A Fusion of Handcrafted Feature-Based and Deep Learning Classifiers for Heart Murmur Detection

Zaria Imran[1], Ethan Grooby[1,2], Vinayaka Vivekananda Malgi[3], Chiranjibi Sitaula[1], Sunil Aryal[3], Faezeh Marzbanrad[1]

[1]Electrical and Computer Systems Engineering Department, Monash University, Australia
[2]Electrical and Computer Engineering Department, The University of British Columbia, Canada
[3]School of Information Technology, Deakin University, Australia

Abstract

*As part of George B. Moody Physionet Challenge 2022, our team **Melbourne Kangas**, proposed an algorithm for identifying abnormal heart sounds from paediatric phonocardiograms (PCGs). We developed a Deep Learning (DL) approach and a handcrafted feature-based approach. The DL classifier was based on bidirectional long-short-term-memory and Mel-frequency cepstrum coefficients from raw PCG signals. The feature-based approach used non-negative matrix factorisation to denoise PCG signals and then extracted the features based on the whole and segmented recordings, followed by feature selection. A random under-sampling boosting classifier for murmur classification and robust boosting classifier for outcome classification were given the subset of features. The feature-based performed better than the DL classifiers on the validation set. The feature-based classifier received a weighted accuracy of 0.632 (29th out of 41 teams) and a challenge cost of 11,735 (3rd out of 39 teams) on the test set. Decision fusion of the two approaches decreased 10-fold cross-validation results.*

1. Introduction

To diagnose patients with cardiovascular disease (CVD), clinicians may use heart auscultation to screen for cardiac diseases in phonocardiograms (PCGs) as it is non-invasive and provides information on congenital and acquired CVDs [1]. Technology for observing cardiac activity has improved; however, studies were limited by insufficient datasets. The new dataset released for the challenge, CirCor DigiScope Phonocardiogram Dataset, hopes to fill the need for a paediatric heart sound dataset with comprehensive patient information [1, 2]. This paper presents a Deep Learning (DL) based classification using Long Short Term Memory (LSTM) and a feature-based classifier to perform heart sound classification as part of the George B. Moody Physionet Challenge[1, 2].

2. Methods

The high-level flowchart of the proposed method is presented in Figure 1. Both DL Method (Section 2.2) and Feature-Based Method (Section 2.3) were implemented and evaluated, then fused as specified in Section 2.4

2.1. Dataset

The dataset was divided into ten folds for cross-validation. Due to having follow-up patients in the second data collection, duplicate patient recordings were kept together[1, 2]. Each fold contained an even distribution of patients with murmur present, absent or unknown per age group where possible. After these folds were created with respect to murmur labels, the distribution was checked for outcome labels and was approximately balanced between Normal and Abnormal coincidentally. All samples were recorded at a sample rate of 4000 Hz [1, 2].

2.2. Deep Learning-Based Method

2.2.1. Preprocessing

Imbalance among the murmur classes was accounted for using random undersampling of the most prevalent class (Absent) and random oversampling of the least prevalent class (Unknown) to match the number of patients in the Present Class. Non-segmented audio was used as input similar to previous works [3]. PCG signals were sampled in 10-second segments with additional zero-padding where required. The location of the recording was ignored. 25 Mel-Frequency Cepstral Coefficients (MFCC) were extracted as suggested by Sitaula et al. [4]. The mean was taken over the temporal dimension forming a one-dimensional MFCC input to the classifier.

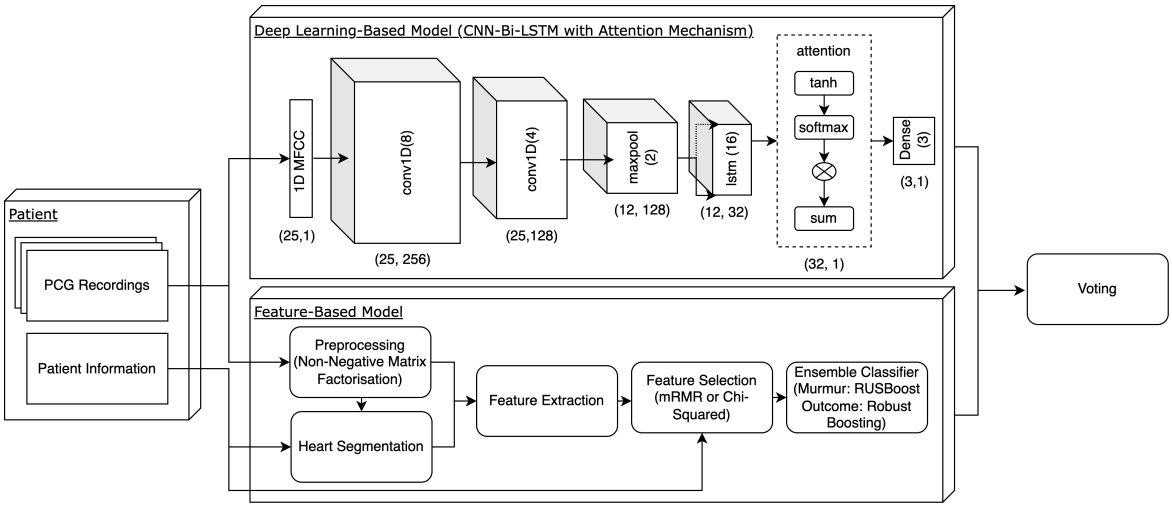

Figure 1. Process of Decision-Fusion of DL-Based Method (Section 2.2) and Feature-Based Method (Section 2.3) with an input of multiple PCG Recordings and Patient Information.

2.2.2. Neural Network Architecture

Two variations of computationally light one-dimensional Convolutional Neural Networks with Long Short Term Memory (1D CNN-LSTM) were trialled. The base CNN contains convolution kernels of size 8 and then 4 consecutively with Rectified Linear Unit (ReLU) activation function. Variations include the addition of an attention mechanism or bidirectionality in the LSTM (Bi-LSTM) [3].

2.2.3. Parameter Selection

Training parameters were manually selected based on training performance curves when changing learning rate, batch size and max epochs. The chosen parameters were 1.00E-06, 32 and 250 respectively. All training used early stopping mechanism based on decreasing validation accuracy with patience of 30 epochs to prevent over-fitting.

2.3. Feature-Based Method

2.3.1. Preprocessing

To remove noise such as lung sounds, and the environment, we used the sound separation method developed in our previous work [5]. This method implements non-negative matrix factorisation (NMF) with a reference database of clean heart and noise sounds, to separate the PCG recording into these respective components [5]. The PCG recording was represented in the time-frequency domain using Short-time Fourier transform with a window size of 512 samples, hop size of 256 samples and a Fast Fourier transform size of 1024 points. Sound separation using NMF, Kullback-Leibler divergence with a sparsity of 0.1 was used in the cost function and 20 basis components for each heart, and noise sound, respectively [5].

2.3.2. Feature Extraction

Firstly, patient information of age, sex, height, and weight were extracted and used as features.

Then features were obtained from both the whole and the segmented heart PCG. Segmentation refers to segmenting the PCG into cardiac cycles, i.e. S1, systole, S2, and diastole. Heart segmentation was performed using a Hidden Semi-Markov Model, which takes into account heart rate range based on age group, which were (1) Neonate: 110-200 bpm, (2) Infant: 70-200 bpm, (3) Child: 60-170 bpm, (4) Adolescent: 40-170 bpm, (5) Young Adult: 40-130 bpm and (5) Unknown: 50-160 bpm [6].

Based on our past work on PCG signal quality, five types of features were extracted from the whole recordings, (1) statistical: variance, skewness and kurtosis of audio/autocorrelation signal, (2) entropy: sample, Shannon, Renyi and Tsallis, (3) power features: total power, various power ratios between 100-1000 Hz, 3 dB bandwidth, 1st, 2nd and 3rd quartiles as well as interquartile range, standard deviation, mean frequency, power centroid and max power, (4) MFCCs: minimum, maximum, mean, median, mode, variance and skewness of a 13-level decomposition in Mel scale and log energy with a window length of 25 ms and overlap length of 15 ms (5) autocorrelation: correlation prominence, sinusoid correlation and Hjorth activity to measure the strength of the signal periodicity [6]. The same features were extracted from segmented signals and averaged for each segment i.e., S1, systole, S2, diastole.

Features based on the best performing models in PhysioNet 2016 Challenge for detection of abnormal heart

sounds, were also extracted [7–10].

2.3.3. Feature Selection

Due to the small dataset in comparison to the number of features extracted, only a subset of features were used to minimise over-fitting. Based on preliminary results, the maximum Relevance Minimum Redundancy (mRMR) algorithm with the mutual information quotient method was used for murmur-based classification [11]. Whereas the Chi-square test (χ^2), which ranked according to p-values, was used for outcome-based classification.

Overall, the top 1-100 features were considered and tested based on the 10-fold cross-validation. The results plateaued around the 40 feature mark, with peak performance using the top 50 features, as reported in Section 3.

2.3.4. Ensemble Classifier Training

Training set features were normalised per fold, with the same scaling and shifting used on the test set features. For murmur-based classification, the following ensemble classifiers were considered: Random Undersampling Boosting (RUSBoost), Bootstrap Aggregation, Adaptive Boosting, Linear Programming Boosting, Totally Corrective Boosting, and Random Subspace Ensemble. For outcome-based classification, the same set of classifiers were considered in addition to Gentle Adaptive Boosting, Adaptive Logistic Regression and Robust Boosting.

Decision tree classifiers were used as weak learners for all classifiers except for subspace, which used discriminant analysis. Using 10-fold cross-validation within the training fold with Bayesian optimisation, the following parameters were optimised when applicable (1) Learning Rate, (2) Number of Learning Cycles, and (3) Minimum Leaf Size for the Base Decision Tree Learner.

For murmur classification, there was a noticeable class imbalance. As demonstrated in our preliminary results, RUSBoost was the best-performing ensemble classifier, which is designed to deal with imbalanced datasets using a combination of random undersampling/oversampling [12]. In RUSBoost, each weak learner is trained on a sampling proportion of recordings with respect to the lowest-represented class, that is, the number of recordings in the unknown murmur class. The Absent:Present:Unknown ratios tested were: (1) 1:1:1, (2) 2:2:2, (3), 3:3:3, (4) 2.5:2.5:1 and (5) 0.75:0.75:0.75.

Finally, the cost function for murmur-based classification was based on the stated competition weight accuracy, with Absent:Present:Unknown cost ratio being 1:5:3. Whereas the cost function for outcome-based classification was such that Abnormal:Normal cost ratio was 10:1.

Training	Validation	Test	Ranking
$\mathbf{0.643 \pm 0.062}^{\mathbf{1}}$	**0.626**	**0.632**	**29/41**
0.600 ± 0.123^{2}	0.460	NA	NA
$0.435 \pm 0.175^{2*}$	NA	NA	NA
0.678 ± 0.180^{3}	0.412	NA	NA
$0.524 \pm 0.117^{3*}$	NA	NA	NA

Table 1. Murmur Detection Task Results. Weighted accuracy metric scores (official Challenge score) for the official test phase. Our final selected entry is in **bold** (team Melbourne Kangas), including the ranking of our team on the hidden test set. We used 10-fold cross-validation on the public training set and one-time scoring on the hidden validation set and hidden test set. [1] Feature-Based Model, [2] CNN Bi-LSTM, [3] CNN LSTM with Attention, * Fusion Results

Training	Validation	Test	Ranking
$\mathbf{11,859 \pm 1,190}^{\mathbf{1}}$	**9420**	**11,735**	**3/39**
$11,958 \pm 2,359^{2}$	11364	NA	NA
$13,897 \pm 2,856^{2*}$	NA	NA	NA
$11,263 \pm 3,232^{3}$	12616	NA	NA
$13,897 \pm 3,919^{3*}$	NA	NA	NA

Table 2. Clinical Outcome Identification Task. Cost metric scores (official Challenge score) for the official test phase. Our final selected entry is in **bold** (team Melbourne Kangas, including the ranking of our team on the hidden test set. We used 10-fold cross-validation on the public training set and one-time scoring on the hidden validation set and hidden test set.[1] Feature-Based Model, [2] CNN Bi-LSTM, [3] CNN LSTM with Attention, * Fusion Results

2.4. Voting Method

The probabilistic outputs were combined using voting methodologies to choose a classification per patient.

- Mean Probabilistic Maximum: All probabilistic outputs were condensed by taking the mean, and a final label was chosen based on the resulting maximum.
- Conditional Mean Probabilistic Maximum: Each probabilistic output was checked for a condition (either murmur present or abnormal outcome). If the condition is met, the final label will be equal to the condition, otherwise, the mean probabilistic maximum is selected.
- Majority Voting: Each probabilistic output is equivocated to a label, and the most frequent label is selected.

3. Results

The CNN LSTM with attention for murmur classification outperformed the feature-based method in cross-validation but with high variation, although had a poor performance on the hidden validation set as shown in Ta-

ble 1. On outcome classification, the feature-based model performed best on both sets as shown in Table 2. For the feature-based model, the best model for murmur classification was RUSBoost with an Absent:Present:Unknown ratio of 2.5:2.5:1 and majority voting. Whereas the best outcome classification model was robust boosting with Conditional Mean Probabilistic Maximum. For both DL and fusion voting methods, the Mean Probabilistic Maximum was found to be most effective. All fusion results show decreased performance. Based on hidden validation set results, the feature-based model performed best and was submitted for evaluation on the hidden test set.

4. Discussion and Conclusions

The DL method showed high variation in cross-validation and the hidden validation set. This variation may be due to under-fitting as the model may be over-simplified. Our LSTM layer contains 16 units, whereas Fernando et al.[3] used 80. Increasing complexity using 2D MFCCs would likely improve performance as it would include the temporal dimension. More fine-tuning is required to increase convergence and decrease the high variation between folds. Mean Probabilistic Output performed best for the DL model likely due to the 10-second segment instead of per recording. The likelihood of classifying as Murmur Present or Abnormal is higher when applying the condition on more recordings.

For feature-based murmur classification, due to the class imbalance, the RUSBoost classifier was most suitable, as shown by superior results. The rationale for the 2.5:2.5:1 ratio performing best is a combination of the murmur present class being the strongest weight in the weighted accuracy, and the unknown class being relatively rare. Therefore, it was appropriate to downsample the absent class to the size of the present class to maximise the training set for each weak learner. For outcome classification, there is minimal class imbalance. Robust boosting avoids over-concentration on a few misclassified observations that can occur in other boosting algorithms, and instead maximises the number of observations with a classification margin above a certain threshold [13]. Conditional Mean Probabilistic Maximum voting method performing the best makes sense as the competition cost function places a strong weighting on correctly identifying abnormal over normal class.

All fusion results show decreased performance compared to individual classifiers. Due to variation in the deep-learning model and fusion, the results are not conclusive.

References

[1] 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. medRxiv 2022;URL https://doi.org/10.1101/2022.08.11.22278688.

[2] 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.

[3] Fernando T, Ghaemmaghami H, Denman S, Sridharan S, Hussain N, Fookes C. Heart sound segmentation using bidirectional lstms with attention. IEEE Journal of Biomedical and Health Informatics 2020;24(6):1601–1609.

[4] Sitaula C, He J, Priyadarshi A, Tracy M, Kavehei O, Hinder M, et al. Neonatal bowel sound detection using convolutional neural network and laplace hidden semi-markov model. IEEE and ACM Transactions on Audio Speech and Language Processing 2022;.

[5] Grooby E, Sitaula C, Fattahi D, Sameni R, Tan K, Zhou L, et al. Noisy neonatal chest sound separation for high-quality heart and lung sounds. arXiv preprint arXiv220103211 2022;.

[6] Grooby E, Sitaula C, Fattahi D, Sameni R, Tan K, Zhou L, et al. Real-time multi-level neonatal heart and lung sound quality assessment for telehealth applications. IEEE Access 2022;1–1.

[7] Abdollahpur M, Ghaffari A, Ghiasi S, Mollakazemi MJ. Detection of pathological heart sounds. Physiological Measurement 2017;38(8):1616.

[8] Kay E, Agarwal A. Dropconnected neural network trained with diverse features for classifying heart sounds. In 2016 Computing in Cardiology Conference (CinC). IEEE, 2016; 617–620.

[9] Bobillo IJD. A tensor approach to heart sound classification. In 2016 Computing in Cardiology Conference (CinC). IEEE, 2016; 629–632.

[10] Homsi MN, Medina N, Hernandez M, Quintero N, Perpiñan G, Quintana A, et al. Automatic heart sound recording classification using a nested set of ensemble algorithms. In 2016 Computing in Cardiology Conference (CinC). IEEE, 2016; 817–820.

[11] Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 2005;27(8):1226–1238.

[12] Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A. RUSBoost: Improving classification performance when training data is skewed. In 2008 19th International Conference on Pattern Recognition. IEEE, 2008; 1–4.

[13] Freund Y. A more robust boosting algorithm. arXiv preprint arXiv09052138 2009;.

Address for correspondence:

Faezeh Marzbanrad
Department of Electrical and Computer Systems Engineering, Monash University. 14 Alliance Lane, Clayton VIC 3800, Australia. Faezeh.Marzbanrad@monash.edu