Heart Murmur Detection and Clinical Outcome Prediction using Multilayer Perceptron Classifier

Kiarash Jalali[1], Mohammad Amin Saket[1], Saman Noorzadeh[1]

[1]Shahid Beheshti University, Tehran, Iran

Abstract

Abnormal sound waves, or heart murmurs in Phonocardiogram (PCG) recordings, are potential indicators of congenital and acquired heart disease in pediatric populations. Detection of these murmurs, and therefore early diagnosis, is usually performed by cardiology specialists. In the 2022 PhysioNet/Computing in Cardiology Challenge, under the team name AKSJ_97BSc, we proposed a Multilayer Perceptron-based model that automatically classifies patient heart murmur status into three categories of present, unknown, or absent based on their metadata and PCG recording. The model also further divides the patients into two outcomes, indicating whether the patient's clinical outcome diagnosed by the medical expert is normal or abnormal. The model was ranked 231 out of 305 submissions with a 0.491 challenge score in the murmur classification category and 161 out of 305 submissions with a 11330.062 challenge score in the outcome classification category.

1. Introduction

Roughly 1% of neonates have congenital heart diseases, which significantly cause morbidity and mortality for several severe disorders [1]. Acquired heart diseases, such as rheumatic fever, are also becoming a serious public health issue in developing countries [2]. One of the common ways to non-invasively detect and diagnose congenital and acquired heart diseases is by studying the heart's mechanical function through Phonocardiogram (PCG) recordings. While this is typically done by cardiac specialists, due to a lack of infrastructure and specialists, especially in developing regions, many patients face difficult challenges in getting an early diagnosis and subsequent treatment. As such, with the data provided by the 2022 PhysioNet/Computing in Cardiology Challenge collected from a pediatric population in northern Brazil, we developed a machine learning model in order to automatically determine whether a patient has heart murmurs or not and potentially help to save more lives with an early diagnosis. In section 2, the model is described along with the features and the method of feature extraction and selection. The training, cross-validation and evaluation phases are then described. In the next section the results are depicted and discussed. Finally, the paper is wrapped up in the conclusions section.

2. Material and Method

2.1. Data

The dataset contains 3163 PCG recordings from 942 patients and their metadata. Each patient has two labels, one that determines their heart murmur status (Present, Unknown or Absent) and the other that shows if the clinical outcome diagnosed by the medical expert is normal or abnormal [3]. The PCG signals are sampled at a rate of 4000 Hz and are typically about 20 to 30 seconds long. There can be up to five PCG recordings for each patient taken from prominent auscultation locations: pulmonary valve (PV), aortic valve (AV), mitral valve (MV), tricuspid valve (TV), and other (Phc). These recordings were collected in a sequential manner by a digital stethoscope.

2.2. Preprocessing

Since the frequency of heart murmurs occur in the 20 to 500 Hz range [4], a fifth-order Butterworth band-pass filter is used in this range for all PCG recordings. Imputation of missing data is done by filling them with the mean value of each column (or feature). Finally, z-score normalization is used to scale the data.

2.3. Feature extraction and selection

The First and one of the essential parts of the algorithm is feature extraction. As well as patient metadata (age, sex, height, weight and pregnancy status), we extracted statistical features from the PCG signals in the time domain (TD), frequency domain (FD), and time-frequency domain (TFD).

In TD, the signal's mean, variance, skewness, and kurtosis were extracted.

In FD, the signal was first transformed by the Fast Fourier Transform (FFT) algorithm, and then the four previously mentioned statistical measures were calculated as FD features. Total Harmonic Distortion (THD), is another feature that has been calculated from the FFT signal and is expected to be higher in murmur signals according to [5].

Another set of features were extracted from the Power Spectral Density (PSD) estimate of the signal using the Welch method. In addition to the four statistical features, the relative power of the signal in the frequency bands 20-130 Hz, 130-400 Hz and 400-500 Hz were extracted as features from the PSD estimate.

Finally, the last set of features were extracted in TFD, specifically with the use of the Wavelet Transform (WT). The Daubechies 7 wavelet was chosen as the mother wavelet and a 5-level decomposition was performed on the PCG signal. The mean and variance of the approximation coefficients and each of the five detail coefficients were calculated as features. In the end we have a total of 151 features for each patient.

After feature extraction, we need to select the features with the best performance for the model. To do this we used a technique called sequential forward floating selection (SFFS) [6]. The algorithm selects one feature with the best performance (after a 5-fold cross-validation) and sequentially adds more features as long as the performance improves. Furthermore, after each iteration, the algorithm removes one feature from the selected pool and checks whether the performance improves or not. This way, we can find the best possible number and combination of features (or at least get very close to them) in a much shorter time compared to testing every possible feature combination. Figures 1 and 2 show the model's performance after performing SFFS going from 1 to 151 features for murmur and outcome classification. The performance metric used in this paper, balanced accuracy, is the arithmetic mean of the recall score for each class. This metric is particularly useful in the case of murmur classification, as it gives each class the same weight regardless of its size and penalized the score if the model is severely biased towards one class.

2.4. Classifier

The classifier is an MLP with four hidden layers with 256, 128, 64, and 32 neurons in each layer, respectively. Admittedly, the number of layers was chosen arbitrarily. However, while the higher number of layers and neurons can potentially cause overfitting, this issue is mitigated by a regularization factor in the classifier, effective feature selection, and a relatively large data set. The MLP classifier uses the "Adam" solver for weight optimization [7]. The activation function is the ReLU function, and the

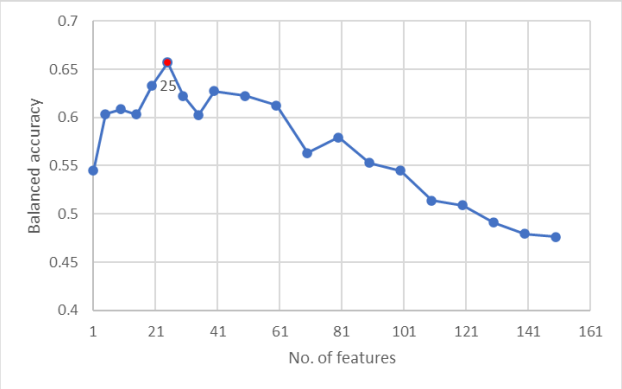

Figure 1. SFFS results for murmur classification. 25 features show the highest performance.

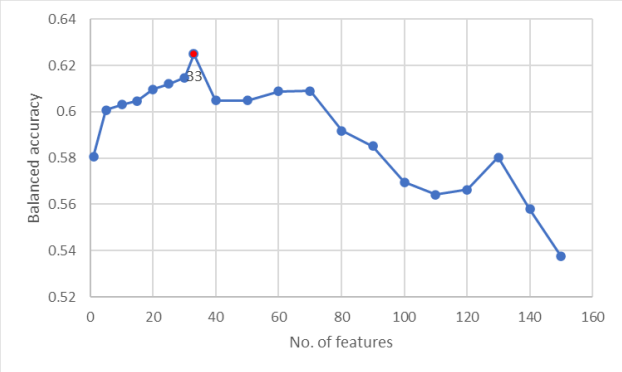

Figure 2. SFFS results from outcome classification. 33 features show the highest performance.

learning rate is 0.001. The output layer uses a softmax function to output the probability of the input belonging to each of the classes.

2.5. Training

As stated before, the challenge dataset is labeled for both heart murmur classification (3 labels) and clinical outcome classification (2 labels). However, the classes in the first case are severely imbalanced, with 695 out of 942 patients belonging to the Absent class, 179 belonging to Present class, and the rest to Unknown class. This could lead to the model being heavily biased towards the majority class. We can see this effect clearly in Table 1, as the recall (averaged after 10-fold cross-validation) for the majority class (Absent) is very high while the recall for the two other classes is extremely low. Therefore, we needed to balance the dataset in order to prevent this issue. One way to do this is through oversampling, but instead of using duplicated samples, by using the Synthetic Minority Oversampling Technique (SMOTE), we can generate new samples based on existing data. The SMOTE algorithm has many variations, and through experimentation, we opted to

use SMOTEENN [8], which uses the Edited Nearest Neighbours (ENN) algorithm to synthesize new samples. Going back to Table 1, the effect of using SMOTEENN is clear. There is a significant increase in the recall of minority classes, and while the recall for the majority class is decreased, this is still a more favorable outcome since the primary function of the model should be detecting the presence of a heart murmur. Thankfully, the issue of imbalance is not present in the two clinical outcome classes, as they are similarly sized to each other. Even so, as our model is shown to be sensitive to data imbalances, we used a random undersampler to even out the size differences.

	Present recall	Unknown recall	Absent recall	Balanced accuracy
Imbalanced dataset	0.2284	0.1619	0.856	0.4154
Balanced dataset with SMOTEENN	0.5035	0.7523	0.4748	0.5769

Table 1. Comparison of the performance between using imbalanced data and balanced data using SMOTEENN.

2.6. Cross-validation

In order to evaluate the performance of the model for our model, we used stratified (to ensure each fold has the same proportion of samples in each class as the whole dataset) 10-fold cross-validation and balanced accuracy (average of the recall for each class) as our performance metric. One thing to note here is that while we used SMOTEENN to generate new samples, we did this only for the training folds and not for the test fold, since evaluating the model's performance with synthetic samples can lead to inaccurate results. Figure 3 shows the final architecture of our model.

3. Results and discussion

Tables 2 and 3 show the final model's results and the PhysioNet challenge score [9] for murmur and outcome classification, respectively. These results are the averaged output of the 10-fold cross-validation. The AUC results for murmur classification are reported for each class by the one-vs-rest strategy. Among 25 features selected by SFFS in heart murmur classification, nine were taken from PSD

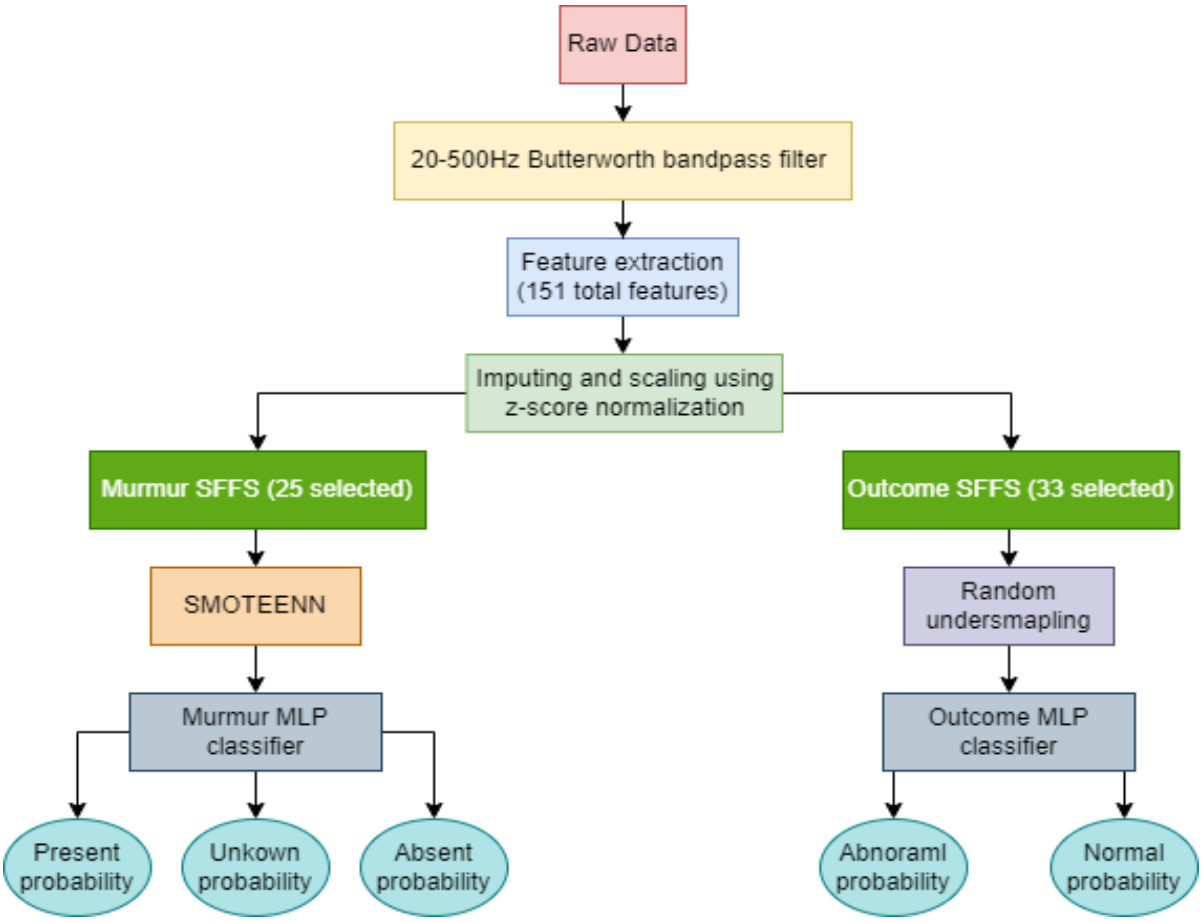

Figure 3. Final model architecture.

estimate, seven from the wavelet coefficients in TFD, and five from the FD. This could indicate that for heart murmur detection, frequency and time-frequency components might lead to more accurate results than TD components. Furthermore, none of the features relating to the patient's metadata were chosen, possibly showing age, gender, height, weight, or pregnancy status might not be effective indicators for heart murmur detection. The same story is mainly true for outcome classification, with only 4 TD features out of 33 being selected by SFFS. Similarly, with the exception of pregnancy status, the other metadata remain unselected.

	Classes		
	Present	Unknown	Absent
AUC	0.645	0.796	0.654
Recall	0.5035	0.7523	0.4748
Balanced accuracy	0.5769		
Challenge score	0.491		

Table 2. Heart murmur classification results.

	Classes	
	Abnormal	Normal
AUC	0.638	0.639
Recall	63.15	60.11
Balanced accuracy	0.6163	
Challenge score	11330.062	

Table 3. Clinical outcome classification results.

Overall, while our approach appears promising, the model's accuracy still leaves much to be desired. One weakness of our model (and possible future research and improvement) is the method we used to extract TD, FD, and TFD features from the PCG recordings. As stated earlier, each of these recordings is about 20 to 30 seconds long and thus, contains several heart cycles. Since heart murmurs typically occur in every cycle, segmenting the PCG recordings so that each segment contains only one heart cycle and extracting features from these segments (rather than on the whole PCG signal as we did in this paper) could lead to finding more murmur-specific components and decreases the amount of noise.

4. Conclusion

In this paper, we described and explored an MLP-based machine learning model for automated heart murmur detection and classification of clinical outcome status. The model was trained with data including patient PCG recordings gathered from a young population in Brazil. We discussed the importance of FD and TFD analysis of PCG signals for heart murmur detection and used a novel oversampling technique to overcome the issue of an imbalanced dataset. With a balanced accuracy of 0.5769 across three labels, the model shows promise and with further investigation and research into its weaknesses, particularly in the feature extraction phase by heart cycle detection and PCG segmentation, the model's implementation in a clinical setting could assist medical staff with congenital and acquired heart disease detection, especially in developing regions where lack of adequate resources and medical specialists prevent patients from receiving early diagnosis and treatment.

Acknowledgments

This work was made possible by PhysioNet, which has provided the data and resources used in this study.

References

[1] Burstein, Danielle S et al. "Significant mortality, morbidity and resource utilization associated with advanced heart failure in congenital heart disease in children and young adults,". American Heart Journal, vol. 209, pp. 9-19, 2019.

[2] S. M. Carvalho et al. "Rheumatic fever presentation and outcome: a case-series report," Revista Brasileira de Reumatologia, vol. 52, no. 2, pp. 241-246, 2012.

[3] J. Oliveira et al. "The CirCor DigiScope Dataset: From Murmur Detection to Murmur Classification" in IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 6, pp. 2524-2535, 2022.

[4] McGee S. Chapter 39 - Auscultation of the Heart: General Principles. In Evidence-Based Physical Diagnosis (Fourth Edition), Elsevier; pp. 327-332, 2018.

[5] Ganguly A, Sharma M. "Detection of Pathological Heart Murmurs by Feature Extraction of Phonocardio-gram Signals.", Journal of Applied and Advanced Research, vol. 2, no. 4, pp. 200–205, 2017.

[6] P. Pudil et al. "Floating search methods in feature slection", Pattern Recognition Letters, vol. 15, no. 1994, pp. 1119-1125, Nov 1994.

[7] Kingma, D.P., Ba, J., "Adam: A Method for Stochastic Optimization",Computing Research Repository, abs/1412.6980, Dec 2014.

[8] Batista, Gustavo et al. "A Study of the Behavior of Several Methods for Balancing machine Learning Training Data.", SIGKDD Explorations, vol.6, pp. 20-29, 2004.

[9] The George B. Moody PhysioNet Challenge 2022, Challenge Scoring section, URL https://moody-challenge.physionet.org/2022/#scoring

Address for correspondence:

Saman Noorzadeh
Shahid Beheshti University, Shahid Shahriari Square, Evin, Tehran, Iran.
saman.noorzadeh@gmail.com