Computer Science and Application 计算机科学与应用, 2020, 10(4), 591-600
Published Online April 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.104061

Hans汉斯

Application of Random Forest Algorithm in Heart Sound Classification

Shuping Sun, Xu Zhang, Tingting Huang, Biqiang Zhang, Hao Chen, Bowen Yang, Hui Li

Nanyang Institute of Technology, Nanyang Henan
Email: shp_sun@yeah.net, nit_xu_zhang@163.com

Received: Mar. 10th, 2020; accepted: Mar. 25th, 2020; published: Apr. 1st, 2020

Abstract

The study aims as utilizing the random forest algorithm to classify heart sounds for diagnosing heart diseases. This paper is organized as follows: the heart sounds are firstly collected via a electronic stethoscope and preprocessed based on the wavelets transform, and secondly the short-time Fourier transform-based (STFT), the frequency domain features and time domain feature are defined and extracted to characterize the features of the first and the second heart sound in time-frequency domain. Finally, the random forest algorithm is employed to classify normal and abnormal heart sounds. The performance evaluation is validated by the achieved accuracy of 93.24% for distinguishing between normal and abnormal signals. Therefore, this study can provide an efficient way to discriminate abnormal sounds for the medical workers or patients.

Keywords

Random Forest Algorithm, Heart Sound, STFT, Feature Extraction

随机森林算法在心音分类中的应用研究

孙树平，张　旭，黄婷婷，张弼强，陈　豪，杨博文，李　辉

南阳理工学院，河南　南阳
Email: shp_sun@yeah.net, nit_xu_zhang@163.com

收稿日期：2020年3月10日；录用日期：2020年3月25日；发布日期：2020年4月1日

摘　要

本研究旨在利用随机森林算法对心音进行分类，为心脏疾病的诊断提供依据。本文结构组织如下：首先通过电子听诊器采集心音，然后基于小波变换对其进行预处理；其次，基于短时傅立叶变换定义并提取

文章引用: 孙树平, 张旭, 黄婷婷, 张弼强, 陈豪, 杨博文, 李辉. 随机森林算法在心音分类中的应用研究[J]. 计算机科学与应用, 2020, 10(4): 591-600. DOI: 10.12677/csa.2020.104061

时频域有效宽度以表征第一和第二心音的时频域特征；最后，采用随机森林算法对心音进行分类研究以区分正常和异常心音信号。通过高达93.24%分类精度验证了本系统区分正常与异常心音可行性。因此，本研究可以为医护人员或患者提供一种有效的异常心音鉴别方法。

关键词

随机森林算法，心音，短时傅里叶变换，特征提取

1. 引言

近年来由于居民生活水平的提高，越来越多的居民患有高血压、血脂异常、糖尿病、超重与肥胖等疾病，导致心血管疾病发病率逐年上升。据《中国心血管病报告 2018》概要[1]表述，中国心血管病患病人数已达 2.9 亿，其中农村和城市心血管死亡占疾病死亡的比率分别约为 46%、43%。如图 1 所示。

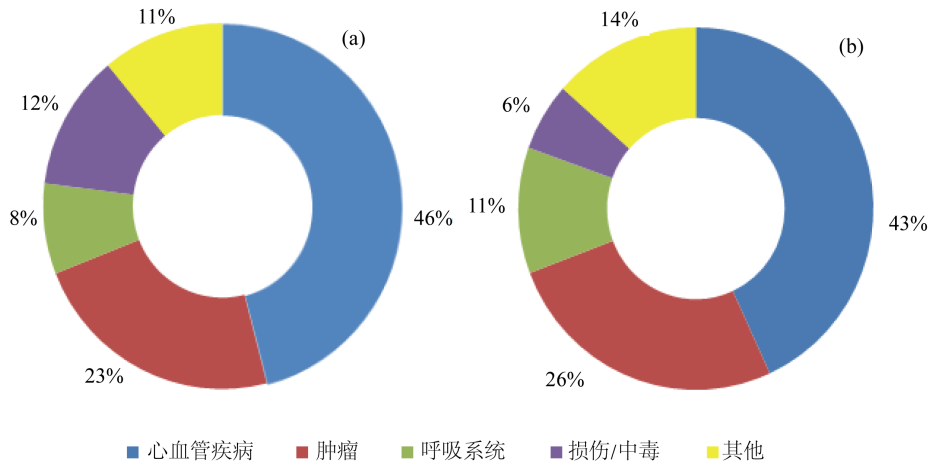

Figure 1. The main cause of death in rural and urban residents in China in 2016
图 1. 2016 年中国农村和城市居民主要疾病死因构成

由图 1 可知心血管疾病是危害居民健康的主要原因之一。心血管疾病在合适的时间发现并得到及时的治疗将会有很高的治愈率，所以在早期诊断出心血管病是非常有益的。由于个体差异、信号噪声的影响、心血管疾病种类繁多等因素，医护人员无法及时准确地对心脏病做出诊断，因此支持心音分类的算法可以极大地促进心脏病的诊断。这可能会导致智能听诊器的发展，智能听诊器可以用很少的设备，有效、经济地检测心脏疾病。目前心音分类的研究主要集中在合适算法的开发上，较为常见分类方法有 K 最近邻(k-Nearest Neighbor，KNN)分类算法[2]，决策树[3]，支持向量机(SVM) [4]，高斯混合模型[5]，人工神经网络(ANN) [6]等。以上常见的分类方法都会存在过拟合且分类准确度不够高的通性。

心音信号性质和检测过程中容易出现噪声，在最近的一项研究中，人们发现基于超声分类算法将杂音分为收缩期杂音、舒张期杂音和连续杂音。以瓣膜狭窄、瓣膜返流、瓣膜功能不全为主的瓣膜缺损杂音多见。但 SVM 和 ANN 对大规模数据训练样本难以实现。由于随机森林算法在任何分类问题中都相对

不容易出现过拟合问题，且对于多种数据，它可以生成准确度较高的分类器。鉴于此，本研究提出基于随机森林算法对心音进行分类研究及应用。

2. 方法

根据研究内容，将研究过程分为以下五个步骤如图 2。

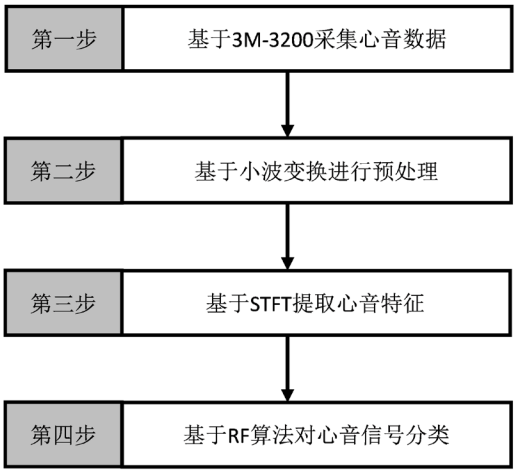

Figure 2. The flow chart of random forest algorithm in heart sound classification
图 2. 随机森林算法在心音分类中的应用流程图

3. 心音的分类研究

心音信号采集预处理

我们采用 3M-3200 电子听诊器对南阳理工学院大学生进行心音数据采集。心音信号分为正常心音信号和异常心音信号两种。正常心音来自 3M-3200 电子听诊器采集的健康成年人的心音信号，表 1 为电子听诊器的功率参数，图 3 是心音信号采集实例图及电子听诊器实物。异常心音[7]来自不同病例，数据库由 20 例正常心音和 80 例异常心音组成。

Table 1. Power parameters of 3M-3200 electronic stethoscope
表 1. 3M-3200 电子听诊器的功率参数

额定最大输出功率 P (W)	间距 d (m)	
	150 kHz~80 MHz	80 MHz~2.5 GHz
		$d = 1.2, \sqrt{P}$
0.01	0.12	0.23
0.1	0.38	0.73
1	1.2	2.3
10	3.8	7.3
100	12	23

Figure 3. 3M-3200 electronic stethoscope
图 3. 3M-3200 电子听诊器

这个阶段的目的是消除杂音和增强心音，可以确保消除噪音和增强相关的心音使它们更容易识别。

小波变换是一种有效的信号处理工具。连续小波变换最适合于信号分析[8] [9]。在研究中，根据心音的频率范围总结，小波分解广泛应用于心音预处理，采样频率为 F_s = 44,100 Hz，采用小波分解方法对心音信号进行滤波，保存 21.5 Hz 以上和 689 Hz 以下的频率分量。在研究[10]中作者检测多贝西小波(DB2-DB10)，哈尔小波、Symlets 小波(Sym2-Sym6)、Coiflet 小波(Coif1-Coif5)和 BiorSplines 小波(Bior1.1-Bior3.3)得出结论，DB10 小波和离散 Meyer 小波给出了心脏声音的最大信噪比(SNR)和最小均方根误(RMSE)。

本研究采用 DB10 作为母波，其波形图如图 4 所示。本研究用分析软件对心音进行降噪，其小波分解图如图 5 所示。图 6 是正常心音信号降噪后的图形。其中蓝色代表的是降噪前的心音信号，左下方蓝色的图是其功率谱图形；相对应的橙色是降噪后的心音信号，右下方橙色的图是其功率谱图形。

心音波形图降噪前后的对比，去掉了一些噪音使得信号的持续时间变得更短、信号变得更密集，功率谱图形降噪后低频率的噪音被消除，功率峰值整体向更高功率方向移动。

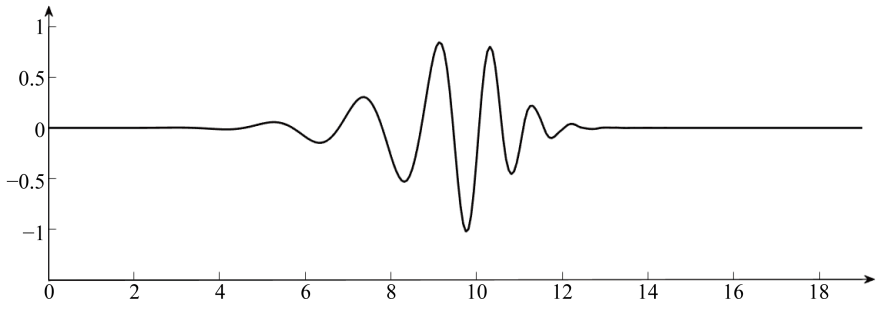

Figure 4. DB10 wavelet
图 4. DB10 小波

4. 随机森林算法应用

4.1. 基于短时傅里叶变换的心音信号特征提取

心音信号特征提取是心音自动解释和心功能障碍诊断的关键环节[11]。特征提取[12]是信号中识别特征属性的重要过程，在心音信号的有效分类中起着重要作用。从全部特征中获得的特征将有助于极大地提高系统的可靠性，提高预测性能。只识别重要性特征可减小随机森林的误差，相对减小过拟合问题。本研究中，基于短时傅里叶变换对第一心音、第二心音进行特征提取。即从时频域信号中提取几个特征，

包括频率宽度 F_w、频率最小值 $F_{\min}$、持续时间 $T_w(s)$。

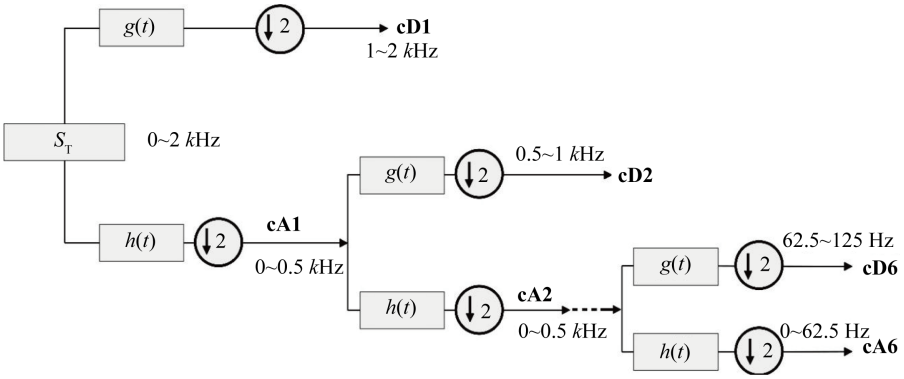

Figure 5. Wavelet decomposition diagram
图 5. 小波分解图

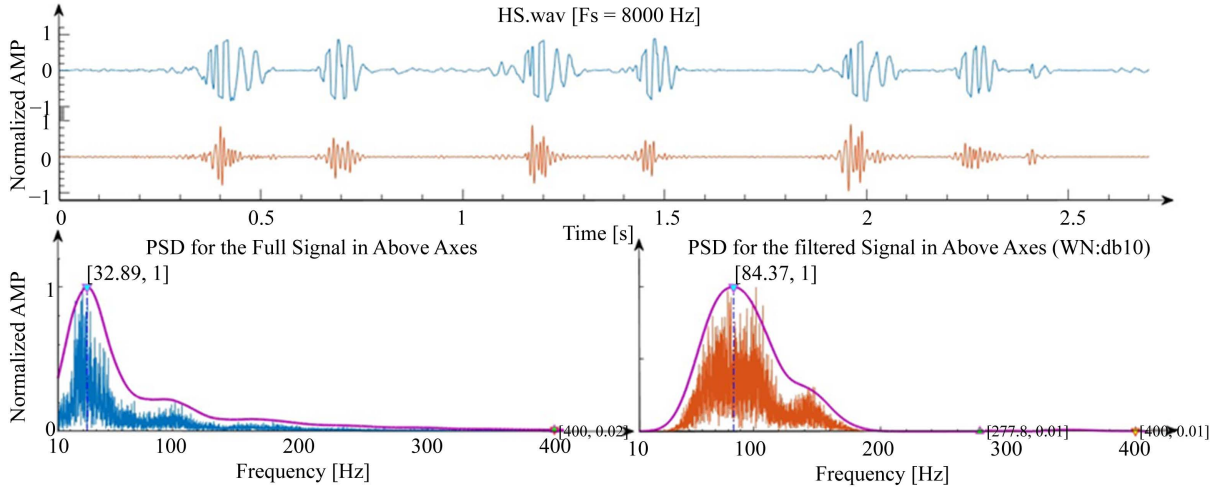

Figure 6. Normal heart sound signal after noise reduction
图 6. 降噪后的正常心音信号

心音信号 $x(t)$ 的短时傅里叶变换定义为[13]：

$$S_F(t,f)=\int_{-\infty}^{+\infty}x(\tau)w'(t-\tau)\mathrm{e}^{-2j\pi f\tau}\mathrm{d}\tau \tag{1}$$

其中 $w(t)$ 是窗口信号，如果其足够窄，可以确保信号在时间 t 内是稳定的。在时域中，窗口信号为 $x_w(t)=x(t)w(t)$。在一定的时间 t，$S_F(t,f)$ 可视为该时刻的频谱。

在实际应用中，连续变换需要离散化，要处理的信号以相等的间隔采样。在研究[14]中，信号 $x(t)$ 的 STFT 离散形式定义为：

$$S_F(t,f)|_{t=m\Delta t f=n/N\Delta t}=S_F(m,n)=\sum_{k=0}^{N-1}x(k\Delta t)w'(k\Delta t-m\Delta t)\mathrm{e}^{-j2nk\pi/N} \tag{2}$$

其中Δt 是采样间隔；N 是采样点的总数；$m,n=0,1,2,\cdots,N-1$。对于合适的窗函数，STFT 频谱能准确地反映时频特性的变化。

时域信号主要是由一个随时间变化的监听设备获得的原始数据。心音信号的采样率为 44.1 kHz。用

于检测心音信号的有效频率范围为 0~2000 Hz。可以直接从时频域信号中提取几个特征，包括频率宽度 F_w、频率最小值 $F_{\min}$、持续时间 $T_w(s)$，如图 7、图 8 所示

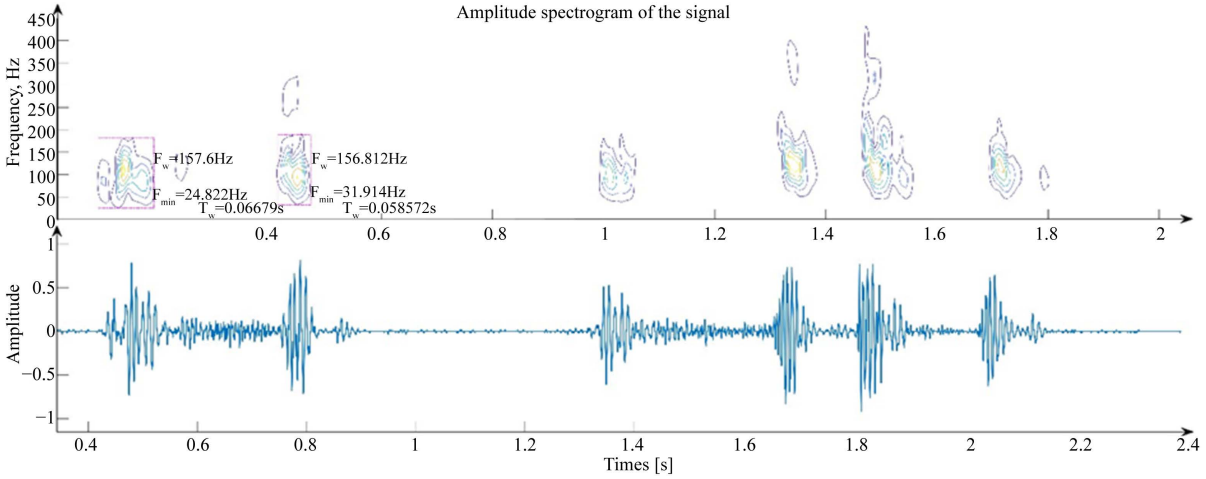

Figure 7. Characteristics of atrial fibrillation cardiac tone signals
图 7. 房颤心音信号的特征

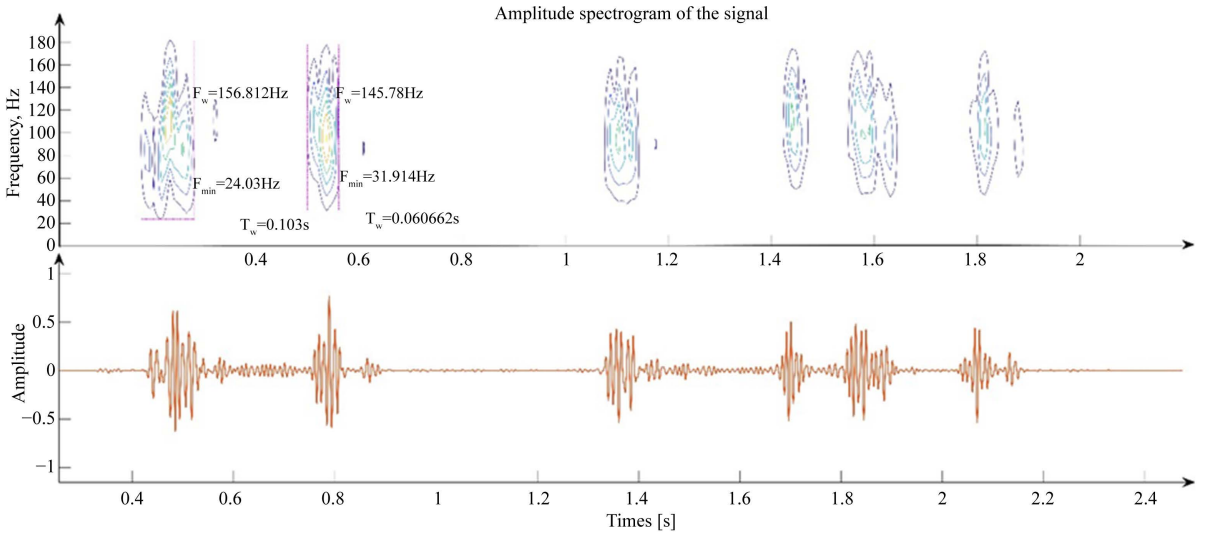

Figure 8. Characteristics of atrial fibrillation cardiac tone signal after noise reduction
图 8. 降噪后房颤心音信号特征

4.2. 基于随机森林算法的心音信号分类

4.2.1. 随机森林算法

在本研究中，通过对心音信号进行预处理得到一个包含 N 个心音特征的验证数据集。心音特征的验证数据集包括第一心音 S1 和第二心音 S2 的频率宽度 F_w、频率最小值 $F_{\min}$、持续时间 $T_w(s)$。将心音信号特征数据集随机划分为 k 个大小相等的子样本。在 k 个子样本中，保留一个单独的子样本作为模型测试的验证数据，其余子样本作为训练数据。交叉验证过程重复 k 次(折叠)，在 k 个子样本中，每个样本都只使用一次作为验证数据。然后对重复交叉验证得到的 k 个结果求平均值，从而产生一个单独的估计。该方法相对于重复随机二次抽样的优点是，所有观测值都同时用于训练和验证，每个观测值只用于验证

一次。通常使用 10 倍交叉验证，如图 9 所示，但一般来说 k 仍然是一个不固定的参数。

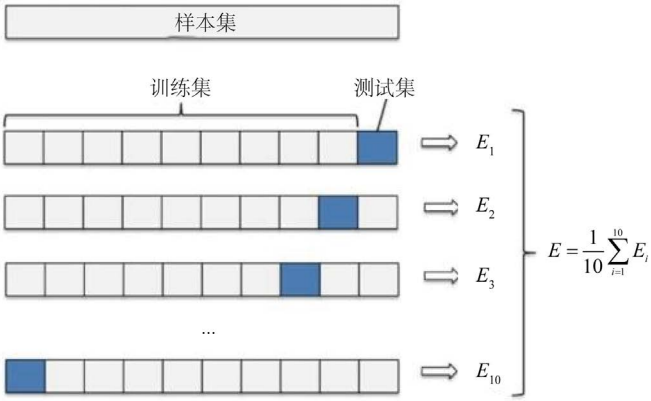

Figure 9. 10 times cross validation
图 9. 10 倍交叉验证

对根节点，树的高度定义为 $H=0$，根节点通过分裂规则分为两个子节点，则这两个子节点对应的高度定义为 $H=1$，以此类推。最大树高 $H_{\max}$ 被设置为方法的超参数或调优参数。它作为一个停止规则，防止算法对数据集进行过度分区，也就是说，树会增长到 $H=H_{\max}$，然后停止。此外，如果节点内部的数据点足够小，则停止树的生长是合理的。设 $N(D_{kl})$ 为节点 D_{kl} 中的数据点的数量。我们将树增长到 $N(D_{kl})\le N_{\min}$，然后停止，其中 $N_{\min}$ 是节点中预定的最小数据点数量，并设置为另一个调优参数。我们结合以上两个停止规则，树停止生长，即 $H=H_{\max}$ 或 $N(D_{kl})\le N_{\min}$。

一棵过于复杂的决策树可决策树进行优化，本研究中采用的能会出现过拟合现象。因此，需要对优化方法为剪枝。选择特征重要性评估值最小的非叶子节点，删除该非叶子节点的左右子节点，若有多个非叶子节点的特征重要性评估值相同小，则选择非叶子节点中子节点数最多的非叶子节点进行剪枝。并计算优化前、优化后及剪枝后决策树的重采样误差和交叉验证误差，将其进行比较。基于随机森林算法的心音分类流程如图 10 所示。

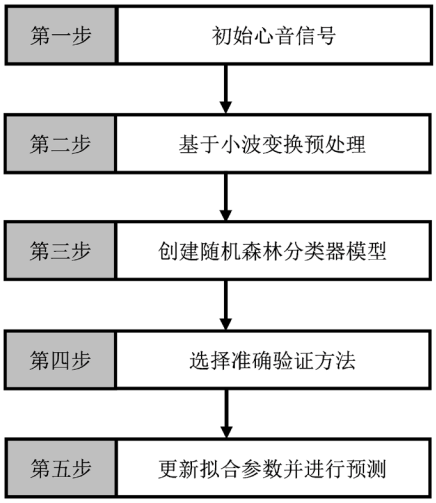

Figure 10. Stochastic forest flowchart
图 10. 基于随机森林算法心音分类流图

4.2.2. 特征重要性评估

随机森林(RF)是 Breiman (2001) [15]开发的一种机器学习算法，包含大量的决策树。而决策树是一个树结构，其每个根节点和子节点都包含有一个测试特征属性，每个叶节点都表示一个类别，既分类结果。

在随机森林生长过程中，可以得到另一种特征重要性测度。在决策树中每个节点 t 分裂是由节点杂质$\Delta R(t)$的减少决定的。节点杂质$\Delta R(t)$为基尼系数。如果节点 t 中有子数据集包含来自 c 类，$gini(t)$的定义为：

$$R(t)=1-\sum_{j=1}^{c}p_j^{\wedge 2} \tag{3}$$

其中 $p_j^{\wedge 2}$ 是 j 类 int 的相对频率。如果 $tint$ 是负偏，$Gini(t)$最小化。节点 t 分裂出两个子节点 t_1 和 t_2，其表示为 $N_1(t)$ 和 $N_2(t)$。分割数据的基尼指数定义为：

$$Gini_{spilt}(t)=\frac{N_1(t)}{N(t)}Gini(t_1)+\frac{N_2(t)}{N(t)}Gini(t_2) \tag{4}$$

特征提供最小 $Gini_{spilt}$ 来选择分割节点。在单一决策树 T_k 中特征重要评分 X_j 为：

$$IS_K(X_j)=\sum_{t\in T_K}\Delta R(t) \tag{5}$$

对随机森林中的所有树 K 进行计算，定义为：

$$IS(X_j)=\frac{1}{K}\sum_{k=1}^{k}IS_k(X_j) \tag{6}$$

4.3. 随机森林分类结果

本研究应用网站采集的离线数据，收集 100 份心音录音，用于训练和测试该系统。其以 DB10 作为母小波，基于小波分解的心音信号预处理；基于短时傅里叶变换的第一心音、第二心音时频分布，结合能量聚集度确定第一心音和第二心音，采用经验窗方法采集心音特征频率宽度 F_w、频率最小值 $F_{\min}$、持续时间 $T_w(s)$；提取得到的特征值，如表 2、表 3 所示。其中，表 2 房颤心音信号提取的特征，表 3 为主动脉瓣回流心音信号提取的特征。在本文中，数据由 100 个样本和 6 个特征组成，其中，训练数据为 80

Table 2. Atrial fibrillation cardiac tone signal extraction
表 2. 房颤心音信号提取

AF02					
S_1			S_2		
F_W	$F_{\min}$	T_W	F_W	$F_{\min}$	T_W
156.812	31.126	0.107	146.568	32.702	0.052
169.420	22.458	0.097	159.176	33.490	0.060
148.932	34.278	0.093	133.960	47.674	0.042
172.572	20.094	0.118	154.448	28.762	0.073
148.144	33.490	0.093	117.412	32.702	0.064
145.780	36.642	0.088	143.416	37.430	0.050
146.568	38.218	0.072	141.052	30.338	0.058
148.932	30.338	0.063	150.508	31.914	0.062
152.084	31.126	0.110	138.324	40.582	0.041
146.568	35.854	0.085	141.840	42.946	0.056

Table 3. Characteristics of heart tone signal extraction from aortic valve regurgitation
表 3. 主动脉瓣回流心音信号提取的特征

DaidoBen_AR_cd1					
S_1			S_2		
F_W	F_{min}	T_W	F_W	F_{min}	T_W
174.936	16.154	0.080	152.084	37.430	0.090
147.356	31.126	0.072	136.324	47.674	0.076
148.144	24.822	0.072	138.688	48.462	0.068
153.660	19.306	0.094	137.112	34.278	0.065
175.724	10.638	0.077	165.480	26.398	0.080
138.688	26.398	0.109	165.480	24.882	0.088
136.329	31.914	0.086	141.840	43.734	0.062
150.508	34.278	0.087	136.324	44.734	0.069
156.812	27.186	0.074	141.052	45.310	0.071
154.478	19.306	0.091	162.328	24.822	0.165
162.382	21.670	0.104	152.872	35.854	0.086

个样本，测试数据为 20 个样本。表 4 为对正常和异常心音信号的识别准确率，其包括精确度和误差分析结果。本文中采用 10 倍交叉验证法，产生 10 个精确度值和误差值，而最终精度为 10 个精度值的平均值，最终结果如下：最终精度 $=(100+100+100+100+100+100+100+100+80+66.66)/10=93.24\%$ 。

Table 4. Accuracy and error
表 4. 精确度及误差

精确度	100	100	100	100	100	100	100	100	100	100	80	66.667
误差	0	0	0	0	0	0	0	0	0	0	20	33.333

5. 结论

本研究针对心音分类，提出一种基于随机森林算法对心音进行分类，将其分为正常心音信号和异常心音信号，结果表明识别精度达到 93.24%。本研究对心音信号进行预处理以及特征提取，采用的软件直接读取音频和 mat 文件；重复选择图形进行分析；同时提取原始信号和降噪后信号的时频特征图。其心音分类结果表明：对于多种数据，随机森林算法可以生成准确度较高的分类器，可以以快速简便的方式为医护人员提供更准确的诊断结果。并且本研究为验证本研究提出方法的有效性，以常见的典型心脏病例房颤、主动脉回流、主动脉狭窄(例)、二尖瓣狭窄和正常心音作为研究对象。其心音分类结果表明：对于多种数据，随机森林算法可以生成准确度较高的分类器，可以以快速简便的方式为医护人员提供更准确的诊断结果。

参考文献

[1] 胡盛寿, 高润霖, 刘力生, 等. “中国心血管病报告 2018”概要[J]. 中国循环杂志, 2019, 34(3): 209-220.

[2] Cover, T.M. and Peter, E.H. (1967) Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, **13**, 21-27. https://doi.org/10.1109/TIT.1967.1053964

[3] King, M.W. and Patricia, A.R. (2014) Data Mining in Psychological Treatment Research: A Primer on Classification and Regression Trees. *Journal of Consulting and Clinical Psychology*, **82**, 895. https://doi.org/10.1037/a0035886

[4] Choi, S. and Jiang, Z. (2010) Cardiac Sound Murmurs Classification with Autoregressive Spectral Analysis and Mul-

ti-Support Vector Machine Technique. *Computers in Biology and Medicine*, **40**, 8-20. https://doi.org/10.1016/j.compbiomed.2009.10.003

[5] Mannini, A. and Angelo, M.S. (2010) Machine Learning Methods for Classifying Human Physical Activity from on-Body Accelerometers. *Sensors*, **10**, 1154-1175. https://doi.org/10.3390/s100201154

[6] Annarumma, M. *et al*. (2019) Automated Triaging of Adult Chest Radiographs with Deep Artificial Neural Networks. *Radiology*, **2019**, Article ID: 180921. https://doi.org/10.1148/radiol.2018180921

[7] Littmann Library. http://www.3m.com/healthcare/littmann/mmm-library.html

[8] Coubes, J.M., Grossmann, A. and Tchanmitchian, P. (1989) Wavelet, Time-Frequency Methods and Phase Space. Springer, Berlin. https://doi.org/10.1007/978-3-642-97177-8

[9] Goupillaud, P., Alex, G. and Jean, M. (1984) Cycle-Octave and Related Transforms in Seismic Signal Analysis. *Geoexploration*, **23**, 85-102. https://doi.org/10.1016/0016-7142(84)90025-5

[10] Ali, M.N., El-Dahshan, E.-S.A. and Yahia, A.H. (2017) Denoising of Heart Sound Signals Using Discrete Wavelet Transform. *Circuits, Systems, and Signal Processing*, **36**, 4482-4497. https://doi.org/10.1007/s00034-017-0524-7

[11] Leng, S., *et al*. (2015) The Electronic Stethoscope. *Biomedical Engineering Online*, **14**, 66. https://doi.org/10.1186/s12938-015-0056-y

[12] 刘翔, 孙静, 赵洋, 等. 基于 MFCC 的心音信号特征提取及识别研究[J]. 电子测量技术, 2018(2): 1-5.

[13] Audone, B., *et al*. (2016) The Short Time Fourier Transform and the Spectrograms to Characterize EMI Emissions. 2016 *International Symposium on Electromagnetic Compatibility-EMC EUROPE*, Wroclaw, Poland, 5-9 September 2016. https://doi.org/10.1109/EMCEurope.2016.7739239

[14] Yeap, Y.M. and Ukil, A. (2016) Fault Detection in HVDC System Using Short Time Fourier Transform. 2016 *IEEE Power and Energy Society General Meeting* (*PESGM*), Boston, MA, 17-21 July 2016. https://doi.org/10.1109/PESGM.2016.7741323

[15] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/A:1010933404324