文章编号：2096-1618(2022)02-0125-06

基于 CNN-LSTM 的柯氏音五时相分类方法研究

秦雷亮， 何培宇， 方安成， 熊 磊， 潘 帆
(四川大学电子信息学院，四川 成都 610065)

摘要：采用柯氏音进行血压测量的人工听诊法是间接测量血压的方法之一。柯氏音有5个时相，分别为弹响音、杂音、拍击音、捂音、消失音。在用人工听诊法时，弹响音开始的第一声对应的袖带压力为舒张压，消失音出现时对应的袖带压力为收缩压。由于孕妇及儿童测量舒张压需以捂音为判断标准，因此，对于不同人群的血压测量而言，识别柯氏音的时相具有重要意义。为精确识别柯氏音五时相，采用 CNN 加 LSTM 的分类模型进行柯氏音五时相的分类。同时为解决过拟合的问题，采用10折交叉验证。结果表明采用 CNN 和 LSTM 的融合模型对柯氏音的5个时相分类的平均准确率为88.69%，相比于单独的 CNN 准确率85.42%和单独的 LSTM 准确率81.39%分别提高了3.27%和7.30%。

关 键 词：柯氏音；CNN；LSTM；10折交叉验证
中图分类号：TP301.6 **文献标志码**：A
doi：10.16836/j.cnki.jcuit.2022.02.002

0 引言

目前中国高血压患者持续增加，据估算有2亿高血压患者。心脏病、脑卒中属于高血压的并发症，高血压成了威胁人们健康的重大疾病[1]。准确测量血压能够了解血压水平、判断血压高低、指导心血管疾病治疗及跟踪病情变化。20世纪初，俄国柯洛特柯夫发现，用袖带绑扎在上臂并将听诊器置于袖带内，然后加压，肱动脉的血管被压瘪，直至血流堵塞后再减压，外压力逐渐降低，可以从听诊器中清晰地听到血液冲开血管并产生与脉搏同步的冲击音，有5种声音的变化，分别为弹响音、杂音、拍击音、捂音、消失音。这个声音被称为柯氏音(Korotkoff Sound)，并用来测量血压。在测量过程中，取心搏产生的能够听到的第一声柯氏音(弹响音)为收缩压，最后一声柯氏音(消失音)为舒张压[2]。但在实际测量中，柯氏音时相与舒张压和收缩压的关系并不是一一对应，例如在测量孕妇和儿童的时候，拍击音向捂音转变的时刻是舒张压[1]。一些研究表明第二时相也可能作为收缩压的测量依据。J Allen等[3]对柯氏音的5个时相频率进行了分析，主要集中在20～300 Hz，然而近几年国内外并没有柯氏音五时相分类的相关研究。随着深度学习的兴起，深度学习作为人工智能中的机器学习的一门分支，在识别图像信息、语音信号等方面取得了重大进展[4]，并且广泛应用于医疗健康领域。深度学习中智能决策在诊断某些疾病的过程中的诊断正确率也比一般医生的水平高，因此越来越多的深度学习方式用于医学诊断及识别。Argha 等[5]采用 Bi-LSTM 将柯氏音分为有声段和无声段；Celle 等[6]从示波器中信号提取特征(袖带压力、柯氏音脉冲能量、柯氏音脉冲能量斜率)使用 GMM-HMM 模型分类估计血压。

因此本文利用卷积神经网络(convolutional neural networks，CNN)[7]和长短时记忆神经网络(long-short term memory，LSTM)[8]对柯氏音信号进行时相分类，先根据脉搏波切分柯氏音信号，将样本转化成语谱图，再将志愿者的整段柯氏音信号样本的语谱图导入含有3个卷积层的卷积神经网络提取特征并尺度变换，最后利用长短时记忆网络对语谱图分类。采用卷积神经网络和长短时记忆网络能够更好地提取频域特征和时相间的时序特征，降低单一特征的时相识别率低的缺点。精确识别五时相也能够协助医生进行血压测量的判断，了解柯氏音时相特征，进而分析形成的原因，对于研究测量血压的医疗设备测量准确率的提高，特别是对测量孕妇及其儿童的血压具有重大意义。

1 研究数据与方法

1.1 研究数据

数据来自临床测试的30个志愿者，志愿者没有任何其他形式的心血管疾病，测试者包括男性13名，女性17名，每人测试9次，共270条数据，每条数据时长

收稿日期：2021-10-11
基金项目：四川省科技计划资助项目(2020YJ0282)

约2 min,每次测试时间间隔为4 min,在进行测试前,每个测试者要求在座椅上休息5 min左右[9-11]。测量步骤:第一步,往袖带里面充气,充气到200 mmHg后以每秒钟2.5 mmHg匀速放气;第二步,通过传音器和压力传感器以2000 Hz的采样频率对柯氏音和袖带压力分别采样保存。表1是研究人员的信息表。

表1 研究人员信息表

信息	平均	标准差
年龄/岁	36	11
身高/cm	169	8
体重/kg	66	13
臂围/cm	27	3

1.2 方法与步骤

图1是采用CNN和LSTM融合算法总体流程图,利用MATLAB制作的GUI[12-13]读取原始柯氏音数据和脉搏波数据。以脉搏波的波峰为切断点对柯氏音信号的每个时相样本进行切分,并做人工时相标注。每个时相被分成多个小样本,将每一段柯氏音生成的语谱图制作成数据集用于提取特征,数据集划分为训练集、测试集和验证集。CNN用于提取语谱图上柯氏音的特征,LSTM处理柯氏音之间的时序信息。

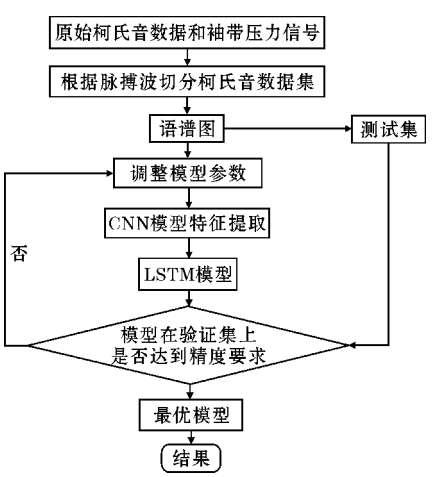

图1 算法流程图

1.2.1 柯氏音信号数据切分

由于脉搏波和柯氏音信号是同步产生,因此根据检测脉搏波的峰值进行切分柯氏音信号。采集的袖带压力信号受噪声的影响需要先进行滤波处理,滤波后的信号为脉搏波。图2为信号滤波前后对比图,对袖带压力信号进行截止频率20 Hz的低通滤波,再进行截止频率0.5 Hz的高通滤波。

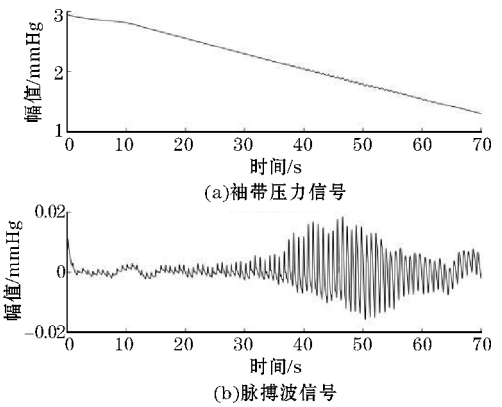

图2 信号滤波前后对比图

图3是柯氏音分割示意图,寻找图3脉搏波的波峰值并记录波峰值的位置,根据波峰的位置在对应的柯氏音样本数据上进行切分,以切分点为中心,左取999个采样点,右取1000个采样点,将柯氏音信号分割成1 s一帧。

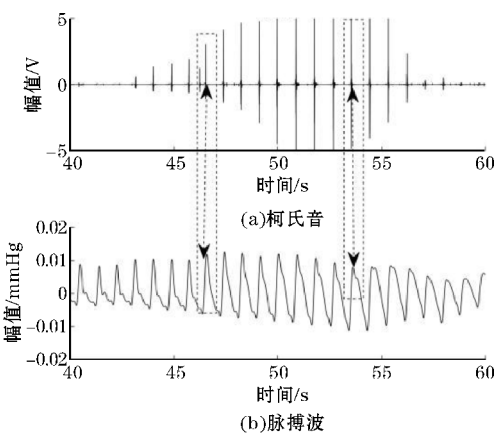

图3 柯氏音分割图

1.2.2 生成语谱图

图4(b)是柯氏音样本,从样本上看各时相的时域图的特征差异并不明显,为了柯氏音信号的特征提取,采用语谱图的方式进行特征提取,先对一维的样本$x(t)$进行加窗、分帧和短时傅立叶变换,则有

$$X(n,f)=\frac{1}{N}\sum_{t=0}^{n-1}h(t)x(t,n)\mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}ft} \tag{1}$$

式中,n为柯氏音样本帧数;f为频率大小;N为窗长大小;t为采样点数;$h(t)$为窗函数,采用汉明窗,表达式为式(2);$x(t,n)$为第n帧的样本。语谱图的表达式为式(3)。

$$h(t)=0.54-0.46\cos\left(\frac{2\pi t}{N}\right)\quad 0\leqslant t\leqslant N \tag{2}$$

$$F(n,f)=20\lg|X(n,f)| \tag{3}$$

通过式(3)可以得到柯氏音样本的语谱图。它直观地表示柯氏音频谱能量随时间的变化。横轴表示时间,纵轴表示频率,一个坐标点代表该时刻频率成分的

能量大小,如图 4(a)是各时相生成的语谱图,柯氏音谐振频率随时间的变化中会在语谱图中呈现出不同程度的能量强弱[14]。

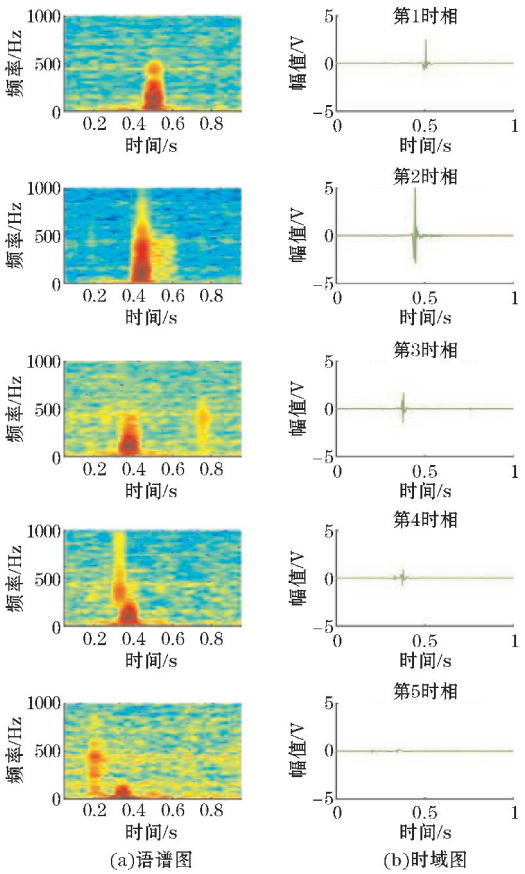

图 4　柯氏音不同时相的语谱图和时域图

1.3　CNN-LSTM 模型

1.3.1　卷积神经网络

CNN 网络模型利用“卷积核”在对图像特征信息提取方面有很高的准确性,与 BP 神经网络各层简单的输入输出相比,CNN 更灵活,能够通过权值共享和偏置降低模型复杂度[15]。CNN 网络是由卷积层和池化层两部分组成。

图 5 是本文 CNN 网络的结构,采用了 3 个卷积层。将数据归一化后生成语谱图输入到卷积神经网络提取特征。

图 5　卷积网络的结构

1.3.2　长短时记忆网络

长短时记忆网络在时间序列数据的处理上有比较好的优势,采用 LSTM 处理柯氏音时序信息,时间步数为 50,LSTM 输出连接到一个全连接层。图 6 是一个 LSTM 细胞单元的内部结构。

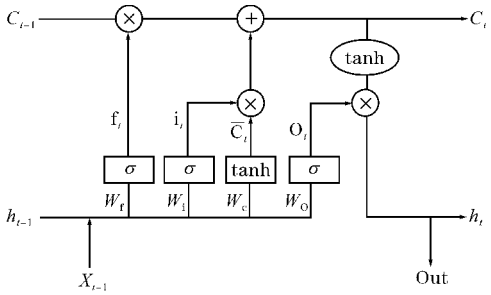

图 6　LSTM 神经单元结构

LSTM 一个细胞单元由遗忘门(f_t)、传输门(i_t)和输出门(o_t)组成。遗忘门表示对 $t-1$ 时刻舍弃信息,输入门是将 t 时刻输入的数据信息添加到记忆信息,输出门表示对当前数据的输出,t 时刻的输入包含了上一时刻 C_{t-1}、x_t 和 h_{t-1},t 时刻输出包含 C_t、O_t 和 h_t,其中 C_t 和 h_t 将作为下一个时刻的两个输入。可以得到式(4)~(9)。

$$f_t=\sigma(\boldsymbol{W}_f[h_{t-1},x_t]+b_f) \tag{4}$$

$$i_t=\sigma(\boldsymbol{W}_i[h_{t-1},x_t]+b_i) \tag{5}$$

$$\bar{C}_t=\tanh(\boldsymbol{W}_C[h_{t-1},x_t]+b_C) \tag{6}$$

$$o_t=\sigma(\boldsymbol{W}_o[h_{t-1},x_t]+b_o) \tag{7}$$

$$C_t=f_t\cdot C_{t-1}+i_t\cdot\bar{C}_t \tag{8}$$

$$h_t=o_t\cdot\tanh(C_t) \tag{9}$$

式(4)~(9)中,$\boldsymbol{W}_f$、$\boldsymbol{W}_i$、$\boldsymbol{W}_C$ 和 $\boldsymbol{W}_o$ 为对应权重矩阵,b_f、b_i、b_C 和 b_o 为对应偏置量,σ 为激活函数[16]。

1.3.3　整体网络结构

柯氏音是语音信号,各时相在时序上的关系紧密相连,因此在提取特征上采用了时相频谱特征和时序特征,两种特征的提取识别相对于单一特征识别率更高,因此采用 CNN 和 LSTM 结合的方式进行分类。图 7是采用的 CNN-LSTM 模型。

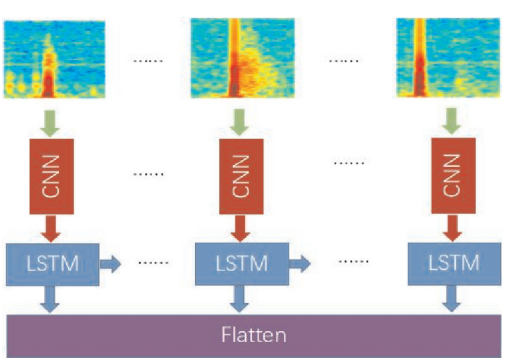

图 7　CNN-LSTM 模型

首先将生成的语谱图以 50 张为一个批次放入三层 CNN 网络进行局部特征提取和尺度变换,初始图片大小为 100×100×3,通过 CNN 网络后变为 12×12×3,然后将数据送入 LSTM 进行时序上的拼接,最后输出到全连接层,通过 SoftMax 函数进行五分类。

2 结果及评估指标

研究的总样本数为 270 条,每个样本 50 张图片,一共训练 13500 张图片。在深度学习中需要大量的数据集进行训练,由于本次实验数据量比较小,故采用 10 折交叉验证。即训练数据量比较小的情况下,将数据集进行多次切割,分成多组训练集和验证集,训练 10 个模型,最后取 10 个模型训练的平均值,验证集数量占总样本的 10%,训练的每一折都进行数据集和验证集的重新分配且训练集和验证集不交叉。从十折交叉验证中得到结果为 5 个时相的分类整体准确率,为区分每一个时相的分类情况,采用了混淆矩阵。混淆矩阵的作用是对分类器分类效果的一种评价指标,将结果显示在一个矩阵里,其中每一列代表每个数据标签的预测时相类别,而每一行代表每个数据标签的真实时相类别。预测数据结果如图 8 所示,第 1 时相 1011 个,预测正确 820 个;第 2 时相 1687 个,预测正确 1414 个;第 3 时相 1030 个,预测正确 693;第 4 时相 992 个,预测正确 522 个;第 5 时相 8629 个,预测正确 8390 个。

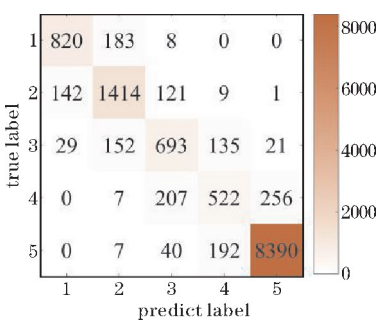

图 8 混淆矩阵

2.1 多指标评价

为客观对分类器的优劣进行评估,采用召回率(recall)、精准度(precision)、准确率(accuracy)、调和平均数(F_β)对分类器进行评估。召回率是指该时相预测准确的数量比上实际该时相的总数(例如图 8 中真实标签为第一时相且预测标签为第一时相的有 820 个,真实标签为第 1 时相总数为 1011,比值为 81.10%);精准度指该时相被预测准确的数量与预测为该时相的总数的比值(如图 8 中第 1 时相预测准确 820 个,预测为第 1 时相总数为 991,比值为82.74%);准确率是指所有时相预测准确的总数与所有数据的比值;调和平均数是用来调和召回率和精准度的一个综合指标。可以得到:

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{10}$$

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{11}$$

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{FN}+\text{TN}} \tag{12}$$

$$F_\beta=(1+\beta^2)\cdot\frac{\text{Precision}\cdot\text{Recall}}{(\beta^2\cdot\text{Precision})+\text{Recall}} \tag{13}$$

以二分类为例,标签 1 为正,标签 0 为负,TP 表示实际标签为正而预测值为正的数量;TN 表示实际标签为负而预测值为负的数量;FP 表示实际标签为负而预测为正的数量;FN 表示实际标签为正而预测为负的数量。召回率和精准度的结果见表 2。通过式(12)得到准确率的值为88.69%。式(13)中,β 的大小是对召回率和精准度的一个偏好,当$\beta>1$时,召回率影响更大,当 $\beta<1$ 时,精准率影响更大,文中 $\beta=1$,即召回率和精准度影响一样大。

表 2 CNN-LSTM 的召回率、精准度、F_1 值 单位:%

	召回率	精准度	F_1
第 1 时相	81.10	82.74	81.91
第 2 时相	83.82	80.20	81.97
第 3 时相	67.28	64.83	66.03
第 4 时相	52.62	60.84	56.43
第 5 时相	97.23	96.80	97.01

3 分析和讨论

3.1 特征选取分析

研究中,时相特征提取是关键的一个环节,柯氏音的时相特征主要集中于声学范畴,包含音色、频率等。单一地在时域提取特征,忽略了时域和频域的相关性,造成特征提取不够细致。因此采用了语音频谱图对柯氏音在时域和频域的可视化表达[17]。

3.2 分类器对比分析

本次研究对 3 个分类器 CNN、LSTM、CNN-LSTM 进行对比。图 9~11 是 3 种模型分别从 F_1 值、精准度和召回率评价指标进行分析对比。从图 9 的 F_1 值对比可以看出 CNN-LSTM 除第 4 时相外的 F_1 值都比

CNN、LSTM 高,第 4 时相 F_1 值略低于 CNN。从图 10 的精准度对比可以看出,CNN-LSTM 各时相的精准度均高于 CNN 和 LSTM。从图 11 的召回率对比可以看出,CNN-LSTM 除第 4 时相外的召回率都比 CNN 和 LSTM 高,第 4 时相明显低于 CNN,同时 CNN 的第 3 时相的召回率却变得异常低,这是由于 CNN 的第 3 时相错分到其他时相数据较多。

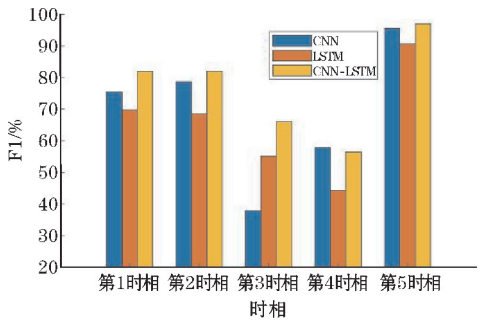

图 9　3 个模型的 F_1 值对比

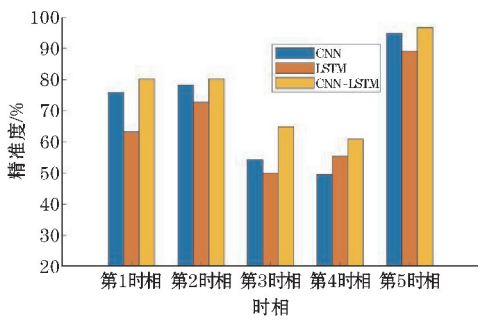

图 10　3 个模型的精准度对比

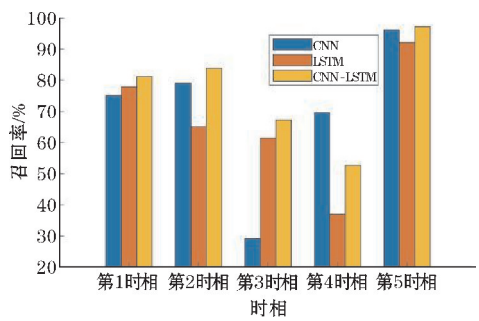

图 11　3 个模型的召回率对比

在神经网络模型选择上,单独的 CNN 对图片特征提取效果显著,单独的 LSTM 能够更好地处理时序信息,因此先进行单独的 CNN 和 LSTM 训练,然后采用融合模型 CNN-LSTM 的深度学习的方法,最后进行模型比较。表 3 是 3 个模型训练的平均准确率和方差的比较,由训练结果可知,CNN-LSTM(88.69%)模型比 CNN(85.42%)模型准确率提高了3.27%,比 LSTM(81.39%)模型准确率提高了7.30%,方差的结果显示模型训练结果的波动性差异不大。

表 3　不同分类器平均准确率和方差的比较

	平均准确率/%	方差
CNN	85.42	0.88
LSTM	81.39	0.95
CNN-LSTM	88.69	0.91

CNN-LSTM 模型的学习训练结果表明 10 折交叉验证的训练和验证达到了一个较好的效果,从混淆矩阵和各项指标每个时相的分类来看第 3 时相和第 4 时相的分类效果比较差。经过讨论和分析,其原因可能在于每个人的柯氏音测量的差异比较明显,部分人的柯氏音第 3 时相和第 4 时相特征差异小,识别率并不高。

4　结束语

文中采用了 CNN 和 LSTM 网络的柯氏音五时相分类模型。先利用 CNN 网络提取柯氏音的局部特征,再利用 LSTM 网络提取时相时序特征,相比单一的 CNN 或 LSTM 分类方法,明显提高了模型的分类准确性。同时所采用的交叉验证的方法,在一定程度上降低了模型过拟合的情况,为后续柯氏音时相的研究及特殊人群的血压测量提供了参考价值。

参考文献:

[1]　王文,张维忠,孙宁玲,等. 中国血压测量指南[J]. 中华高血压杂志,2011,19(12):1101-1115.

[2]　黄青霞,曹云云,程晓萍,等. 二次改良柯氏音法在心血管疾病病人血压测量及管理中的应用研究[J]. 护理研究,2019,33(19):3333-3337.

[3]　J Allen, A Murray. Time-frequencyanalysisKorotkoffof sounds[C]. IEE Colloquium on Time-Frequency Analysis of Biomedical Signals 1997.

[4]　王锡山. 未来医学时代——人工智能诊疗[J]. 中华结直肠疾病电子杂志,2017,6(4):349-352.

[5]　Argha, Celler, Lovell. A Novel Automated Blood Pressure Estimation Algorithm Using Sequences of Korotkoff Sounds[J]. IEEE Journal of Biomedical and Health Informatics,2021,25(4):1257-1264.

[6]　B G Celler, P N Le, A Argha. GMM-HMM-Based Blood Pressure Estimation Using Time-Domain Features[J]. IEEE Transactions on Instrumentation and Measurement,2020,69(6):3631-3641.

[7]　Yin Wenfeng, Yang Xiuzhu, Lin Zhang, et al. ECG

monitoringsystem integrated with IR-UWB radar based on CNN [J]. IEEE Access, 2016, 4(99): 6344-6351.

[8] Latif, Usman, Rana, et al. Phonocardiographic sensing using deep learning for Abnormal heartbeat detectio[J]. IEEE Sensors Journal, 2018, 18(22): 9393-9400.

[9] 胡欣宇,王昕波,赵召龙,等. 基于柯氏音法与示波法结合的新型血压测量系统[J]. 软件,2017,38(3):78-81.

[10] 李刚,王宏,林凌. 基于血压形成原理的血压测量方法及应用[J]. 计量技术,2007(10):16-19.

[11] 张政波,吴太虎. 无创血压测量技术与进展[J]. 中国医疗器械杂志,2003,27(3):196-199.

[12] K Tara, A K Sarkar, M A G Khan, et al. Detection of cardiac disorder using MATLAB based graphical user interface (GUI)[C]. 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC),2017:440-443.

[13] S Y Wong, S Yong Lim. Towards a Livelier Electromagnetic Education with an Interactive MATLAB-based GUI[C]. 2019 IEEE Asia-Pacific Conference on Applied Electromagnetics (APACE),2019:1-4.

[14] 李响,李国正,邓明君,等. 基于语音频谱图像特征的人体疲劳检测方法[J]. 仪器仪表学报,2021,42(2):123-132.

[15] H Yanagisawa, T Yamashita, H Watanabe. A study on object detection method from manga images using CNN[C]. 2018 International Workshop on Advanced Image Technology (IWAIT),2018:1-4.

[16] 熊一橙,徐炜,张锐,等. 基于 LSTM 网络的长江上游流域径流模拟研究[J]. 水电能源科学,2021,39(9):22-24.

[17] ZhangY, DaisS, Song W, et al. Exposing speech resampling manipulation by local texture analysis onspectrogramimages [J]. Electronics, 2020, 9(1):1-16.

Research on Five-phase Classification Method of Korotkoff Sounds based on CNN-LSTM

QIN Leiliang, HE Peiyu, FANG Ancheng, XIONG Lei, PAN Fan

(College of Electronic Information of Sichuan University, Chengdu 610065, China)

Abstract: The manual auscultation method of blood pressure measurement using Ko-sound is one of the methods of indirect blood pressure measurement. Offset sound of Ko-sound has five phases, such as namely, snapping sound, noise, slapping sound, cover sound, and disappearing sound. In the manual auscultation method, the cuff pressure corresponding to the first sound of the buzzing sound is the systolic pressure, and the cuff pressure corresponding to the disappearing sound is the systolic pressure. As for the measure diastolic blood pressure of pregnant women and children, it is necessary to use the muffled sound as the criterion. Therefore, it is of great significance to identify the phase of Korotkoff sounds for blood pressure measurement of different groups of people. In order to accurately identify the five-tempo of the offset, the classification model of CNN with LSTM is used to classify the five-tempo of the offset. At the same time, to solve the problem of over-fitting, 10-fold cross-validation is used. The results show that the fusion model of CNN and LSTM has an average accuracy of 88.69% for the five-phase classification of skew sounds, which is 3.27% higher than the accuracy of CNN alone of 85.42% and 7.30% the accuracy of LSTM alone of 81.39%.

Keywords: Korotkoff sounds; CNN; LSTM; 10-fold cross-validation