Two-stage Classification for Detecting Murmurs from Phonocardiograms Using Deep and Expert Features

Sara Summerton[1], Danny Wood[1], Darcy Murphy[1], Oliver Redfern[2], Matt Benatan[3], Matti Kaisti[4], David C Wong[1]

[1] University of Manchester, Manchester, UK (note - full address at end of paper)
[2] University of Oxford, Oxford, UK [3] Independent Researcher, Manchester, UK [4] University of Turku, Turku, Finland

Abstract

Detection of heart murmurs from stethoscope sounds is a key clinical technique used to identify cardiac abnormalities. We describe the creation of an ensemble classifier using both deep and hand-crafted features to screen for heart murmurs and clinical abnormality from phonocardiogram recordings over multiple auscultation locations. The model was created by the team Murmur Mia! for the George B. Moody PhysioNet Challenge 2022.

Methods: Recordings were first filtered through a gradient boosting algorithm to detect Unknown. We assume that these are related to poor quality recordings, and hence we use input features commonly used to assess audio quality. Two further models, a gradient boosting model and ensemble of convolutional neural networks, were trained using time-frequency features and the mel-frequency cepstral coefficients (MFCC) as inputs, respectively. The models were combined using logistic regression, with bespoke rules to convert individual recording outputs to patient predictions.

Results: On the challenge validation set, our classifier scored 0.737 for the weighted accuracy and 11828 for clinical outcome challenge metric. This placed 28/305 and 188/305 on the challenge leaderboard, for each scoring metric, respectively.

1 Introduction

Global morbidity burden in childhood caused by heart disease is disproportionately distributed in low- and middle-income countries (LMICs) [1]. While mortality from congenital heart disease (CHD) has declined globally, the decline has been slower in poorer countries [2]. Similarly, over 300,000 people each year die from acquired heart diseases such as rheumatic heart disease (RHD) and over 80 percent of these deaths are in LMICs [3]. CHD is a structural abnormality of the heart or great vessels, affecting 1% of live births [4]. RHD is a common sequalae of rheumatic fever infection, resulting in heart valve damage [5]. Both CHD and RHD are easily diagnosed by cardiac ultrasound (echocardiography). Echocardiography requires specialist operators and clinical interpretation, both of which are scarce in severely resource-limited healthcare systems [5]. However, early diagnosis is essential to improve outcomes in children with heart disease.

Before the widespread introduction of echocardiography, structural heart disease was diagnosed through cardiac auscultation [4]. The opening and closing of heart valves during the normal cardiac cycle are clearly audible with a stethoscope. Abnormal, turbulent blood flow across the heart valves can also be heard as a "whoosh-ing" sound, known as a murmur. While not all murmurs are pathological, they can be indicative of structural defects. A phonocardiogram (PCG) is an audio recording obtained from an electronic stethoscope. Signal processing and machine-learning of PCG data could provide an objective way to identify potential cardiac pathology. In contrast to echocardiography, electronic stethoscopes require little specialist training, providing a cost-effective tool to screen populations for CHD and RHD.

The 2022 George B. Moody PhysioNet Challenge was to develop a heart murmur classifier from PCG recordings. Scoring metrics were designed to discourage underprediction of murmurs where pathology was subsequently confirmed by echocardiography.

2 Methods

The dataset for this challenge was the CirCor DigiScope dataset [6]. The task and dataset are described in detail in [6], [7].

Each recording in the training set was annotated with salient points (commencement of S1, systole, S2, diastole) by clinical experts. These annotations were used to train a hidden Markov model to fulfil the same role in our classifier, the outputs of which were used for feature extraction.

We used an ensemble of three classifiers to predict the presence of murmur, and two to predict clinical outcome. The classifiers were trained separately on individual recordings, and combined to output predictions on a per-patient basis using a set of hand-crafted rules, as depicted in Figure 1. Gradient boosting classifiers were trained using SKlearn and neural networks were trained using Pytorch; the regression coefficients used to combine the models were calculated using SKlearn. We describe each component of our ensemble below.

2.1 Data Preprocessing

Demographic variables were missing for 69 participants. We imputed missing data using iterative imputation. We restricted the imputed data to plausible values.

In order to improve classifier discrimination of audible murmurs, training labels were reassigned on a per-recording basis. That is, if the patient had recordings from the mitral valve (MV) and tricuspid valve (TV), but the murmur was only marked as audible in MV, only the MV recording was assigned 'Present' and the recording from TV was given the label 'Absent' for training purposes.

Our approach makes use of [S1, S2] segmentation. As these annotations are not available in validation and test datasets, we created our own annotations using Springer's hidden Markov model approach [8]. We modified Springer's algorithm to make use of demographic data from each patient in predicting the heart rate, as well as tweaking the filters to have a narrower passband, using the frequencies suggested by [9]. We also ported the method to Python (available at: https://github.com/EchoStatements/Springer-Segmentation-Python).

2.2 Unknown Detector

We assumed, a priori, that at least some proportion of participants labeled 'Unknown' would be due to poor signal quality for the duration of the audio recording. We created the Unknown detector to recognize these poor quality recordings, using a set of hand-crafted recording-level features, previously described and used by Zabihi et al. [10].

The probabilities of {Present, Unknown, Absent} from this model were saved on a per-recording basis, yet only the probability of 'Unknown' contributed to the final patient label, as described in 2.5.

2.3 Hand-crafted Feature Classifier

We created a second stochastic gradient boosting model to detect murmurs from a single PCG recording. Inputs to this model were a set of hand-crafted time-frequency features. The annotation described in 2.1 was used to isolate sections of the recording that corresponded to the same stage in the cardiac cycle, then each section was further subsegmented into five equal partitions in time. We applied a Hamming window to each subsegment to reduce edge effects, then took the Fourier transform, in order to calculate the relative amplitude of frequency bands from 5 through 400 Hz. The bins used were determined experimentally through Shapley value analysis in order to maximize the resolution of bands containing the most useful information for the classifier, typically in the 5-25 Hz and 80-180 Hz ranges. The average of the extracted time-frequency features for each subsegment was taken across the entire recording, as well as the mean and standard deviation of the cardiac cycle stage. This gave 80 time-frequency features, which were included with the patient demographic data and heart rate to give a total of 86 inputs to the gradient boosting model.

Two of these gradient boosting models were trained: one against murmur labels, and the other for clinical outcome. The probabilities of {Present, Unknown, Absent} were returned by one model, and {Abnormal, Normal} by the other, on a per-recording basis.

2.4 Convolutional Neural Network Classifier

Our final classifier was an ensemble of convolutional neural networks (CNNs) using mel-frequency cepstral coefficient (MFCC) inputs. Data processing followed the prior work of Rubin et al. [11]. We adapted their CNN model architecture to include demographic information (see Figure 2). Inclusion of such tabular information has been associated with modest increase in performance for similar problems [12]. The CNN was trained to perform binary classification, with {Present, Unknown} both assigned to the same label.

We created an ensemble of five of these CNNs by freezing the model weights before the last layer. We concatenated the models and combined their outputs through two fully connected layers, reducing the network bandwidth from $(5 \times 512) \rightarrow 512 \rightarrow 1$.

2.5 Combining the Base Models

Our combination rules introduced a set of hyperparameters which were used to balance the contributions of each model in order to maximize performance on the challenge scoring metrics. The unknown detectors described in 2.2 were applied first; a patient was classified as an outlier only if the respective model reached a threshold confidence ($>$90% for 'Unknown' murmur, $>$70% for 'Abnormal' outcome). For murmur score, we used linear regression to determine the optimal contribution of the gradient boosting and CNN ensembles to the recording label. The normalized geometric mean of the confidence levels for each recording was used to set the final patient-level classifica-

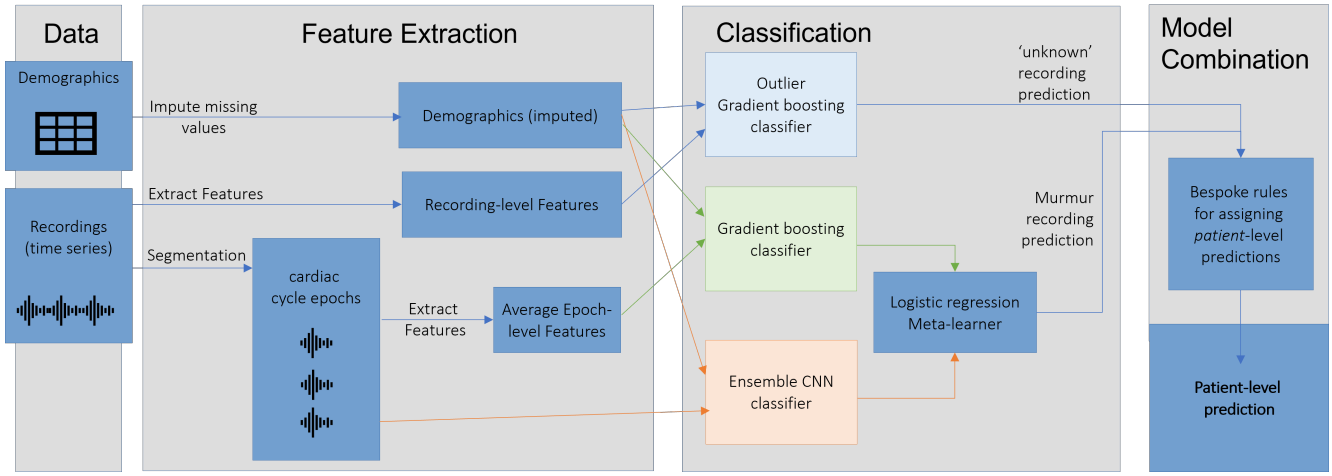

Figure 1. Overall system diagram, showing how patient data were processed, the base classifiers used, and how they were combined to provide a patient-level prediction.

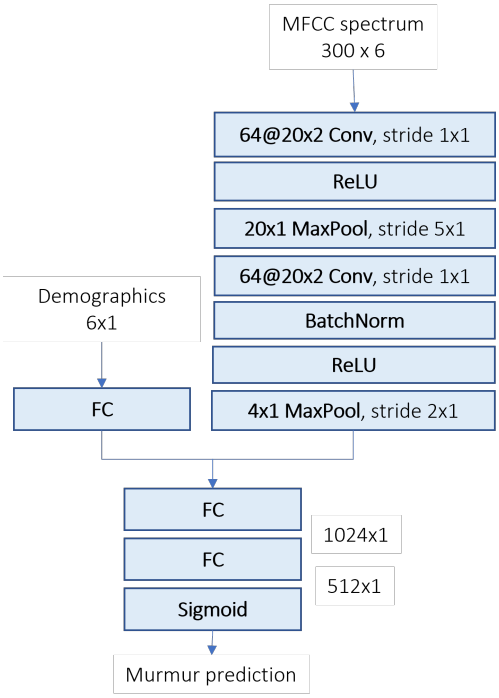

Figure 2. CNN architecture for PCG recording classifier.

tion for both murmur score and outcome, with a series of threshold parameters to allow a single recording to "overrule" the mean. This ensured that recordings with murmurs which may only be detected in certain locations were given priority when assigning a patient-level classification.

We optimized these parameters, as well as the method of combining recording-level predictions to produce a patient-level prediction, via a constrained grid search using the bespoke challenge metrics, described in 2.6.

Table 1. Performance metrics for Murmur ensemble and constituent models on the training data from 5-fold cross validation. The combined model was trained to weigh murmurs more heavily, so while the unweighted per-recording accuracy is lower than the individual models, it performed better on the weighted scoring metric.

Model	Per-rec. Acc.	Weighted Acc.
Outlier G. Boost.	$81.7\% \pm 1.3$	0.50 ± 0.03
Murmur G. Boost.	$86.5\% \pm 1.5$	0.63 ± 0.04
Murmur CNN	$88.2\% \pm 0.7$	0.65 ± 0.03
Combined Model	$84.9\% \pm 0.6$	0.75 ± 0.03

2.6 Model Evaluation

Five-fold cross-validation was used to assess the performance of the ensemble models. We used the challenge metrics to evaluate performance. For murmur classification, this was a weighted accuracy in which the Murmur and Unknown classes were weighted by 5 and 3 respectively. For clinical outcome, this was a bespoke metric (defined in full in [7]) that aimed to balance costs of clinical time against clinical errors.

3 Results

The training cross-validation performance of the constituent models is shown in Table 1 and 2 for Murmur and Outcome labels, respectively. Training and validation set performance of the combined models is shown in Table 3.

We note that accuracy at a per-epoch and per-recording level did not necessarily translate to the same per-patient accuracy. While this may be a limitation of our rules for combining models, we also note that murmur sounds are

Table 2. Performance metrics for Outcome model and constituent models on the training data from 5-fold cross validation. The two models were combined on a per-patient basis.

Model	Per-rec. Acc.	Score
Outlier G. Boost.	$59.0\% \pm 2.1$	14232 ± 1475
Outcome G. Boost.	$59.3\% \pm 2.1$	13767 ± 1242

Table 3. Murmur (weighted accuracy) and clinical outcome (cost) challenge scores for the final selected entry for team Murmur Mia!. Test set scores are from 5-fold cross validation on the public training set; scores from the hidden validation data were provided during the official phase. Ranking is our chosen model out of all models submitted.

Metric	Training	Valid.	Ranking
Weighted accuracy	0.753 ± 0.030	0.737	28/305
Clinical outcome	11965 ± 655	11828	188/305

not always audible in all auscultation locations. Clear murmur sounds can depend on the type of murmur, the physiology and position of the patient, and whether they are inhaling and exhaling.

4 Discussion

We have developed an ensemble of models that used both expert and learned features. In both cross-validation, and on validation data, this blended approach performed better than relying solely on a deep learning model.

Our approach was tailored to detect murmurs, regardless of other contextual information. However, we know clinically that the presence of a murmur may not be the only or even main factor used by a medical professional to decide if a patient needs further treatment. For instance, in many cases, murmurs are 'innocent' and of no clinical concern [13]. It is likely that our focus on murmur, regardless of 'innocence' is one of the reasons why our approach was relatively poor at determining clinical outcome.

In future work, we intend to improve the murmur model by exploring how to effectively extract time-domain features, for instance, using a ResNet. We also wish to consider how effective relabelling might improve prediction of clinical outcomes.

References

[1] Coffey S, Roberts-Thomson R, Brown A, Carapetis J, Chen M, Enriquez-Sarano M, Zühlke L, Prendergast BD. Global epidemiology of valvular heart disease. Nature Reviews Cardiology 2021;18(12):853–864. ISSN 17595010.

[2] Wu W, He J, Shao X. Incidence and mortality trend of congenital heart disease at the global, regional, and national level, 1990-2017. Medicine United States 2020;99(23). ISSN 15365964.

[3] Rwebembera J, Beaton AZ, de Loizaga SR, Rocha RT, Doreen N, Ssinabulya I, Okello E, Fraga CL, Galdino BF, Nunes MCP, Nascimento BR. The Global Impact of Rheumatic Heart Disease. Current Cardiology Reports 2021;23(11). ISSN 15343170. URL https://doi.org/10.1007/s11886-021-01592-2.

[4] Van Der Linde D, Konings EE, Slager MA, Witsenburg M, Helbing WA, Takkenberg JJ, Roos-Hesselink JW. Birth prevalence of congenital heart disease worldwide: A systematic review and meta-analysis. Journal of the American College of Cardiology 2011;58(21):2241–2247. ISSN 15583597.

[5] Marijon E, Mocumbi A, Narayanan K, Jouven X, Celermajer DS. Persisting burden and challenges of rheumatic heart disease. European Heart Journal 2021;42(34):3338–3348. ISSN 15229645.

[6] Oliveira JH, Renna F, Costa P, Nogueira D, Oliveira C, Ferreira C, Jorge A, Mattos S, Hatem T, Tavares T, Elola A, Rad A, Sameni R, Clifford GD, Coimbra MT. The CirCor DigiScope Dataset: From Murmur Detection to Murmur Classification. IEEE Journal of Biomedical and Health Informatics 2021;1–12. ISSN 21682208.

[7] Reyna MA, Elola A, Oliveira J, Renna F, Gu A, Sadr N, Alday EAP, Kiarashinejad Y, Mattos S, Coimbra MT, Sameni R, Rad AB, Clifford GD. Heart Murmur Detection from Phonocardiogram Recordings: The George B. Moody PhysioNet Challenge 2022 2022;URL https://moody-challenge.physionet.org/2022/.

[8] Springer DB, Tarassenko L, Clifford GD. Logistic regression-hsmm-based heart sound segmentation. IEEE Transactions on Biomedical Engineering 2015;63(4):822–832.

[9] Giordano N, Knaflitz M. A novel method for measuring the timing of heart sound components through digital phonocardiography. Sensors 2019;19(8):1868.

[10] Zabihi M, Rad AB, Kiranyaz S, Gabbouj M, Katsaggelos AK. Heart sound anomaly and quality detection using ensemble of neural networks without segmentation. Computing in Cardiology 2016;43:613–616. ISSN 2325887X.

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

[12] Zhao Z, Fang H, Relton SD, Yan R, Liu Y, Li Z, Qin J, Wong DC. Adaptive lead weighted resnet trained with different duration signals for classifying 12-lead ecgs. In 2020 Computing in Cardiology. IEEE, 2020; 1–4.

[13] Biancaniello T. Innocent murmurs. Circulation 2005; 111(3):e20–e22.

Address for correspondence:

Sara Summerton

Kilburn Building, Oxford Road, M13 9PL, Manchester, UK

sara.summerton@postgrad.manchester.ac.uk