2020 15th IEEE International Conference on Signal Processing (ICSP) | 978-1-7281-4480-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICSP48669.2020.9320916

Classification of Abnormal Heart Sounds with Machine Learning

Erin B. Evangelista
Department of Engineering
Trinity College
Hartford, CT, USA
erin.evangelista@trincoll.edu

Fabiana Guajardo
Department of Engineering
Trinity College
Hartford, CT, USA
fabiana.guajardo@trincoll.edu

Taikang Ning
Department of Engineering
Trinity College
Hartford, CT, USA
taikang.ning@trincoll.edu

***Abstract*—This paper discusses the utilization of machine learning techniques in separating normal and abnormal heart sounds of public heart sound data sets. The focus of the study is to examine the information value of commonly used heart sound features that characterize heart sound signal of interest. We adopted two information measures, entropy and Gini index. They were utilized to rank the relevance of extracted heart sound features that would assist accurate classification. We examined 26 different heart sound features and ranked them accordingly. Using these two information measures, we have shown that with a minimum of six highest information valued heart sound features, satisfactory classification accuracy can be achieved. Our results were obtained by employing the ensemble decision-tree algorithm in supervised classification with both 80-20 and 90-10 splits cross-validation.**

Keywords—machine learning, classification, information measure, heart sound.

I. Introduction

Heart disease has been the leading cause of death worldwide since the 1920's. Due to its critical importance, cardiac diseases and disorders should be diagnosed at an early stage prior to detrimental and fatal impacts [1]. The most popular and the first diagnosis comes from detecting abnormal heart sounds of a patient. Cardiac auscultation is the main diagnostic method doctors use for detecting heart disease. Since it relies on the use of a stethoscope and the doctor's hearing, the diagnosis provided by cardiac auscultation is purely qualitative. When attempting to detect abnormal heart sounds, such as heart murmurs, what may seem for one doctor to be a soft heart murmur could be a rough heart murmurs for another. Thus, the assessments provided by cardiac auscultation are subjective and sometimes provide an inaccurate diagnosis.

Machine learning and artificial intelligence has boomed in the last three decades as researchers see its potential in detecting patterns that humans may not be able to immediately see and its ability to consistently solve problems. Amongst other health-related research, scientists have explored using machine learning for the diagnosis of heart murmurs using phonocardiograms, or heart sound signals [2, 3, 7-10]. The quality of the diagnosis depends on the quality and quantity of the input data; thus, the higher the quality and quantity of the input data, the more accurate and objective the diagnosis will be.

The underlying study presented the application of machine learning and the decision tree technique–to assist heart sound diagnosis and provide an objective assessment that can expedite detecting abnormal heart sounds with satisfactory accuracy at an early stage prevents the possibility of heart disease. To achieve the goal, heart sound statistical features are extracted from heart sound files [3]. While many features can be extracted from the signal of interest, we are also concerned about the relative feature relevance that is useful to assist classification. The approach we adopted is to rank all extracted features with quantitative information measures and then gradually adding them to the ML decision tree classification to achieve satisfactory detection accuracy [4]. The two information measures are *Entropy* and *Gini Impurity index*. We computed the *entropy gain* and *Gini gain* for each heart sound feature in the feature pool and rank them on two lists, respectively. That is, features are ordered based on their information values to assist determining normal and abnormal heart sounds in decision-tree classification.

This paper makes use of the 2016 PhysioNet Challenge data set (https://physionet.org) of abnormal and normal heart sounds. PhysioNet is an open-source research resource for a vast collection of various physiologic signals. The particular heart sound data set thousands of heart sounds, ranging from 5 to 120 seconds, labeled as abnormal or normal. No description is given as regard to how these heart sound episodes were scored, except the normal and abnormal labeling tags attached to the normal heart sounds came from healthy patients and the abnormal sounds came from patients with pathological diseases. Within the heart sound data, 3,240 sounds were used for training and testing of the decision tree machine learning model while 301 sounds were used for the validation of the model. Within the training data, there were a total of 665 abnormal heart sounds and 2575 normal heart sounds; within the validation data, there were 151 abnormal heart sounds and 150 normal heart sounds.

978-1-7281-4480-1/20/$31.00 ©2020 IEEE

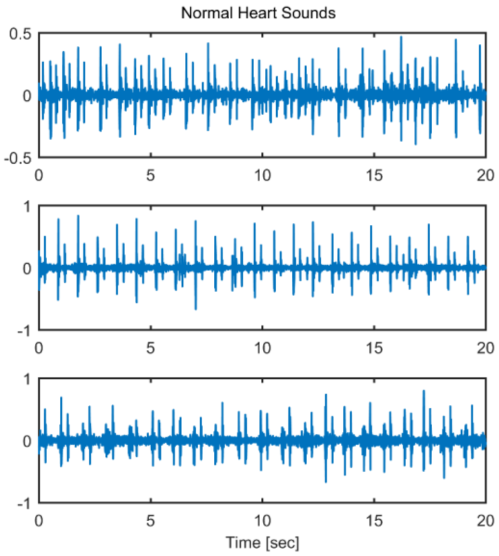

Figure 1. Examples of normal heart sounds

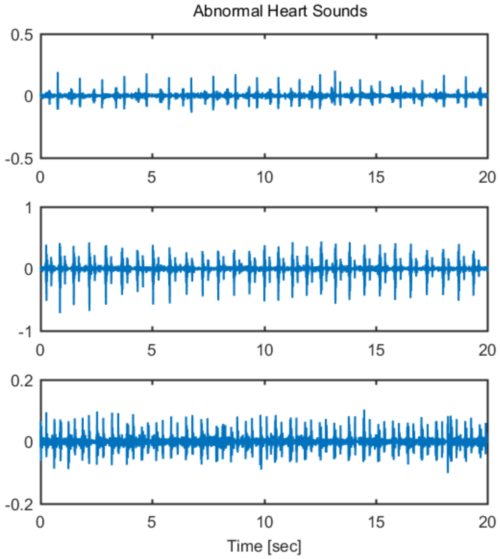

Figure 2. Examples of abnormal heart sounds

In a heartbeat cycle, normal heart sounds consist of two distinct sounds: the first heart sound (S1) comes from the simultaneous closing of the mitral and tricuspid valves; the second heart sound (S2) results from the closing of the pulmonary and aortic valves. The systole is the time between S1 and S2, when the heart contracts and forces blood to flow out of the heart, and the diastole is the time between S2 and S1, when the heart relaxes and blood refills the heart. The time between consecutive S1 (or S2) sounds constitutes one cardiac cycle. The duration of a systole of a normal heart sound ranges from 300 milliseconds to 400 milliseconds, while the duration of diastole is between 500 milliseconds to 600 milliseconds. A typical cardiac cycle generally lasts between 800 to 1000 milliseconds [1]. It's worth to note that a cardiac cycle duration varies from person to person and is affected by the heart when it is measured. Examples of normal heart sounds are shown in Fig.1.

Heart disease may cause irregular heartbeats and abnormal openings and closings of the heart valves. These alternations lead to abnormal turbulent blood flow and result in heart murmurs, which can occur in both systole and diastole. Examples of abnormal heart sounds are given in Fig.2. To avoid misunderstanding, we should be reminded that examples displayed in Fig.1 nor Fig.2 are only small samples of a wide variety of possible heart sound scenarios [5]. In addition to the cardiac cycle and durations of systole and diastole, the appearance of murmurs, clicks, and splitting of S1 or S2, etc. are key features that are indicatives of abnormal heart sounds. Useful heart sound features are required to capture these abnormal deviations.

II. Heart sound features

The default features of the heart sounds from the *PhysioNet* data base were processed using a built-in feature extraction algorithm. A total of 26 features were extracted. They *are mean, median, standard deviation, mean absolute deviation, percent quantiles, median and interquartile range (IQR), skewness, kurtosis, signal entropy, spectral entropy, the dominant frequency, dominant frequency magnitude, dominant frequency ratio*, and *Mel Frequency Cepstral Coefficients (MFCCs)*. MFCCs can accurately represent the shape (envelope) of the short term power spectrum variation. MFCCs have been widely used as the main features in speech and speaker recognition. They are also utilized in our study.

Based on the fact that human's hearing is more sensitive to variations in low pitch frequencies than high frequencies, a reasonable approach to incorporate this varying hearing sensitivity is to adjust heart sound frequency features to match more closely to human's hearing. The Mel scale is an approach to translates the signal frequency (*f* in Hz) to human perceived pitch, Mel Scale (*Mels*) by the following:

$$M(f) = 1125 \ln\left(1 + \frac{f}{700}\right) \qquad (1)$$

Using Mels, we can utilize the extracted heart sound frequency features as physicians normally experienced in cardiac auscultation.

Although there are 26 extracted heart sound features from the PhysioNet, they do not carry the same amount of diagnostic information to classify heart sounds; the large number of features greatly increases the computation burden. In machine learning based classification, a forever endeavor is to find a balance between the required computation burden and acceptable accuracy. Dimensionally reduction is a constant challenge for all ML tasks. In our line of study, we also need to effectively determine the most relevant heart sound features to adopt in a decision-tree algorithm that will allow a compromise between efficient computation and satisfactory accuracy of heart sound classification [6]. While there are many information measures developed for various signals types, we have chosen Entropy and Gini Impurity Index in our study. They are by no means the best information measures but they prove to be effective for the said task.

A. Entropy

Entropy is the measure of variance within a feature. The higher the entropy of a data set, the lower the possibility is of having the data lead to a solid conclusion. In (2), the probability (p_n)of the parent node, or the given classification label, is denoted as p and the number of instances is n.

$$Entropy = -\sum_{n=1}^{N} p_n \log(p_n) \qquad (2)$$

The information gain (IG) determines the information that can be used to expedite the decision tree approach and can be derived from entropy. The I.G. is calculated by finding the difference between the parent entropy in (2) and weighted children entropy.

$$IG = E_{\text{parent}} - \sum_{n=1}^{N}(\text{weighted average}) * E_{\text{child}} \qquad (3)$$

Each feature leads to more one children and the information gain (I.G.) is calculated for each child feature. A feature with higher I.G. can better be more effective to assist decision-tree classification.

B. Gini Index

The Gini Impurity Index (GI) is a measure of how often a randomly chosen element from the set would be incorrectly labeled. Similar to the information gain calculation, the GI index determines an optimal predictor variable, or feature, to split the data from the root node. The Gini impurity index is calculated as follows.

$$Gini\ Index = 1 - \sum_{n=1}^{N} p_n^2 \qquad (4)$$

The Gini Gain (G.G.) uses the Gini impurity index calculation to decide on features which would expedite classification processing while maintaining high accuracy. weighted average of each child node Gini impurity index is subtracted from the parent GI.

$$\text{G.G.} = G_{\text{parent}} - \sum_{n=1}^{N}(\text{weighted average}) * G_{\text{child}} \qquad (5)$$

We used the information gain (I.G.) and Gini index gain (G.G.) measures to generate two rank lists. The top 10 ranked heart sound features in classification are listed, respectively, in Tables I and II.

By taking the most important features, the ensemble decision tree algorithm would be able to classify the heart sounds faster. The top ten is made up of the top five within time domain and the top five within the frequency domain. We compared overall classification accuracy results obtained from the features based on the G.I. list vs. the those of the G.G. list as seen in Table III.

III. Decision Tree classification

The decision tree machine learning technique is a supervised algorithm. Supervised learning presents the final data set from the machine learning program with its corresponding solutions and labels. This allows for the data to be properly identified and checked. The ensemble decision tree technique, also known as random forest, takes a collection of different decision trees to classify the data. A decision tree [6, 11] resembles a flowchart in that each internal node of the tree represents a 'test' on an attribute and each branch represents the outcome of this 'test', and in which each leaf node represents a class label. The technique uses multiple layers of features before reaching a final classification. The entropy and information gain of the features inputted in the decision tree technique can be

TABLE I. Ten Highest Information Gain Features

Features Ranked by I.G.	
Feature Name	***Information Gain***
75% percentile	0.084279
Standard deviation	0.079332
Median value	0.060556
Mean absolute deviation	0.057767
Mean value	0.55789
MFCC 4	0.18457
MFCC 5	0.092784
MFCC 1	0.087917
MFCC 2	0.066604
Spectral entropy	0.062423

TABLE II. Ten Highest Gini Index Features

Features Ranked by Gini Gain	
Feature Name	***Gini Gain***
Signal interquartile range	0.35291
Sample kurtosis	0.34854
Mean value	0.34586
Standard deviation	0.34565
Median value	0.34376
MFCC 10	0.34978
Spectral entropy	0.34745
Dominant frequency value	0.34332
MFCC 8	0.34083
MFCC 6	0.33822

calculated in order to determine which features are best to act as the beginning, upper level nodes of the tree.

In order to train and test the program, 3,547 heart sound files were retrieved from the PhysioNet and use. Within the heart sound data, 3,240 sounds were used for training and testing while 301 sounds were used for the validation of the decision tree machine learning model. Within the training data, there were a total of 665 abnormal heart sounds and 2575 normal sounds. In the validation set, there were 151 abnormal sounds and 150 normal. The heart sounds ranged from five seconds to 120 seconds.

To process the features, an ensemble decision tree algorithm was used for its capability for binary classification and to maintain consistency through a Matlab classification program. The ensemble decision tree algorithm processed all the heart sounds within contained in the training and validation set. As there was less data representative of abnormal heart sounds, the cost for misclassification was set to a ratio of 20 to 1 to compensate for the fewer number of observations in the abnormal class. The decision tree algorithm also tunes the hyperparameters for the model using Bayesian optimization, which is a method that searches for the very minimum of a hyperparameter function. Bayesian optimization was utilized as it results in lower prediction error.

Cross-validation for both an 80-20 split (k-fold = 5) and 90-10 split (k-fold = 10) was also conducted. Both 80-20 and 90-10 cross-validation splits were used to more closely examine the classification performance. If there was a decrease in the accuracy with the 90-10 split, then the program could potentially be overfitting to the training data. In addition, these two were tested to understand which would result in higher classification accuracy for both the abnormal and normal heart sounds.

IV. RESULTS AND DISCUSSION

The features which elicited the largest information gain (I.G.) values are displayed in Table I and the largest G.G. values are displayed in Table II. The first five in the tables come from the time domain and the last five from the frequency domain. From the original dataset with all 26 features, the ten features extracted from the information gain calculation were collected in a new dataset for analysis. As seen in the two tables, the ten features from the information gain are not exactly the same as those obtained from the Gini index gain. Some features do rank on both top-10 lists. They are *mean value, median value, standard deviation*, and *spectral entropy*.

TABLE III. ACCURADY OF CROSS-VALIDATION

Feature number	***80-20 split***		***90-10 split***	
	I.G.	***Gini G.***	***I.G.***	***Gini G.***
6	95.127	82.415	98.093	97.458
7	95.551	94.280	98.093	98.093
8	94.915	94.280	98.305	98.517
9	95.127	96.398	97.881	98.093
10	94.915	96.186	98.729	97.881
26	96.398		98.941	

With supervised learning, decision-tree and random forest analysis were performed. The results are summarized in Table III. The table shows the overall classification accuracy, i.e., the correct classification of abnormal and normal cases divided by the total number of test cases. We evaluated the accuracy performance using 80-20 and 90-10 cross-validation splits for the following situations: (1) using all 26 default features; (2) using different top ranked features in Tables I and II. We also evaluated the accuracy using 6 to 10 to ranked features. The extensive simulation results are summarized in Table III. We verified the accuracy of the trained random forest model against the validation folder from the PhysioNet data. Our results show that the 90-10 cross-validation achieves better results than 80-20 split, i.e., there is no overfitting worry with our approach. The results in Table III demonstrate that with carefully selected heart sound features with relevant information, satisfactory classification accuracy can be achieved even with significantly reduced number of features.

REFERENCES

[1] Noncommunicable Diseases and Mental Health. World Health Statistics 2018: Monitoring Health for the SDGs, Sustainable Development Goals (pp. 7). Geneva: World Health Organization.

[2] Davis, S. Mermelstein, P. "*Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences*." IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. 28 No. 4, pp. 357-366, 1980

[3] A. Hamidah, R. Saputra, T. L. R. Mengko, R. Mengko, and B. Anggoro. "Effective Heart Sounds Detection Method Based on Signal's Characteristics," *2016 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS),* 2016, 1-4.

[4] Noncommunicable Diseases and Mental Health. World Health Statistics 2018: Monitoring Health for the SDGs, Sustainable Development Goals (pp. 7). Geneva: World Health Organization.

[5] A. Gharehbaghi, M. Borga, B. Sjöberg, & P. Ask. "A Novel Method for Discrimination between Innocent and Pathological Heart Murmurs," *Medical Engineering and Physics,* 2015, *37*(7), 674-682.

[6] Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd ed., by Aurélien Géron (2019), O'Reilly Media, Inc.

[7] K. Chen, A. Mudvari, F. G. G. Barrera, L. Cheng, and T. Ning, "Heart Murmurs Clustering Using Machine Learning," Proc. *14th IEEE Int. Conf. Sig. Proc*. pp. 94-98, Beijing, China, Aug.12-16, 2018.

[8] P. R. Hegde, M. M. Shenoy, and B.H. Shekar. "Comparison of Machine Learning Algorithms for Skin Disease Classification Using Color and Texture Features," *Advances in Computing, Communications and Informatics (ICACCI), 2018 International Conference on*, 2018, 1825-828.

[9] K. Bhatia, S. Arora, and R. Tomar. "Diagnosis of Diabetic Retinopathy Using Machine Learning Classification Algorithm," *Next Generation Computing Technologies (NGCT), 2nd International Conference on*, 2016, 347-51.

[10] M. Amrane, S. Oukid, I. Gagaoua, and T. Ensari. "Breast Cancer Classification Using Machine Learning." *Electric Electronics, Comp.r Sci., Biomed. Engineerings' Meeting (EBBT),* 2018, 1-4.

[11] M. I. Jordan, and T. M. Mitchell. "Machine learning: Trends, perspectives, and prospects." *Science*, 2015, 349(6245), 255-260.