Early Diagnosis of Heart Disease Using Classification And Regression Trees

Amir Mohammad Amiri and Giuliano Armano

***Abstract*— Early diagnosis of heart defects are very important for medical treatment. In this paper, we propose an automatic method to segment heart sounds, which applies classification and regression trees. The diagnostic system, designed and implemented for detecting and classifying heart diseases, has been validated with a representative dataset of 116 heart sound signals, taken from healthy and unhealthy medical cases. The ultimate goal of this research is to implement a heart sounds diagnostic system, to be used to help physicians in the auscultation of patients, with the goal of reducing the number of unnecessary echocardiograms and of preventing the release of newborns that are in fact affected by a heart disease. In this study, 99.14% accuracy, 100% sensitivity, and 98.28% specificity were obtained on the dataset used for experiments.**

I. Introduction

CARDIAC auscultation is widely used by physicians to evaluate cardiac functions in patients and detect the presence of abnormalities. It is however a difficult skill to acquire. Nowadays signals produced by the heart are not only heard using a stethoscope but also observed as phonocardiograms (PCG) on a screen. Phonocardiography is the recording of sonic vibrations of heart and blood circulation.

Heart murmurs are often the first sign of pathological changes of heart valves, and they are usually found during auscultation in primary health care. Two types of murmur can be observed during auscultation: pathological and normal. The latter, in newborns, is called innocent murmur [1].

Newborns often have heart murmurs that mostly originate from normal flow patterns, with no structural or anatomic heart vessel abnormalities and are referred to as innocent murmurs. Conversely, murmurs may be created by abnormal flow patterns in the heart and vessels, resulting from congenital heart disease, including regurgitation or stenosis of heart valves or left to right shunt lesions at the atrial, ventricular, or great arterial levels.

The most common cause of murmurs in newborns is when a specific condition called patent ductus arteriosus (PDA) occurs, which is often detected shortly after birth, most commonly in premature newborns. This is a potentially serious condition in which blood circulates abnormally throughout the ductus arteriosus. In most cases, the only symptom of PDA is a heart murmur, which lasts until the ductus closes on its own, for healthy newborns typically shortly after birth. Sometimes, especially in premature newborns, it may not close on its own, or it may be large and permit too much blood to pass through the lungs, which can place extra strain on the heart, forcing it to work harder and causing a rise in blood pressure in the arteries of the lungs. If this is the case, a medication or, rarely, surgery may be needed to help close the PDA.

Amir Mohammad Amiri and Giuliano Armano are with the Department of Electrical and Electronic Engineering (DIEE), University of Cagliari, Italy (email: amir.amiri@diee.unica.it).

An innocent heart murmur still requires an echocardiogram for reassurance, even though the cost of an echocardiogram is not negligible. The result of this practice is a misallocation of healthcare funds. While it is clearly important to prevent healthy newborn being sent for echocardiogram, it is also important to avoid that a newborn that has a pathological heart murmur is sent home without proper treatment [2].

Prior works performed on heart murmur are concerned with various stages of life, but our work is on newborns only. The classification at this stage is very important because heart murmurs in newborns are more difficult to diagnose [3, 4].

In this paper, we propose a method for automatically classify PCG data. The method utilizes Classification and Regression Trees (CART) to identify pathological murmurs. Feature extraction has been performed in time and frequency such as maximum, peak to peak, Shannon energy and bispectrum, or both (Wigner bispectrum [5]).Notably, features extraction was very effective to improve experimental results. Our results show an accuracy which significantly improves the current state-of-the-art on this specific problem [3, 6]. Improvements have been obtained also in terms of sensitivity and specificity.

II. Methods

A. Pre-processing

Filtering. Pre-processing of heart sounds occurs in two steps: i) filtering and ii) segmentation. Filtering of heart sounds is performed with the goal of removing the unwanted noise. The recording of PCG usually has a sampling frequency higher than 8 kHz. In the event that the environment influences the recording activity, noise is coupled into the PCG. To avoid unpredictable effects brought by noise, filtering becomes important for later processing. Since the main spectrum of first and second heart sound (S1 and S2 respectively) occurs within the range of 200 Hz, the system filters the original heart sound using a 3rd order band-pass Butterworth filter, with cut-off frequencies at 50 Hz and 200 Hz. An electronic stethoscope has been used to record heart sounds, giving rise to a dataset at 44 kHz and subsequently converted to 4k Hz.

Heart murmur segmentation. The second step of preprocessing is a segmentation method aimed at identifying

the heart sound components S1 and S2 and timing interval between them.

Although the detection can also be manual, we used to identify S1 and S2 with an automatic procedure. The segmentation method is based on the timing between high amplitude components. The fact that the time interval that occurs between S1 and S2 (systole) is always less than the one between S2 and S1 (diastole) is the basis for this process.

Heart sound signals still have very complicated patterns, with numerous small spikes that have little impact on diagnosis but may influence the location of S1 and S2. Peak conditioning was performed for the obtained peaks using wavelet transform, which enabled the process of cycle detection.

We used the Wavelet transform based on Complex Morlet Wavelet (CMW) for finding peak locations (see figure 1). CMW is a kind of Wavelet transform, which are a powerful tool in time-frequency analysis for PCG signals (see Fig. 1).

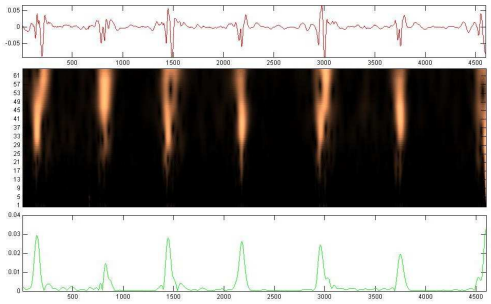

Fig. 1. Peak detection: (top) original signal (middle) scale colored (bottom) coefficients line.

K-means uses an iterative method that minimizes the sum of distances from each object to its cluster centroid, over all clusters. Each class of heart murmurs contains distinctive information in time and frequency domains. This stage involves the extraction of each cardiac cycle of the PCG signal, the formed after the peak detection and conditioning stages. As systolic (S1-S2) and diastolic (S2-S1) murmurs occur within the time intervals that were calculated by the peak conditioning process, these time intervals were clustered into two clusters[7].

Cluster 1 and cluster 2 occur consecutively and indicate a single cardiac cycle. The smaller time interval of each cycle was then identified as systole while the other interval was identified as diastole. After peak detection and condition, cardiac cycle is identified by using k-means, a non hierarchical clustering algorithm in which observation are divided in k mutually exclusive clusters.

We extracted each single cycle of PCG signals using clusters, as shown in Fig. 2 for normal heart sound, heart sound with Systolic murmur and heart sound with Diastolic murmur, respectively.

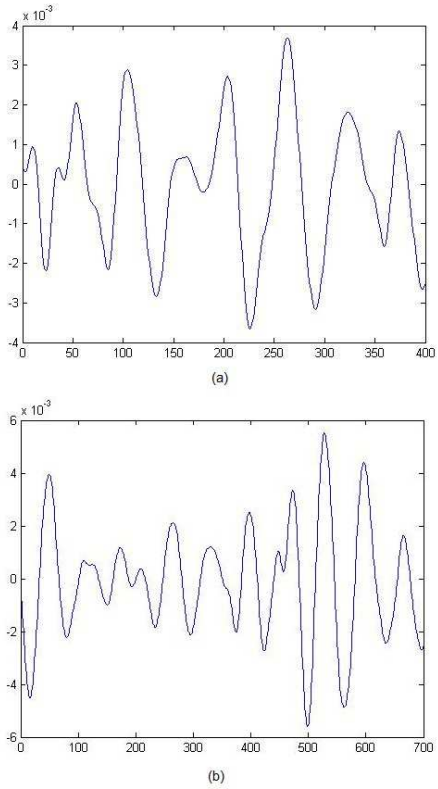

Fig. 2. (a) Normal systolic murmur and (b) normal diastolic murmur.

B. Feature extraction

This phase is focused on extracting signal features that better highlight the properties of the PCG signal, with the goal of identifying those that are more suitable for classification purposes.

This part consists of two major steps: feature extraction and feature selection. In the former step we extracted several features including Maximum value amplitude, Peak to Peak, Energy Shannon, Bispectrum and Wigner bispectrum.

The latter step (i.e.feature selection) was aimed at reducing the size of the feature vector. In particular, we used gains and importance variable in CART to measure the score each of variable.

To calculate the importance score of a variable, CART looks at the improvement measure of each variable, in its role as a surrogate to the primary split. The values of these improvements are summed over each node and are scaled according to the best performing variable [8]. In particular, the variable with the highest sum of improvements is scored 100, while other variables have lower score. Importance variable scores (IVS) are summarized in Table I.

Table I lists all variables used and not used in the tree building process. A score is attached to each variable, and is based on the improvement each variable makes as a surrogate to the primary splitting variable. Variable importance measure highlights variables whose significance is masked or hidden by other variables in the tree building process.

TABLE I

SCORE OF IMPORTANCE VARIABLE

No	Feature	Improvement	IVS
7	Peak to Peak	0.26727	100.000
1	Maximum	0.24242	98.3110
8	Shannon Energy	0.22668	82.6663
9	Bispectrum, C1	0.20649	72.6115
12	Wigner Bispectrum	0.16474	54.7318
4	Absolute Negative Area	0.14060	48.2417
11	Bispectrum, C2	0.03019	43.8223

C. Classification And Regression Trees

Early diagnosis of heart murmurs in newborns is a novel application of CART for clinical and physiological data. CART developed by Breiman et al. (1984), is a nonparametric statistical method that creates binary decision trees. It is a step-by-step process in which a decision tree is constructed by either splitting each node on the tree in two daughter nodes.

The realistic objective of partitioning is to find partitions of the data such that terminal nodes are as such homogeneous as possible. The quantitative measure of node homogeneity is called the impurity function. The simplest idealization of the impurity function is the number of patients who meet an objective criteria divided by the total number of patients in the node. Ratios close to 0 or 1 are considered more pure.

To partition a node, CART examines all possible splits of the explanatory variables. In general, the number of possible splits for ordinal or continuous variables is 1 less the number of distinctly observed values. A potential split is judged by its reduction of the impurity function for both daughter nodes it creates. The partitioning iteratively continues by splitting each node in two daughter nodes and continues until the tree is saturated that is, until no further partitions can be found [9].

The decision tree for predicting heart murmurs is reported in Figure 3. We start at the top of the tree and follow different branches, depending on conditions involving the predictor variables. Trees with multiple layers of splits may be conceptualized as describing interactions between predictor variables. Once we arrive at an end-point of the tree, we used 12 nodes and variables classified in two classes (classes 0 and 1 were Innocent and pathological murmurs respectively [10]).

We calculated the likelihood ratio (LR) to obtain sensitivity and specificity on a tree, defined as follows:

$$LR+ = \frac{sensitivity}{1 - specificity} \quad (1)$$

$$LR- = \frac{1 - sensitivity}{specificity} \quad (2)$$

The interpretation of likelihood ratios is intuitive: the larger the positive likelihood ratio, the greater the likelihood of heart disease; the smaller the negative likelihood ratio, the lesser the likelihood of heart disease.

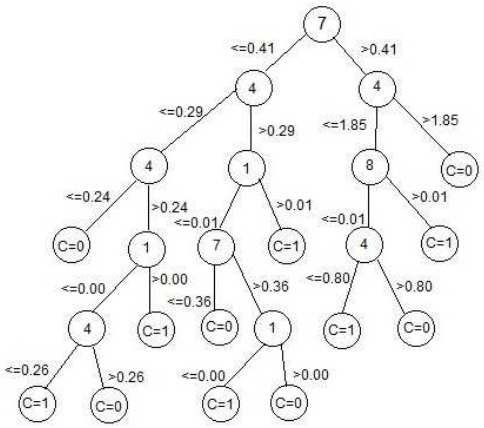

Fig. 3. Illustration of decision tree structure.

III. EXPERIMENTS AND RESULTS

This section reports experimental results and discusses the application of CART to early diagnosis of heart disease in newborns. K-fold cross validation (K=10) has been used as training and test strategy. Experiments have been run using an implementation of CART provided by Salford System Inc, USA.

Results are shown in Table II in the form of a confusion matrix, together with percentage classification accuracy. It can be seen that out of 58 normal signals, 57 were correctly classified as normal, and 1 was misclassified as pathological. As for 58 pathological signals, they were correctly classified as pathological without misclassification. A detailed analysis of the misclassified example showed that it was in fact very difficult to classify, even by human experts.

TABLE II

CLASSIFICATION RESULT OF HEART DISEASE IN NEWBORNS

Actual Group	Normal	Pathological	Percent Correct
Normal	57	1	98.28%
Pathological	0	58	100%
Average/Overall	116		99.14%

Summarizing, 99.14% accuracy, 100% sensitivity and 98.28% specificity were obtained by CART, when used to distinguish between the 116 innocent and pathological heart murmurs in newborns.

Let us point out that, for this system, both high sensitivity and specificity are important. In particular, high sensitivity reduces the number of newborns with innocent murmurs who are identified as pathological murmur and sent to echocardiogram for further testing. More importantly, high specificity reduces the number of newborns with pathological murmurs that are identified as innocent murmurs and have been released with a potentially deadly heart condition.

For each fold, learning has been performed in two steps: growing and pruning. It is worth noting that pruning has been performed provided that decision tree error curve did not trespass the threshold of 1%.

The CART decision tree error curve archived automated growing of a too large tree, followed by automated pruning to find the right-sized tree [11]. The rationale for the growing/pruning process is illustrated in the error curve (Fig. 4).

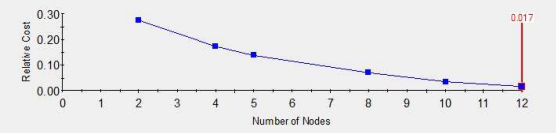

Fig. 4. CART decision tree Error curve.

Fig. 4 shows a curve which outlines the relationship between classification errors and tree size. The scale is always between 0 and 1, so it is called a relative error curve. A tree with a relative error of 0 or nearly 0 is usually too good to be true. The proposed model shows excellent performance for application of diagnosis of heart disease.

In a Receiver Operating Characteristic (ROC) curve for a binary classification problem, the true positive rate (Sensitivity) is reported as function of the false positive rate (100-Specificity) for different cut-off points. ROC curve are reported in Fig. 5a and 5b are targeted for innocent and pathological murmurs, respectively.

A predictive model with perfect performance has an area under ROC curve equal to 1. We obtained, on average, an accuracy of 0.99 the ROC curve highlights the excellent performance of CART to discriminate of heart murmurs.

IV. Conclusions

The proposed methods proposes novelties in both segmentation of heart sound and application of CART. We demonstrated that CART and a suitable data encoding have significant potential for the classification of heart sound data as innocent or pathological murmurs in newborns. Given an unknown heart sound, the system output its classification. This information can be very useful for a physician to decide whether or not to release a newborn or send her/him for an echocardiogram.

This technology is for high-volume screening of newborns suspected of having a heart disease. The software system proposed in this work can be considered the first release of a diagnostic tool able to support physicians in their diagnostic task.

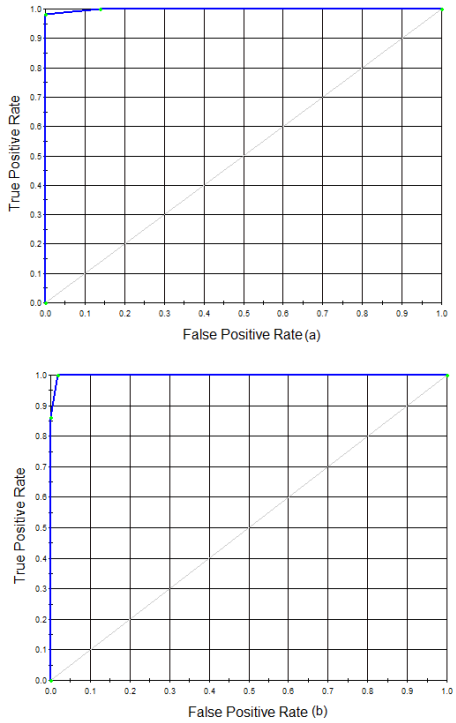

Fig. 5. ROC curve of innocent (a) and pathological murmurs (b) classified.

References

[1] C. Ahlstrom, P. Hult, P. Rask, J.E Karlsson, E. Nylander, U. Dahlstro *"Feature extraction for systolic heart murmur classification"*, Annals of Biomedical Engineering, Vol.34, No. 11, November 2006.

[2] R. Shandas, L. Valdes-Cruz and Roop L. Mahajan. *"Artificial Neural NetworkBased Method of Screening Heart Murmurs in Children"*, Circulation is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231.2001.

[3] S. L. Strunic, F. Rios-Gutierrez, R. Alba-Flores, G. Nordehn, S. Bums. *Detection and Classification of Cardiac Murmurs using Segmentation Techniques and Artificial Neural Networks*, Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2007.

[4] E. Loukis and M. Maragoudakis *Heart murmur identification using random forest in assistive environments* PETRA'10, June 23 - 25, Samos, Greece, 2010

[5] F. Rios-Gutierrez, R. Alba-Flores, S. Strunic *Recognition and Classification of Cardiac Murmurs using ANN and Segmentation*, Electrical Communications and Computers (CONIELECOMP), 22nd International Conference on, IEEE2012.

[6] A.M. Amiri and G. Armano *Diagnosis and classification of systolic murmur in newborns*, The 10th IASTED international Conference on Signal processing, pattern recognition and applications, Innsbruck Austria, 2013.

[7] C. N. Gupta, R. Palaniappan, S. Swaminathan, and S. M. Krishnan, *Neural network classification of homomorphic segmented heart sounds*, Applied Soft Computing 7 286297 (2007).

[8] A. Kumar Banerjee, N.Arora, U.S.N Murty, *Classification and Regression Tree (CART) Analysis for Deriving Variable Importance of Parameters Influencing Average Flexibility of CaMK Kinase Family*, Electronic Journal of Biology, Vol. 4(1):27-33, 2008.

[9] M. Rabinoff, C.M.R. Kitchen, I.A. Cook, and A.F. Leuchter, *Evaluation of Quantitative EEG by Classification and Regression Trees to Characterize Responders to Antidepressant and Placebo Treatment*, The Open Medical Informatics Journal, 5, 1-8, 2011.

[10] H. Zhang and B.H. Singer, *A Practical Guide to Tree Construction*, Springer Series in Statistics, Springer Science Business Media, LLC 2010.

[11] Steinberg, D., and P. Colla. 1995,*CART: Tree-structured non-parametric data analysis*, San Diego, Calif., U.S.A.: Salford Systems.