基于深度学习Transformer和迁移学习的声诊体质辨识研究*

门韶洋[1]，陈绿洁[1,2]，黄晓梅[3]，文晓冰[1]，林传权[4]，张洪来[1,5**]

（1. 广州中医药大学医学信息工程学院 广州 510006；2. 中山大学生物医学工程学院 深圳 518107；3. 广州中医药大学针灸康复临床医学院 广州 510006；4. 广州中医药大学科技创新中心 广州 510006；5. 广州中医药大学中医证候全国重点实验室 广州 510006）

摘 要：目的 中医体质证型的辨识，在中医“治未病”方面起着重要作用。目前，对于湿热质及平和质的辨识多以问卷形式判定，主观因素影响较大，针对中医湿热质及平和质辨识问题，本文通过语音研究自动化体质辨识任务，为中医体质临床辨识提供辅助。方法 基于深度学习Transformer和迁移学习，设计一种纯粹注意力机制模型，用于中医声诊体质辨识。采集了34名受试者的700条语音，对语音数据预处理得到对应梅尔频谱图，并利用基于公共数据集预训练的Transformer模型来提升模型对音频分类的性能。结果 实验结果准确率为83.33%，曲线下面积（Area under curve，AUC）为92.16%，灵敏度为80.25%，特异性为87.03%，与使用卷积神经网络（Convolutional neural network，CNN）来构建的深度学习模型相比性能更佳。结论 本文湿热质和平和质辨识模型Transformer取得了更优的辨识效果，表明其可提高中医声诊体质识别效率，能够推动体质辨识客观化和智能化发展。

关键词：中医声诊 体质辨识 深度学习 Transformer

DOI: 10.11842/wst.20240920003 CSTR: 32150.14.wst.20240920003 中图分类号：R-058 文献标识码：A

中医学关于体质的论述可以追溯至《黄帝内经》，后世医家也有散在的相关论述，但未成理论体系，近现代关于“中医体质学说”的概念是由王琦[1]提出的：中医体质是指人体生命过程中，由先天遗传和后天获得所形成的，个体在形态结构和功能活动方面所固有的、相对稳定的特性，与心理性格具有相关性。湿热质的人群性格多急躁易怒，平素面垢油光，易生痤疮粉刺，易患疮疖、黄疸、火热等病证[2]。调查发现湿热质与糖尿病前期、肥胖、功能性便秘等密切相关，对某些湿热性疾病有一定的倾向性，对于防治高血脂、糖尿病、肥胖病、中风等病具有重大意义[3]。

然而，目前传统的中医体质辨识主要以问卷调查形式为主，主观性较强且较繁琐，需要具有中医专业知识的医护人员进行操作，使得其操作性受到一定的限制[4]。随着现代科技的发展，人工智能技术在辅助中医临床诊疗方面取得一定的进展[5]，有学者对不同体质人们发出的元音等语音信号研究发现，体质辨别与语音特征有着密切联系，其中语音信号的部分时域特征和频域特征可作为体质辨别的客观指标[6-7]，利用客观语音信号识别体质具有一定优势[8]。随着音频分

收稿日期：2024-09-20

修回日期：2024-11-26

* 科技部“中医药现代化研究”重点专项（2019YFC1710402）：辨证论治四诊延伸性电子化数据采集系统及设备集成，负责人：张洪来；国家自然科学基金委员会专项项目（T2341009）：基于“唾液生物标志物-舌苔微生物-舌象映射脾运化功能”构建2型糖尿病的智能预警模型，负责人：林传权；广东省基础与应用基础研究基金委员会面上项目（2023A1515011316）：基于声诊知识图谱和联合神经网络的多源声音智能识别方法研究及其在中医体质辨识中的应用，负责人：门韶洋。

** 通讯作者：张洪来（ORCID:0000-0003-1283-008X），研究员，博士生导师，博士，主要研究方向：临床疗效中医评价、中医诊疗方法规范化的研究。

类技术和人工智能技术的发展，以深度学习Transformer模型和迁移学习为代表的方法在自然语言处理、语音识别等领域已取得了一定的成效，将其应用于体质辨识任务，有助于辅助临床医师辨识中医体质，同时推动体质辨识客观化、现代化和智能化发展。

1 研究现状

随着人工智能的发展，当利用机器学习和深度学习分析音频时，可得到性能较好的诊断[9-10]。通过数据进行深度学习，实现中医体质辨识的智能化，提高辨识准确度和速度，以便建立客观化的衡量标准。以传统中医理论为基础，结合现代化技术手段，使得声诊研究由定性分析转为定量分析，由主观判断转到客观演绎。在使用语音识别体质方面，孙乡等[11]对50位受试者的语音特征参数作相关性统计和回归分析，发现湿热质大致与音高、音强、语速皆有正相关性，符合实证中的“容易心烦急躁”。穆怀喜[12]通过对体质音频分析发现梅尔倒谱系数和线性预测倒谱系数能在一定程度上区分湿热质及平和质，应用支持向量机（Support vector machine，SVM）和后向传播（Back propagation，BP）神经网络进行分类，Florence等[13]通过100名受试者的韵母和五音发音研究发现平和体质的声学特征符合声调和谐、柔和圆润等，湿热体质的语速加快，为探讨声音与体质提供客观依据。利用深度学习辨识体质方面，Lai等[14]收集了48名受试者的声音，利用残差神经网络（Residual networks，ResNet）对气虚体质和平和体质分类，准确率达81.5%。然而，现有的研究并没有采用深度学习的方法来辨识平和质与湿热质。

音频分类的目标是把音频映射到对应的类别上，性能依赖于音频的特征提取，随着人工智能领域研究的发展，基于深度学习的音频分类取得较大的进展。卷积神经网络（Convolutional neural networks，CNN）在这个领域被广泛运用[15-17]，旨在学习直接映射从声波或频谱图到相应的标签，并通过网络深度和广度设计进一步提高其性能。然而，传统CNN随着输入或梯度信息通过许多层，误差的逆传播过程中会存在梯度消失和梯度爆炸问题，ResNet网络[18]解决了这一问题，而密集卷积网络（Dense convolutional network，DenseNet）[19]在基于ResNet的思想上，采用一种更密集的连接方式，更好地缓解了梯度消失，有效减少了参数数量。注意力机制和卷积模块的集成，可以克服卷积网络局部性的不足[20-21]。然而，基于CNN的模型为获得不同任务的最佳性能，通常需要调优架构，训练时需要大量计算资源，耗时长。最近，纯粹基于自注意力机制的神经网络，Transformer已被证明在各种音频数据集上优于使用CNN构建的深度学习模型分类任务[22-25]。

近年来，音频分类中的迁移学习主要集中在大型的音频数据集上预训练模型，如AudioSet、Million Songs等数据集。Hershey等[26]使用视觉几何群（Visual geometry group，VGG）和ResNet等模型对音频进行分类，在AudioSet训练集上预训练了模型（也称为VGIST）并被应用[27-28]。与此不同的是，我们不仅对海量图像数据集ImageNet迁移学习以代替随机初始化，还在大量音频数据集AudioSet中进行迁移学习。因此本文提出使用Transformer和基于ImageNet和AudioSet预训练的音频信息，直接对语音数据处理，以提升体质辨识效果。

2 基于Transformer预训练模型的体质辨识方法

2.1 模型构建

在本研究中，基于深度学习Transformer对采集的语音数据进行分类[22]。本研究的网络架构如图1所示。在该实验中，将t秒音频的频谱图分割成12[100t-16/10]个大小为16×16的块，在时间和频率维度均重叠6。通过线性投影层将每个16×16的块展平成尺寸为768的一维嵌入特征序列，由于这些信号没有位置信息，因此将尺寸同样为768的可训练位置编码增加到每一个块嵌入特征上，使其能获得二维音频频谱图的空间结构信息。

在一维嵌入特征序列的开始位置附加[CLS]标记，将最终序列输入Transformer[29]中，由于本实验为分类任务，因此仅使用编码部分，将其设置为由12个相同的层组成[30]，为防止梯度消失，加快训练速度，加速收敛，每个层内部均有残差链接以及归一化操作，同时也包含了多头注意力模块以及全连接前馈网络，最后计算相应的映射，得到最终的输出。

2.2 多头注意力机制

自注意力机制对当前位置编码时，会过度集中于自身的位置，因此Transformer提出多头注意力机制。一维特征序列输入至多头注意力层，由多个自注意力

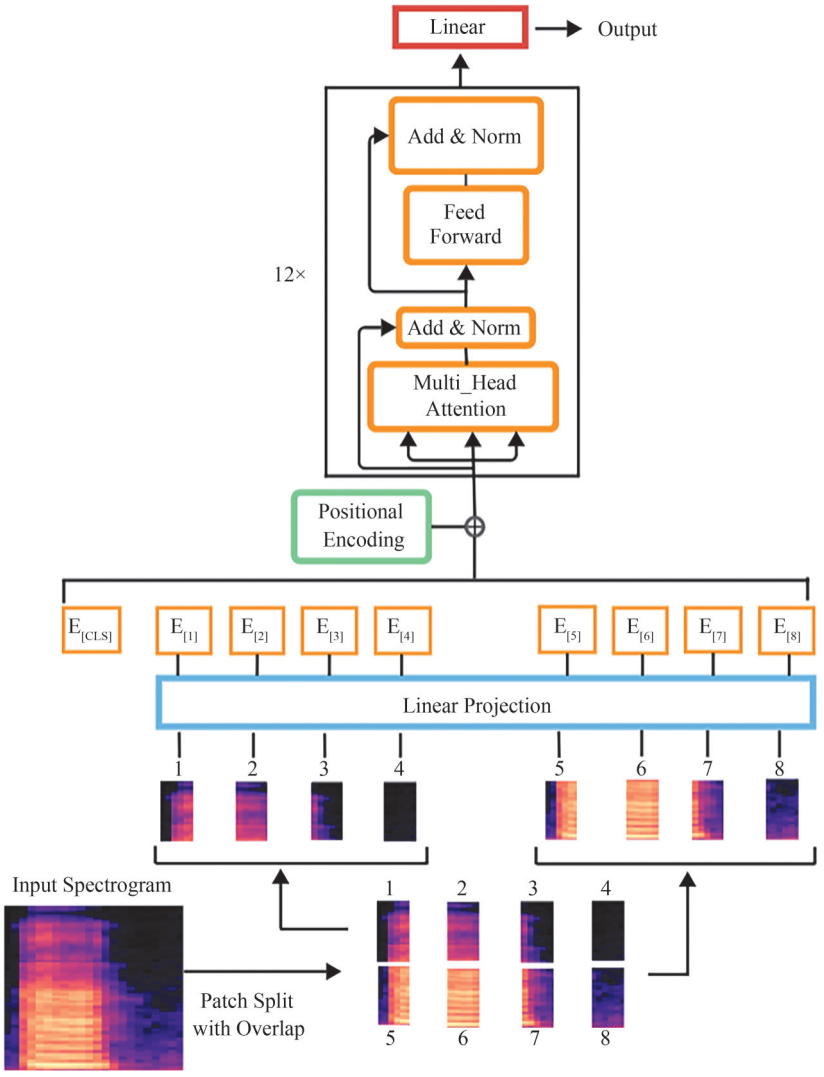

图1 基于音频频谱Transformer网络的体质分类模型

层组成，其输出为每个自注意力层的输出拼接而成，自注意力层获得输入后，会对其做不同的线性变换分别生成查询向量（Query，Q）、键向量（Key，K）和值向量（Value，V）三个矩阵，计算公式如下：

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

其中，d_k为Q，K矩阵的列数，即向量维度。

在本实验中，将head设置为12[31]，即由12个自注意力层组成，如图2所示Q、K、V三个固定值，分别通过一个线性层映射，使用缩放点积注意力（Scaled dot-product attention）评分函数，最后将每个头的输出连接，再映射成和单头一样的输出。由Gong等[22]对比ImageNet预训练Transformer在平衡和完整AudioSet实验以及更改块形状大小消融研究可知，选用ImageNet

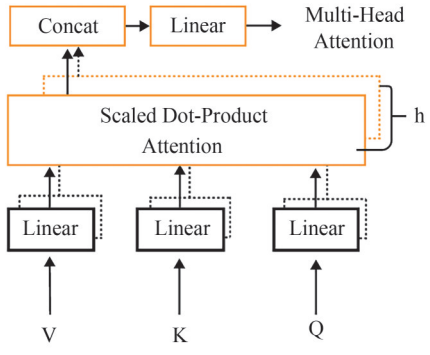

图2 多头注意力层网络结构

预训练的Transformer代替随机初始化的Transformer可降低Transformer对域内音频数据的需求，且选用16×16的块是最优解决方案，因此，本实验采用

ImageNet预训练的Transformer和16×16块。

3 实验数据及预处理

3.1 数据集

所有研究对象均来自广州中医药大学在校学生，纳入标准为受试者年满18岁，且无急性疾病、无急性疼痛、近3个月没服用过任何药物。所有受试者均签署知情同意书，均按照由中华中医药学会发布的《中医体质分类与判定》测评，该量表的9个亚量表的内部一致性系数为0.72-0.82，实用性、再现性、尺度内部一致性的性能评价良好[32]。每一问题按5级评分，计算原始分及转化分，依照标准判定体质类型。本实验共纳入34名受试者录入音频700条，其中平和质16名、湿热质18名，两组受试者平均年龄均为18-23岁。

在文献调查的基础上，本研究收集了两种语音内容。第一种语音内容为“海、云、心、十、口、月、东、泥、方、美”，宋雪阳等[33]将其用于声诊客观化分析。第二种是元音“a,o,e,i,u”，马天才等[34]使用元音来分析中医所定义的虚实证型，雍小嘉等[6]、孙乡等[11]利用元音对比分析体质声音特征。

采集语音过程中，环境噪声小于35 dB，要求受试者放松并以自然舒适的姿势坐着。麦克风型号为AKG HSC271，其信噪比为22 dB，受试者嘴巴与麦克风之间的距离约为5 cm，使用Praat 6.1.321采集声音，采样频率为44100 Hz，通道为单声道。

对湿热质和平和质的具体语音参数进行了描述性分析，从“海”与湿热体质以及平和体质的语音参数结果可以看出，湿热质的第一共振峰均值比平和质小，具体结果见表1。

3.2 数据预处理

梅尔频谱图是用于根据时间、频率和振幅描述声音的特征，主要应用在处理与语音分析相关的任务[35]。其横轴为时间，纵轴为梅尔频率。如图3所示，首先将t秒的音频缓冲到选取窗口长度个样本的帧中，帧与帧之间有样本重叠，每10 ms将25 ms汉明窗应用于每个帧，然后将该帧转由时域转换为频域。将从128维梅尔滤波器组输出的光谱值相加，合并通道产生大小为128×100 t的梅尔频谱图。

将处理后的频谱图随机分成5等份，以便五折交叉验证法验证模型，判断出该语音主人为湿热质或是平和质。图4显示的是不同声音的波形图和频谱图，上侧为平和体质受试者“a”发音的语音波形、频谱图及梅尔谱图，下侧为湿热体质受试者“a”发音的语音波形、频谱图及梅尔谱图。

4 实验与讨论

4.1 损失函数与评价指标

使用交叉熵损失函数（Cross entropy）作为损失函

表1 受试者录音内容“海”语音参数结果

体制	音长	音高	音强	第一共振峰	第二共振峰
平和质	0.43±0.08	134.13±36.80	61.80±3.37	786.21±122.20	1737.07±244.10
湿热质	0.46±0.07	136.56±40.20	61.09±3.54	739.30±90.70	1721.13±174.50

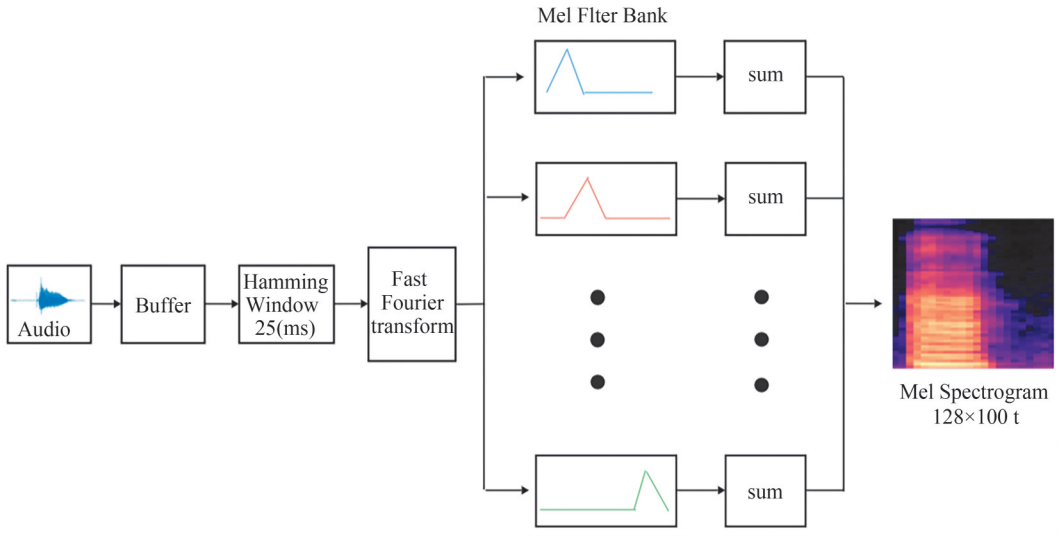

图3 语音预处理过程

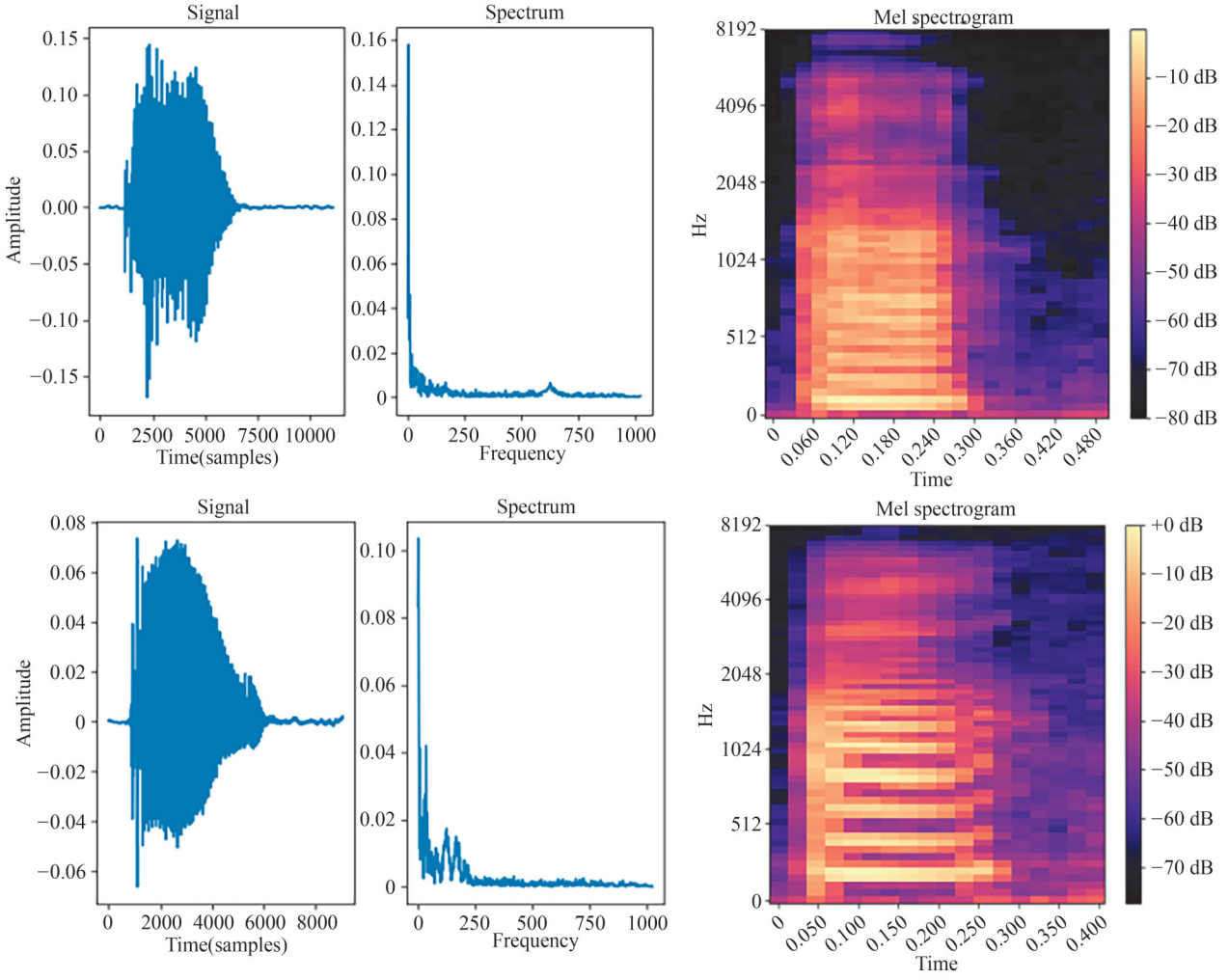

图4 湿热质与平和质的语音波形图、频谱图和梅尔谱图

数，交叉熵损失函数常用于分类问题，交叉熵能够衡量同个随机变量中的两个不同概率分布的差异，交叉熵的值越小，模型预测效果越好，即最小化交叉熵可提高模型的学习精度。

$$H(P,Q) = -\sum_{i=1}^{n} p(x_i)\log(q(x_i)) \quad (2)$$

其中$H(P,Q)$为交叉熵损失函数的值，$p(x_i)$为真实标签，$q(x_i)$为模型预测为正类的概率。

为了定量地对湿热体质和平和体质的分类效果进行评估，使用了4个评价分类指标：接受者操作特性（Receiver operating characteristic，ROC）曲线下面积（Area under curve，AUC）、准确率（Accuracy，Acc）、灵敏度（Sensitivity，Sen）和特异性（Specificity，Spe），相关原理公式如下：

（1）AUC为ROC曲线下的面积，介于0.1和1之间，作为数值可以直观地评价分类器的好坏，值越大越好。ROC曲线的横坐标是假正率，表示为负样本中错误预测为正样本的概率。纵坐标为真正率，表示正样本中预测正确的概率。

（2）Acc为预测正确的样本在所有样本中的比例。

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \quad (3)$$

其中TP为正样本且预测为正，TN为负样本且预测为负，FP为正样本且预测为负，FN为负样本且预测为负。

（3）Sen为所有正例中被分对的比例，衡量了模型对正例（平和质）的识别能力。

$$Sen = \frac{TP}{P} \quad (4)$$

（4）Spe为所有负例中被分对的比例，衡量了模型对负例（湿热质）的识别能力。

$$Spe = \frac{TN}{N} \quad (5)$$

4.2 实验结果与讨论

利用Ray tune提供的网格搜索技术求得最佳超参数[36]，发现模型学习率为1e-5，权重衰减为1e-3为最佳参数，使用了批处理大小为48的Adam优化器。

4.2.1 模型的效果实验

进行了实验比较使用CNN来构建的深度学习模型和Transformer模型对湿热质和平和质音频分类的性能。由于ImageNet预训练的Transformer代替随机初始化权重，所以对ResNet、DenseNet、DenseNet-Attention模型同样使用ImageNet预先训练的权重进行初始化，结果如表2所示，五折交叉验证的平均准确性为82.02%，Transformer模型结构将分类的准确性提高3.04%，且训练时间明显较短。实验证明CNN构建的深度学习模型可以较好地辨识湿热质和平和质，但Transformer模型的表现优于基于CNN构建的模型。

4.2.2 迁移学习效果实验

对比附加公共数据集AudioSet进行预训练的Transformer模型和ResNet、DenseNet、DenseNet-Attention模型后两种体质的分类结果，对比结果见表3，模型的ROC曲线见图5。无AudioSet预训练的Transformer模型记为Transformer-X；有预训练的Transformer模型记为Transformer-Y，图中展示了5个模型的ROC曲线，其中，Transformer-Y的曲线最靠近左上角，表明其分类性能最好，实验中Transformer-Y相比于Transformer-X的分类总体准确率提高了1.31%，证明了附加AudioSet数据集训练的有效性。

表2 不同模型结果对比

Model	AC(%)	AU(%)	Sen(%)	Spe(%)	时间(min)
ResNet	78.98	87.28	77.26	83.96	54
DenseNet	79.92	87.56	77.15	84.66	56
DenseNet-Attention	80.25	88.02	78.05	85.21	52
Transformer	82.02	88.74	78.19	85.65	45

表3 有无附加音频数据AudioSet训练模型性能对比

Model	ACC(%)	AUC(%)	Sen(%)	Spe(%)
ResNet	79.26	88.26	78.35	84.48
DenseNet	80.45	88.93	78.46	84.85
DenseNet-Attention	81.69	89.56	79.21	83.97
Transformer-X	82.02	88.74	78.19	85.65
Transformer-Y	83.33	92.16	80.25	87.03

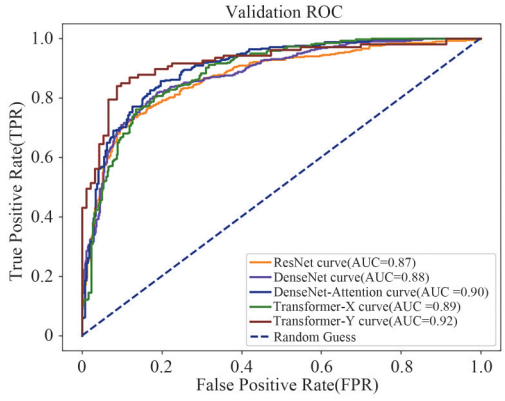

图5 各模型的ROC曲线

5 小结

目前，传统的体质判定大多基于问答的量表，这种方式会受到受试者的主观偏差，同时增加了临床医生的工作强度进而可能影响判定结果。本研究创新性地引入声音客观数据。通过采集声音样本，运用机器学习和深度学习技术，深入挖掘声音中潜在的特征，从而精准判定受试者的体质是平和质还是湿热质。在本文中，提出了基于深度学习Transformer和迁移学习的声诊体质辨识模型。该模型是一个无卷积、纯粹基于注意力机制的模型，其不仅结构简洁明了，并且能够高效地捕捉音频信号中的关键信息，从而在处理音频数据时展现出了卓越的性能，最终对不同体质的声音特征进行准确识别。实验结果表明，所提出的模型在判断湿热质和平和质方面具有较高的准确性，其准确率达到了83.33%。同时，基于灵敏度和特异性指标的分析，也进一步证实了本研究采用的方法能够有效地减少主观因素的干扰，获得相对客观的判定结果。由此表明声音特征作为一种客观指标，在中医体质辨识领域具有巨大的应用潜力，能够快速、高效地实现湿热质和平和质的判定，显著提高辨识效率。

本研究也存在一定的局限性。在人群多样性方面，尚未充分考虑到性别、年龄等因素对声音特征和体质判定结果的影响。不同性别和年龄的人群在发声习惯、生理结构等方面存在差异。此外，在音频数据处理过程中，发现部分音频数据存在冗余现象。针对这些问题，在未来的研究中，将性别、年龄等因素全面纳入数据标准，积极扩大数据集的规模，广泛收集不同类型人群的声音样本，通过这种方式最大程度地排除这些变量对实验结果的不利影响，进一步提高模型分类的准确性和鲁棒性，为中医体质辨识提供更加精准、可靠的技术支撑。

参考文献

1 王琦. 中医体质学. 北京: 中国中医药出版社, 2021:2.

2 王琦. 9种基本中医体质类型的分类及其诊断表述依据. 北京中医药大学学报, 2005, 28(4):1-8.

3 鲁明源. 湿热体质与冠心病——冠心病危险因素的中医评析. 山东中医药大学学报, 2003, 27(1):16-20.

4 卢海莎, 何小丽, 田华, 等. 中医体质辨识研究进展. 中医药管理杂志, 2023, 31(19):202-204.

5 张鸣然. 人工智能在中医体质辨识中的应用展望. 通讯世界, 2019, 26(2):281-282.

6 雍小嘉, 赵刚, 郭佳. 青年与老年气虚及平和体质人群声音特征对比分析. 河南中医, 2018, 38(3):411-413.

7 郑贤月, 梁嵘, 王召平, 等. 平和体质女性的五音频率与年龄的相关性研究. 中国民族民间医药, 2009, 18(3):35-36.

8 Alghowinem S, Goecke R, Wagner M, *et al*. Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors. *IEEE Trans Affect Comput*, 2018, 9(4):478-490.

9 Folland R, Hines E, Dutta R, *et al*. Comparison of neural network predictors in the classification of tracheal-bronchial breath sounds by respiratory auscultation. *Artif Intell Med*, 2004, 31(3):211-220.

10 Sola-Soler J, Jane R, Fiz J A, *et al*. Automatic classification of subjects with and without Sleep Apnea through snoring analysis. *Annu Int Conf IEEE Eng Med Biol Soc*, 2007:6094-6097.

11 孙乡, 杨学智, 李海燕, 等. 成人语音特征与9种体质的相关性研究. 中国中医基础医学杂志, 2012, 18(4):447-449, 454.

12 穆怀喜. 基于中医体质的声象特征研究. 天津: 天津大学硕士学位论文, 2012.

13 ReclusFlorence, 李海燕, 朱庆文, 等. 闻诊研究——成人语音与体质的相关性. 中华中医药学会对外交流与合作分会成立大会暨中医药国际交流合作学术研讨会. 中华中医药学会, 2010.

14 Lai T, Guan Y, Men S, *et al*. ResNet for recognition of Qi-deficiency constitution and balanced constitution based on voice. *Front Psychol*, 2022, 13:1043955.

15 Dong M. Convolutional neural network achieves human-level accuracy in music genre classification. 2018 Conference on Cognitive Computational Neuroscience. September 5-8, 2018. Philadelphia, Pennsylvania, USA. Cognitive Computational Neuroscience, 2018.

16 Guzhov A, Raue F, Hees J, *et al*. ESResNet: environmental sound classification based on visual domain models. 2020 25th International Conference on Pattern Recognition (ICPR). January 10-15, 2021, Milan, Italy. IEEE, 2021:4933-4940.

17 Oord A. v. d, Dieleman S, Zen H, *et al*. Kalchbrenner, SeniorA., and KavukcuogluK., "Wavenet: A generative model for raw audio," arXiv preprint arXiv:1609.03499, 2016.

18 He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition. 2016 IEEE Conf Comput Vis Pattern Recognit CVPR, 2020.

19 Huang G, Liu Z, Van Der Maaten L, *et al*. Densely connected convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). July 21-26, 2017, Honolulu, HI, USA. IEEE, 2017:2261-2269.

20 Wang Q, Wu B, Zhu P, *et al*. ECA-net: efficient channel attention for deep convolutional neural networks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 13-19, 2020, Seattle, WA, USA. IEEE, 2020:11531-11539.

21 Hu J, Shen L, Albanie S, *et al*. Squeeze-and-excitation networks. *IEEE Trans Pattern Anal Mach Intell*, 42(8):2011-2023.

22 GONG Y, CHUNG Y A, GLASS J. AST: audio spectrogram transformer. Interspeech 2021. ISCA, 2021.

23 Koutini K, Schlüter J, Eghbal-Zadeh H, *et al*. Efficient training of audio transformers with patchout 2021: 2110.05069. https://arxiv.org/abs/2110.05069v3.

24 Chen K, Du X, Zhu B, *et al*. HTS-AT: a hierarchical token-semantic audio transformer for sound classification and detection. 2022: 2202.00874. https://arxiv.org/abs/2202.00874v1.

25 Gong Y, Khurana S, Rouditchenko A, *et al*. CMKD: CNN/transformer-based cross-model knowledge distillation for audio classification. 2022: 2203.06760. https://arxiv.org/abs/2203.06760v1.

26 Hershey S, Chaudhuri S, Ellis D P W, *et al*. CNN architectures for large-scale audio classification. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). March 5-9, 2017, New Orleans, LA, USA. IEEE, 2017:131-135.

27 Xie H, Virtanen T. Zero-shot audio classification based on class label embeddings. 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). October 20-23, 2019. New Paltz, NY, USA. IEEE, 2019.

28 Shi L, Du K, Zhang C, *et al*. Lung sound recognition algorithm based on VGGish-BiGRU. *IEEE Access*, 2019, 7:139438-139449.

29 Vaswani A, Shazeer N, Parmar N, *et al*. Polosukhin. Attention is all you need. In NIPS, 2017.

30 DosovitskiyA, BeyerL, KolesnikovA, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR, 2021.

31 Touvron H, Cord M, Douze M, *et al*. Training data-efficient image transformers & distillation through attention. 2020: 2012.12877. https://arxiv.org/abs/2012.12877v2.

32 朱燕波, 王琦, 薛禾生, 等. 中医体质量表性能的初步评价. 中国临床康复, 2006(3):15-17.

33 宋雪阳, 许朝霞, 王寺晶, 等. 121例肺结节患者的语音共振峰初探. 世界科学技术-中医药现代化, 2019, 21(12):2904-2908.

34 马天才, 庄燕鸿, 颜建军, 等. 基于小波包声音信号分析技术的中医虚实证声诊特征分析. 咸阳: 中华中医药学会中医诊断学分会第十次学术研讨会论文集, 2009:6.

35 Sueur J. Mel-frequency cepstral and linear predictive coefficients. Sound Analysis and Synthesis with R. Cham: Springer International Publishing, 2018:381-398.

36 Liaw R, Liang E, Nishihara R, *et al*. Tune: a research platform for distributed model selection and training. 2018: 1807.05118. https://arxiv.org/abs/1807.05118v1.

Research on Sound Diagnosis Constitution Identification Based on Deep Learning Transformer and Transfer Learning

MEN Shaoyang[1], CHEN Lyujie[1,2], HUANG Xiaomei[3], WEN Xiaobing[1], LIN Chuanquan[4], ZHANG Honglai[1,5]

(1. School of Medical Information Engineering, Guangzhou University of Chinese Medicine, Guangzhou 510006, China; 2. School of Biomedical Engineering, Sun Yat-Sen University, Shenzhen 518107, China; 3. Clinical Medical College of Acupuncture Moxibustion and Rehabilitation, Guangzhou University of Chinese Medicine, Guangzhou 510006, China; 4. Science and Technology Innovation Center, Guangzhou University of Chinese Medicine, Guangzhou 510006, China; 5. State Key Laboratory of Traditional Chinese Medicine Syndrome, Guangzhou University of Chinese Medicine, Guangzhou 510006, China)

Abstract: Objective The identification of TCM constitution plays an important role in "treating and preventing diseases" of TCM. At present, the identification of damp-heat constitution and balanced constitution is mostly determined by questionnaire, and subjective factors have a great influence. Aiming at the identification of damp-heat constitution and balanced constitution in TCM, this paper utilizes voice signal to automatically realize the constitution identification task, in order to provide assistance for the clinical identification of TCM constitution. Methods Based on deep learning Transformer and transfer learning, a pure attentional mechanism model was designed for the identification of constitution in TCM sound diagnosis. We collected 700 voices from 34 subjects, pre-processed the voice data to obtain the corresponding Mayer spectrum diagram, and used the Transformer model pre-trained based on the public data set to improve the performance of the model for audio classification. Results The accuracy of the experimental results was 83.33%, the AUC was 92.16%, the sensitivity was 80.25%, and the specificity was 87.03%. Compared with the Convolutional Neural Network (CNN), the performance of the deep learning model was better. Conclusion In this paper, the damp-heat constitution and balanced constitution identification model Transformer has achieved better identification effect, indicating that it can improve the efficiency of TCM acoustic diagnosis of constitution identification, and promote the objective and intelligent development of constitution identification.

Keywords: Sound diagnosis of TCM, Constitution identification, Deep learning, Transformer

（责任编辑：刘玥辰）