by the clinician, or predicted by our parallel HSMM algorithm) is taken as a direct indication of an abnormal clinical outcome. For both these predictors, we assume all cases that are predicted 'unknown' are referred on for follow-up because they cannot be confidently screened out. Both the murmur algorithm and the clinician are specific, but their sensitivity is poor, indicating that many patients have heart disease that does not produce an audible signature. Both the murmur algorithm and clinician labels achieve a Youden's index (0.328 and 0.312 respectively) that is higher than is achieved by the CatBoost model optimised for the challenge cost function. However, the CatBoost challenge score (11050) is significantly better than the murmur detection algorithm (13642) and the clinician label (16083). These two latter scores are worse than a hypothetical random classifier on the random diagonal line of the ROC, shown in Fig 8, and achieves a score of 12581.

The colourmap in Fig 8 shows that the overall effect of the cost function is to heavily prioritise sensitivity over specificity. The figure also illustrates that the main effect of the CatBoost algorithm has been to shift the operating point from a specific area (i.e. the murmur detection performance) to a sensitive area. This is highly beneficial to lower the challenge cost function, but may not be practical for widespread screening. Many population-level screening programs prioritise a high specificity over sensitivity [26], because this is crucial to maintaining a high positive predictive value when operating over a low prevalence population. An algorithm with low specificity could lead to a very large number of false positive referrals overwhelming secondary care cardiac services, such as echocardiography.

Challenge test results. On the challenge test set for the clinical outcome task, the algorithm achieved a micro-averaged accuracy of 0.602 and an overall challenge cost score of 11144. This was the winning score. As for the murmur detection task, these test results do not differ significantly from the training results (0.569 and 11050 respectively), again suggesting that the model has not significantly overfit to the training data.

In general, across challenge entries, the performance of the algorithm at detecting clinical outcome was poor compared to murmur detection. The CirCor dataset only provides a binary label of disease and does not provide additional labels on its nature or severity [8]. Given the low sensitivity of clinical auscultation, it is likely that many of the patients recruited have diseases that do not produce audible murmurs or other abnormal sounds. Phonocardiography is a useful tool to detect many structural heart diseases but should be combined with non-invasive cardiac screening tests (e.g. electrocardiography) to provide a more sensitive test for heart

dysfunction. It may therefore be beneficial to focus algorithm designs on diseases that are known to produce abnormal sounds (e.g. valvular heart disease, septal defects) rather than training models to predict a general abnormality.

Analysis of the PhysioNet 2022 challenge

A notable observation of the challenge results is that the ranking of entries changed significantly between the validation and test sets. The final scores also showed a significant difference between the validation and the training and test sets. The validation set (10% of the data) was smaller than the test set (30%) and teams were allowed up to 10 submissions to the validation set. Teams, therefore, optimised their algorithm for best performance on the validation set, which, given the relatively small total dataset, was not entirely representative of the final test set. In particular, the prevalence of the 'abnormal outcome' in the validation set (0.383) was significantly lower than in the test set (0.507) [8].

A novel part of the 2022 challenge was the use of the custom cost function for the binary clinical outcome task. The challenge organisers argue that traditional metrics for binary classifiers, such as area under the receiver operating characteristic curve, weigh all examples equally and are not optimised for particular clinical contexts [27]. The 2022 cost function was designed to represent the key issues in the deployment of an algorithm in low-cost screening environments [8]. However, one potential limitation of solely using a custom loss function is that results from the challenge cannot be easily compared to other studies applying machine learning to PCG analysis. The cost value can also mask whether the algorithms are actually identifying diagnostic features. Approximately half of the challenge teams (19) achieved a worse performance on the test set than a random classifier. An optimum random model that achieves a sensitivity of 80% and a specificity of 20% (thus lying on the diagonal line of a receiver operating characteristic graph) on the test set would achieve a challenge cost score of 13168. In Fig 9, we compare the training and test scores for all the official challenge entries.

Fig 9 also shows that shows that many algorithms overfit to the training set with far worse performance on the final testing set. Many teams used deep learning algorithms commonly deployed in other areas such as speech recognition. However, the small size of the dataset makes training generalisable models a challenge. Although an RNN is used in this work, it is constrained to perform a specific task within the segmentation model and does not generate the final prediction of clinical outcome. Unlike many other teams, our approach does not retrain a completely new model to predict clinical outcome. The CatBoost model was designed to leverage the murmur predictions from the parallel HSMM and combine them with limited patient biometrics. Therefore, it had a limited feature set to train on and the risk of overfitting was low. However, one limitation of this approach is that the murmur detector algorithm is only trained to detect audible abnormal sounds. It is possible that some of the abnormal examples contain inaudible time-frequency features that would be missed by the murmur detection algorithm and hence the CatBoost model. A model trained directly to predict clinical outcome could detect these features.

Conclusion

We present a novel algorithm to detect and classify heart murmurs that was the winning entry in the 2022 PhysioNet challenge. The model uses a hybrid approach combining a recurrent neural network with parallel hidden semi-Markov models to accurately segment and classify signals, even in the presence of noise and murmurs. Compared to many other algorithms described in the literature and used in the challenge, our model is lightweight and can easily be interpreted by a clinician.
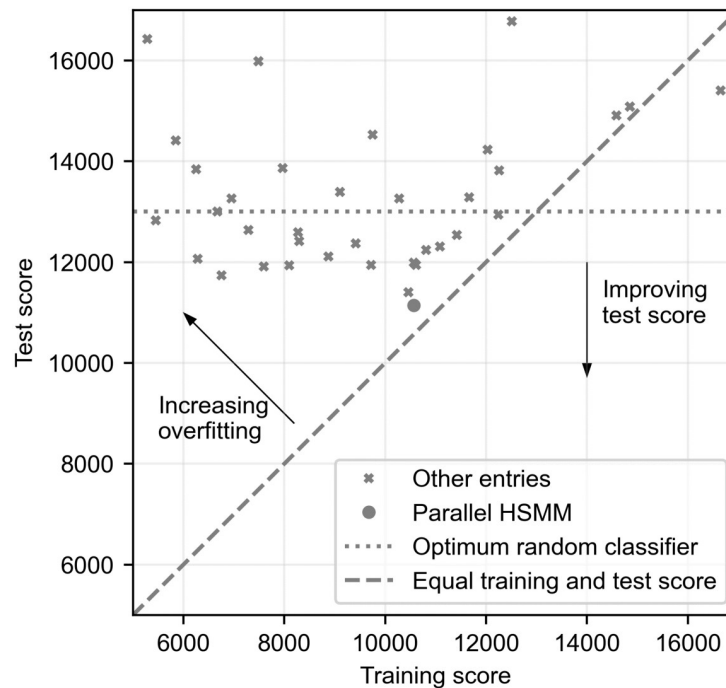
Fig 9. Training and test performance of all the official entries on the clinical outcomes task. A lower score is better. Many algorithms had significantly better performance on the training set, indicating overfitting has taken place. The diagonal line shows ideal performance where training and test performance are equal. The parallel HSMM algorithm achieves the best score and shows little evidence of overfitting. For context, the performance of a random classifier at an optimal operating point is also shown.

https://doi.org/10.1371/journal.pdig.0000436.g009

On the murmur detection task, the model won the first prize in the challenge with a sensitivity of 92.7% and 77.6% for the 'murmur present' and 'murmur absent' classes respectively. The algorithm also won the first prize in the clinical outcome task. However, its accuracy was reduced compared to the murmur detection task because many of the abnormal patients did not have audible pathological sounds. More specific labels of disease were not available in the CirCor dataset, and future work could investigate the accuracy of these approaches on a per-disease basis. The dataset labels also did not differentiate between physiological and pathological murmurs, meaning that the full clinical significance of the murmur detection task is unclear.

The structure of the challenge allowed an independent comparison of different machine learning algorithms. However, it was not possible to directly compare the performance of these algorithms with a general clinician using a stethoscope, as the CirCor dataset did not include a prospective evaluation of blinded clinicians' performance. Future studies should prospectively compare general clinicians with these algorithms, to investigate any performance increase and provide early health economic evidence for replacing clinician auscultation with AI.

Whilst the CirCor data used in this work is a significant improvement over previous open-access datasets, future studies are needed to validate the use of automated heart sound analysis in specific clinical pathways. The relative importance of sensitivity and specificity will significantly differ depending on health system priorities and the availability of follow-up tests such as echocardiography.

The algorithm additionally predicts signal quality, so a user can be asked to make a repeat recording if their stethoscope was incorrectly held or if there is substantial bodily or environmental noise. However, the algorithm's predictions often disagreed with the signal quality label assigned by the clinician. For widespread and low-cost screening, it is essential that heart sound data can be reliably gathered by an unskilled operator. Future studies should explore the usability of electronic stethoscopes and investigate if automatic signal quality assessment can aid this process.

Supporting information

S1 Fig. Improved estimate of heart rate using RNN output. Compared here are the methods of Schmidt et al. [23] (left column) and our approach (right column). Schmidt et al. take a homomorphic envelope (middle left) of the signal and then compute its autocorrelation (bottom left). They then search for a peak in a specified range to estimate the heart rate. We use a range of 30-180 bpm for both methods in this dataset because of the faster paediatric sounds. However, this example heart sound (top left) has significant noise which corrupts the envelope and therefore gives a noisy autocorrelation where the correct peak is difficult to find. Our approach instead uses the output of the RNN to create a signal that shows the probability the signal is not in diastole, $P(q_t \neq \text{diastole}|x_{1:T}, \theta)$, (i.e. the summed probability of the S1, S2, systole, and systolic murmur states, middle right). This is a much cleaner signal than the homomorphic envelope, so its autocorrelation (bottom right) is much clearer and the correct peak corresponding to the signal period is easy to find.
(TIF)

Acknowledgments

Author Contributions

Conceptualization: Andrew McDonald, Anurag Agarwal.

Formal analysis: Andrew McDonald, Mark J. F. Gales, Anurag Agarwal.

Investigation: Andrew McDonald, Mark J. F. Gales, Anurag Agarwal.

Methodology: Andrew McDonald, Mark J. F. Gales, Anurag Agarwal.

Project administration: Anurag Agarwal.

Resources: Anurag Agarwal.

Software: Andrew McDonald.

Supervision: Mark J. F. Gales, Anurag Agarwal.

Validation: Andrew McDonald, Mark J. F. Gales.

Visualization: Andrew McDonald, Mark J. F. Gales, Anurag Agarwal.

Writing – original draft: Andrew McDonald.

Writing – review & editing: Andrew McDonald, Mark J. F. Gales, Anurag Agarwal.

References

1. World Health Organisation. Cardiovascular diseases fact sheet; 2023. Available from: https://www.who.int/health-topics/cardiovascular-diseases.

2. Gaede L, Di Bartolomeo R, van der Kley F, Elsässer A, Iung B, Möllmann H. Aortic valve stenosis: what do people know? A heart valve disease awareness survey of over 8, 800 people aged 60 or over. EuroIntervention. 2016; 12(7):883–889. https://doi.org/10.4244/EIJY16M06_02 PMID: 27283409

3. Coffey S, Roberts-Thomson R, Brown A, Carapetis J, Chen M, Enriquez-Sarano M, et al. Global epidemiology of valvular heart disease. Nature Reviews Cardiology. 2021; 18(12):853–864. https://doi.org/10.1038/s41569-021-00570-z PMID: 34172950

4. Watkins DA, Johnson CO, Colquhoun SM, Karthikeyan G, Beaton A, Bukhman G, et al. Global, Regional, and National Burden of Rheumatic Heart Disease, 1990–2015. New England Journal of Medicine. 2017; 377(8):713–722. https://doi.org/10.1056/NEJMoa1603693 PMID: 28834488

5. Constant J. Essentials of Bedside Cardiology: With a Complete Course in Heart Sounds and Murmurs on CD. 2nd ed. Totowa, New Jersey: Humana Press; 2003.

6. Gardezi SKM, Myerson SG, Chambers J, Coffey S, d'Arcy J, Hobbs FDR, et al. Cardiac auscultation poorly predicts the presence of valvular heart disease in asymptomatic primary care patients. Heart. 2018; 104(22):1832–1835. https://doi.org/10.1136/heartjnl-2018-313082 PMID: 29794244

7. Clifford GD, Liu C, Moody B, Millet J, Schmidt S, Li Q, et al. Recent advances in heart sound analysis. Physiological Measurement. 2017; 38(8):E10–E25. https://doi.org/10.1088/1361-6579/aa7ec8 PMID: 28696334

8. Reyna MA, Kiarashi Y, Elola A, Oliveira J, Renna F, Gu A, et al. Heart murmur detection from phonocardiogram recordings: The George B. Moody PhysioNet Challenge 2022. PLOS Digital Health. 2023; 2 (9):e0000324. https://doi.org/10.1371/journal.pdig.0000324 PMID: 37695769

9. McDonald A, Gales M, Agarwal A. Detection of Heart Murmurs in Phonocardiograms with Parallel Hidden Semi-Markov Models. In: 2022 Computing in Cardiology Conference (CinC). Computing in Cardiology; 2022. Available from: http://dx.doi.org/10.22489/CinC.2022.020.

10. Kay E, Agarwal A. DropConnected neural networks trained on time-frequency and inter-beat features for classifying heart sounds. Physiological Measurement. 2017; 38(8):1645–1657. https://doi.org/10.1088/1361-6579/aa6a3d PMID: 28758641

11. Liu C, Springer D, Li Q, Moody B, Juan RA, Chorro FJ, et al. An open access database for the evaluation of heart sound algorithms. Physiological Measurement. 2016; 37(12):2181–2213. https://doi.org/10.1088/0967-3334/37/12/2181 PMID: 27869105

12. Bentley PJ, Nordehn G, Coimbra MT, Mannor S. The PASCAL Classifying Heart Sounds Challenge 2011 (CHSC2011) Results; 2011. Available from: http://www.peterjbentley.com/heartchallenge/index.html.

13. Oliveira J, Renna F, Costa PD, Nogueira M, Oliveira C, Ferreira C, et al. The CirCor DigiScope dataset: from murmur detection to murmur classification. IEEE Journal of Biomedical and Health Informatics. 2021; 26(6):2524–2535. https://doi.org/10.1109/JBHI.2021.3137048

14. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation. 2000; 101(23):e215–e220. https://doi.org/10.1161/01.CIR.101.23.e215 PMID: 10851218

15. Oliveira J, Renna F, Costa P, Nogueira M, Oliveira AC, Elola A, et al. The CirCor DigiScope Phonocardiogram Dataset (version 1.0.3); 2022. Available from: https://doi.org/10.13026/tshs-mw03.

16. Springer DB, Tarassenko L, Clifford GD. Logistic Regression-HSMM-Based Heart Sound Segmentation. IEEE Transactions on Biomedical Engineering. 2016; 63(4):822–832. PMID: 26340769

17. Kay E. Cardiac Acoustics: Understanding and Detecting Heart Murmurs. PhD Thesis, University of Cambridge; 2018.

18. Messner E, Zohrer M, Pernkopf F. Heart Sound Segmentation-An Event Detection Approach Using Deep Recurrent Neural Networks. IEEE Transactions on Biomedical Engineering. 2018; 65(9):1964–1974. https://doi.org/10.1109/TBME.2018.2843258 PMID: 29993398

19. McDonald A. A recurrent neural network and parallel hidden Markov model algorithm to detect heart murmurs. GitHub repository; 2023. Available from: https://github.com/am2234/parallel-hsmm-murmur.

20. Nussbaumer M, Agarwal A. Stethoscope acoustics. Journal of Sound and Vibration. 2022; 539:117194. https://doi.org/10.1016/j.jsv.2022.117194

21. Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation; 2014. Available from: https://arxiv.org/abs/1406.1078.

22. Kingma DP, Ba JL. Adam: a Method for Stochastic Optimization. In: International Conference on Learning Representations 2015. San Diego; 2015. p. 1–15.

23. Schmidt SE, Holst-Hansen C, Graff C, Toft E, Struijk JJ. Segmentation of heart sound recordings by a duration-dependent hidden Markov model. Physiological Measurement. 2010; 31(4):513–529. https://doi.org/10.1088/0967-3334/31/4/004 PMID: 20208091

24. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features; 2017. Available from: https://arxiv.org/abs/1706.09516.

25. PhysioNet Challenge organisers. George B. Moody PhysioNet Challenge 2022; 2022. Available from: https://physionetchallenges.org/2022/.

26. Maxim LD, Niebo R, Utell MJ. Screening tests: a review with examples. Inhalation Toxicology. 2014; 26 (13):811–828. https://doi.org/10.3109/08958378.2014.955932 PMID: 25264934

27. Reyna MA, Nsoesie EO, Clifford GD. Rethinking Algorithm Performance Metrics for Artificial Intelligence in Diagnostic Medicine. JAMA. 2022; 328(4):329. https://doi.org/10.1001/jama.2022.10561 PMID: 35802382