•智能交叉科学与工程•

DOI:10.12454/j.jsuese.202400537

本刊网刊

基于双路DCGAN数据生成和分类–回归网络的抑郁症检测

卢静雪，李鸿燕*，郑睿超，秦睿臻

(太原理工大学 电子信息与光学工程学院，山西 太原 030024)

摘　要：为了改善基于语音的抑郁症检测研究中存在的特征提取繁杂、数据扩充方式较为单一及回归预测时预测偏差不可控的问题，本文从特征构建、数据增强以及网络架构3个方面进行改进，提出一种采用双路深度卷积生成对抗网络（dual path deep convolution generative adversarial network，DP–DCGAN）进行数据增强和分类–回归网络的抑郁检测模型。特征构建部分用于构建具有时频特性及线性、非线性特性的二维特征图以充分表征抑郁症。数据增强部分提出DP–DCGAN网络进行数据增强，增加特征图的多样性以提高模型的鲁棒性及泛化性；并基于空间和频域特性两方面提出评价指标对生成特征图进行筛选，保留高质量的生成特征图。网络架构部分提出分类–回归网络，通过减小预测置信区间降低预测偏差；针对分类网络中的残差网络，引入多尺度卷积以增强特征间的信息交互，使残差网络充分感知特征图中所蕴含的多层次信息。采用准确率（accuracy）、均方根误差（RMSE）和平均绝对误差（MAE）作为分类–回归网络的评价指标，最终在AVEC2014数据集上，本文模型的四分类准确率达到了94.73%，RMSE和MAE分别为4.55和1.11，与现有的抑郁症检测模型相比具有明显优势。

关键词：语音；抑郁症检测；生成对抗网络；分类–回归；残差网络

中图分类号：TP39　　**文献标志码：**A　　**文章编号：**2096-3246（2026）02-0057-12

2021年全球疾病负担（GBD）数据库显示，全球抑郁症人数约超3.32亿[1]，且仍在持续增长中。到2030年，抑郁症将极有可能成为世界第一大精神障碍类疾病[2]。随着抑郁症逐渐成为一种不容忽视的重大疾病，研究人员开始从图像、语音、文本等方面入手进行抑郁症检测的研究。由于语音能够较为直观地反映人类的情绪状态和心理健康状况，因此许多研究人员对基于语音的抑郁症研究更为关注。

目前，基于语音的抑郁症检测研究工作主要集中在以下两方面：特征提取及构建分类–回归网络模型。在特征提取方面，多数研究采用特征提取工具（如OpenSMILE[3]）进行现有情感特征集的提取[4–7]，所提取的特征多达几十种，维数多达上千维，所提取特征较为繁杂且缺乏对特征有效性的验证。刘振宇[8]通过实验表明，各类语音特征在检测抑郁症方面的有效性存在差异，仅部分语音特征为有效特征；Tian等[9]提取了14类音频特征进行单特征、双特征及多特征组合的对比实验，结果表明，3类特征组合信息之间的互补性接近饱和，证明适当类别的特征组合可以有效提高抑郁症识别的准确性。因此，本文基于抑郁症患者音频信号特点，对现有语音特征进行针对性提取，共提取了6类代表性特征，并通过实验验证了此6类特征在抑郁症检测方面均展现出积极作用，证明了所提特征的有效性。

在网络模型方面，随着深度学习技术的兴起，神经网络模型被广泛应用到基于抑郁症检测的研究中。Lu等[10]将由通道注意力与空间注意力相结合构成的卷积注意力模块（CBAM）以残差结构的形式与残差块相连，使用自制抑郁症数据集进行抑郁症严重程度检测，其中，男性、女性受试者在消极情绪下的四分类识别准确率均高于积极和中性情绪，且分类准确率最高的均为轻微抑郁类别，分别达到了71%和94%。Ishimaru等[11]将音频特征之间的相关性表示为图结构，并使用图卷积神经网络学习语音特征，之后进行抑郁严

收稿日期：2024－07－10　　**修回日期：**2024－11－25　　**网络出版日期：**2025－03－24

基金项目：国家自然科学基金项目（62201377）；山西省回国留学人员科研资助项目（2022–072）

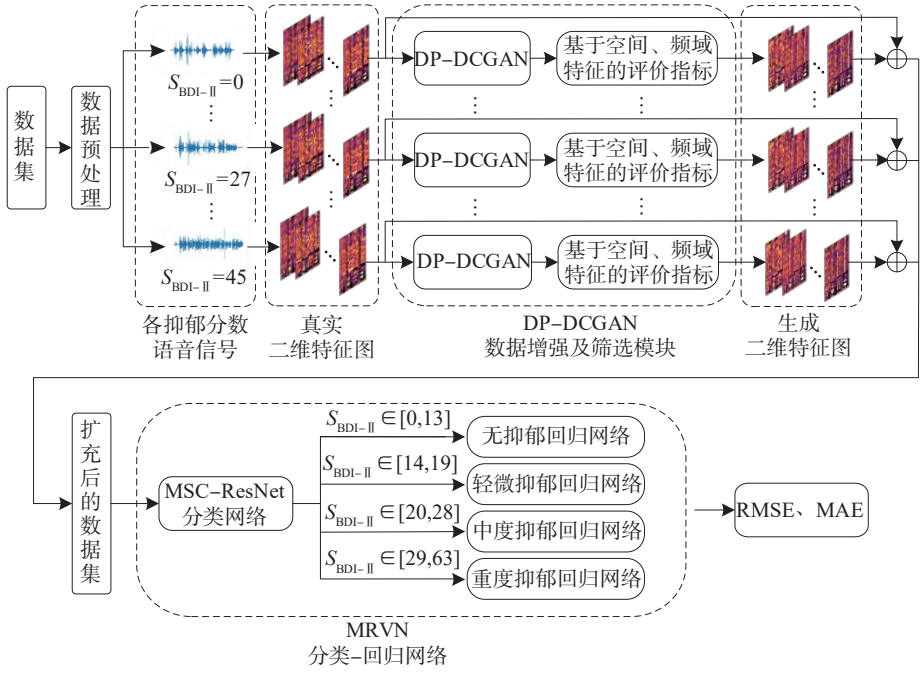
作者简介：卢静雪（2001—），女，硕士生．研究方向：语音情感识别．E-mail：ljx_0404@163.com

* **通信作者：**李鸿燕，教授，E-mail：tylihy@163.com

重程度分类，在DAIC–WOZ数据集上，四分类的精度分别为98.36%、95.79%、94.68%、93.52%。Sardari等[12]采用基于端到端卷积神经网络的自动编码器技术，从原始音频数据中学习高度相关和判别性特征，并在DAIC–WOZ数据集上进行二分类，F分数提高了7%以上。Wang等[6]将共振峰特征输入卷积神经网络、长短时记忆(LSTM)网络和注意力机制融合的网络模型中进行深层特征提取，并由LSTM网络进行抑郁分数预测，在基于E–DAIC改进数据集中，其均方根误差(RMSE)和平均绝对误差(MAE)分别为5.01和3.62。Pan等[13]提出一种多特征深度监督声纹对抗网络(MFDS–VAN)，对提取到的多种声学特征及原始音频波形进行融合编码，并通过声纹对抗网络进行特征增强，以减小特定个体声纹的影响，同时映射到回归网络中进行抑郁分数预测：在AVEC2013数据集上，其RMSE和MAE分别为9.43和7.29；在AVEC2014数据集上，其RMSE和MAE分别为9.44和7.33；在AVEC2017数据集上，其RMSE和MAE分别为5.34和4.27。目前，基于抑郁症检测的研究通常针对单一的分类或回归问题[14–15]，但由于抑郁样本数量有限且模型复杂度较高，往往会出现过拟合的问题[16]。同时，在对抑郁症患者进行回归预测时，通常在整个抑郁量表范围内进行[17]，忽略了预测偏差不可控的问题。基于此，研究人员尝试将分类与回归思想相结合，以达到降低预测偏差的目的。例如：Dong等[18]提出一种回归区间缩放算法，即由多个分类器对深度协调特征进行分类，通过分类结果定位回归区间，在该区间内进行抑郁分数预测；Ma等[16]提出一个由粗分类器和精细回归器组成的两阶段框架，在视频模态下取得了良好效果，但在音频模态下表现不佳。目前，将分类和回归相结合进行抑郁分数预测的方法较少，模型性能有待提高。

另外，由于抑郁症涉及个人隐私问题，数据收集存在一定难度。例如，在AVEC2014数据集中，仅包含150名受试者的视频数据，且随着抑郁分数的提高，抑郁症患者的数量呈现逐渐下降的趋势，抑郁程度越高，样本数量越少。因此，现有数据集样本量有限且数据分布不均衡的问题使数据增强成了目前研究中不可或缺的一环。现有研究大多采用噪声注入、波形位移等常规方式对语音信号进行数据增强[19–20]，这些方法在一定程度上提升了模型的泛化能力，然而，现有数据增强技术在应对数据多样性和复杂性的挑战时，往往存在局限性。近年来，生成对抗网络(GAN)[21]在图像生成方面表现出的优异性能使研究人员尝试将其引入抑郁症数据集的扩充中。Yang等[22]使用条件GAN对AVEC2016[23]数据集进行数据增强，证明了GAN在基于音频的数据增强方面的有效性。GAN在对抗训练过程显著增加了生成样本的多样性[24]，多样化的合成数据将进一步增强模型在面对稀缺样本时的鲁棒性与泛化能力，能够改善传统数据增强方法的局限性。

基于此，本文提出一种基于深度卷积生成对抗网络(DCGAN)和分类–回归网络的抑郁症检测模型。图1为本文模型整体框架。本文主要工作如下：



注：$S_{BDI-Ⅱ}$为贝克抑郁量表Ⅱ(BDI–Ⅱ)分数。

图1 本文模型整体框架

Fig. 1 Overall frame of proposed model

1）提取包含梅尔频率倒谱系数（MFCC）在内的6类语音特征，构建具有时频特性及线性、非线性特性的二维特征图；

2）提出双路深度卷积生成对抗网络（dual path DCGAN，DP–DCGAN）对二维特征图进行数据增强，并从空间特征和频域特征两方面提出评价指标对生成特征图进行筛选；

3）提出一种结合多尺度卷积残差网络（MSC–ResNet）和VGG（visual geometry group）网络的分类–回归抑郁症检测模型MRVN（MSC–ResNet and VGG etwork），使输入数据在更小、更均匀的尺度上进行抑郁分数预测，减少预测误差。

1　二维特征图构建

相较于健康人群，抑郁症患者的语音信号往往呈现出能量较低、节奏变化缓慢以及声强较低等特点[9]。因此，选用短时能量、过零率及声强特征作为抑郁症患者语音时域特征的表征。同时，MFCC特征在抑郁检测中已被证实为有效特征[19–20]，因而在此基础上将Teager能量算子（TEO）[25]融入MFCC特征构成MFCC–TEO特征。将基于频域TEO的非线性运算引入MFCC特征能够进一步凸显能量分布的差异性[26]，使MFCC–TEO特征比MFCC特征具有更优的抑郁症表征能力。MFCC–TEO特征能够更有效地捕捉和表征抑郁症患者语音的频域及非线性特性，特征提取过程如下。

在对语音信号进行预加重、分帧、加窗操作后，先进行N点快速傅里叶变换（FFT），得到各点k的幅值$a(k)$；然后利用式（1）计算各点TEO值TEO($a(k)$)；最后计算各点的能量谱值并输入Mel滤波器组，取对数运算及离散余弦变换即可得到MFCC–TEO系数。

$$\begin{aligned}&\mathrm{TEO}(a(k))=a(k)^2-a(k+1)a(k-1);\\&k=1,2,\cdots,N/2\end{aligned}\tag{1}$$

此外，提取线性预测倒谱系数（LPCC）用于表征抑郁症患者语音的线性特性。有研究表明，抑郁症患者与对照组之间在Jitter特征上具有明显差异[27]，抑郁症患者语音波形的抖动随抑郁程度的增加而增大[28]。因此，选取Jitter特征用于捕捉抑郁症患者声带周期性振动的变化。

对上述特征使用统计函数HSF（high level statistics functions）计算HSF值，包含最小值、最大值、极差、平均值、标准差、平均绝对偏差、偏度系数、峰度系数、四分位数及四分位数的间隔等，在提取特征整体信息的同时降低特征维度。最终，一条语音样本得到770维特征，表1为提取的语音特征名称及维数。

表1　提取的语音特征名称及维数

Tab. 1　Feature names and dimensions of extracted speech features

语音特征类别	语音特征名称	执行统计函数后的维数	总维数
韵律特征	短时能量	14	770
	过零率	14	
	声强	14	
谱相关特征	MFCC–TEO	182	
	MFCC–TEO_Δ	182	
	MFCC–TEO_ΔΔ	182	
	LPCC	168	
语音质量特征	Jitter	14	

注：MFCC–TEO_Δ、MFCC–TEO_ΔΔ为MFCC–TEO特征的一阶差分和二阶差分特征。

由于所提取特征的范围值差异较大，而较大的尺度差异会使网络对各个特征的关注度不同。为了使网络较为平衡地学习到各类特征，对所提取特征采用式（2）统一进行“max–min”归一化，将归一化后的770维特征通过尾部补0的方式转化成大小为28×28的二维特征图。

$$y=y_{\min}+\frac{(y_{\max}-y_{\min})(x-x_{\min})}{x_{\max}-x_{\min}}\tag{2}$$

式中，x、y分别为归一化前、后的数据，$x_{\max}$、$x_{\min}$为归一化前数据的最大值和最小值，$y_{\max}$、$y_{\min}$为归一化后目标区间的最大值、最小值。

2　数据增强及筛选模块

2.1　DP–DCGAN网络

采用DP–DCGAN对AVEC2014数据集[5]进行数据增强。图2为DP–DCGAN架构。

生成器部分由100×1×1维的随机噪声作为输入，经4层转置卷积层后输出生成特征图，且前3层转置卷积层后均加入批归一化（batch normalization）层和Relu激活层，第4层转置卷积层后使用Tanh激活函数。

判别器部分不同于DCGAN中的单路卷积，采用双路卷积。两路具有不同大小感受野的卷积核（3×3和5×5）构成一组3×3卷积层和一组5×5卷积层分别对输入特征图进行特征提取，小卷积核用于提取局部细节信息，大卷积核用于提取更广泛的全局信息。随后，通过一层3×3卷积层对融合后的特征信息进行深层特征提取，使网络学习到更复杂、更高级别的特征表示。通过增加所提取特征信息的丰富性，网络能够更有效地判别输入特征图是真实特征图还是生成特征图，以提升判别器性能。

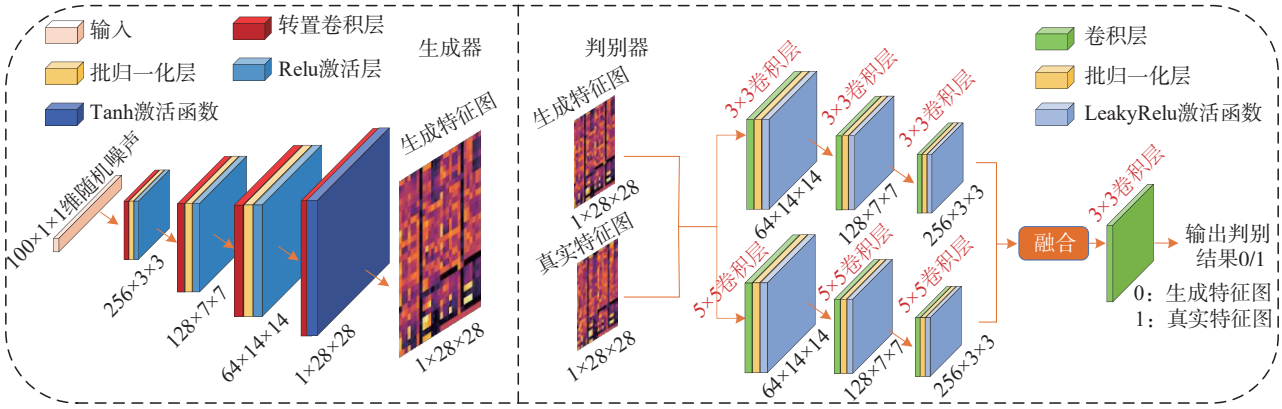


图2 DP–DCGAN架构

Fig. 2 Architecture of DP–DCGAN

2.2 评价指标

由于GAN在数据生成过程中具有极强的随机性，生成图像的质量可能有所波动，而可靠的评价指标有助于剔除在不稳定训练阶段生成的低质量样本，保留的高质量样本则可最大限度地提高模型训练的效果和性能。目前，评价生成图像质量的常用指标有IS（inception score）[29]和FID（Fréchet inception distance）[30]，但这两种指标均存在一定问题。例如：IS指标过于依赖ImageNet数据集，在评估由其他数据集训练生成的图像时结果并不可信；FID指标则无法避免过拟合的问题。因此，以上两种评价指标并不适用，本文提出两种评价指标对生成特征图进行筛选。

2.2.1 基于图像二维熵信息的评价指标

图像熵反映了图像中所包含的平均信息量的大小，包括一维熵和二维熵[31]，但图像的一维熵信息仅能表示图像灰度值的聚集性特征，无法表征空间特性，因此采用二维熵信息作为评价指标，度量特征图之间基于空间特性的相似性。

在计算图像二维熵信息时，采用图像中某一点的像素值i与该点的领域均值j组成特征二元组(i,j)，在图像中同一特征二元组(i,j)发生的概率P_{ij}为：

$$P_{ij}=\frac{f(i,j)}{W\times H} \tag{3}$$

式中，$f(i,j)$为同一特征二元组(i,j)在图像中出现的次数，W和H为图像的宽和高。

图像的二维熵值E为：

$$E=-\sum_{i=0}^{255}\sum_{j=0}^{255}P_{ij}\cdot \mathrm{lb}\,P_{ij} \tag{4}$$

针对某一抑郁分数，假设该抑郁分数中真实特征图总数量为m，生成特征图数量为n。真实特征图用$\boldsymbol{t}_r$表示，下标r为真实特征图序号，$r=1,2,\cdots,m$；生成特征图用$\boldsymbol{t}_s$表示，下标s为生成特征图序号，$s=1,2,\cdots,n$。筛选过程如下：

1）计算该抑郁分数中所有真实特征图的二维熵值$E(\boldsymbol{t}_r)$，构建二维熵信息直方图，由于多数抑郁分数对应的二维熵信息直方图均大致服从高斯分布，因此采用高斯曲线进行拟合。

2）计算生成特征图的二维熵值$E'(\boldsymbol{t}_s)$。

3）图3为二维熵信息筛选标准。b_1和b_2为二维熵信息筛选范围的下限和上限值。图3中阴影部分面积越小，生成特征图与真实特征图的相似性越高。在筛选时，为了确保保留的生成特征图的质量及数据量，采用面积百分比$P(b_1\le E'(\boldsymbol{t}_s)\le b_2)=80\%$作为筛选标准，当$E'(\boldsymbol{t}_s)$在$[b_1,b_2]$中时保留，不在时舍弃。

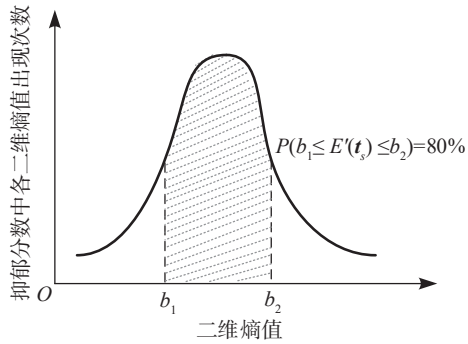


图3 二维熵信息筛选标准

Fig. 3 Two-dimensional entropy information screening criteria

2.2.2 基于FFT和欧几里得距离的评价指标

Yang等[22]采用FFT与余弦距离作为基于频域的图像间相似性的度量指标，考虑到余弦距离仅关注方向信息，忽略了两个向量间的长度差异，因此本文采用FFT与欧几里得距离作为图像间相似性的评价指标，具体步骤如下。

1）针对某一抑郁分数，采用FFT分别计算出该抑郁分数中真实特征图与生成特征图的频谱，将其重塑为一维频谱向量$\boldsymbol{S}(\boldsymbol{t}_r)$和$\boldsymbol{S}'(\boldsymbol{t}_s)$，其维度均为$D$。

2）将第r张真实特征图的频谱向量$\boldsymbol{S}(\boldsymbol{t}_r)$表示为

$(t_{r1},t_{r2},\cdots,t_{rD})$，其中心点$C_r$的坐标为$(C_{r1},C_{r2},\cdots,C_{rD})$，通过式(5)计算出该抑郁分数下所有真实特征图的频谱向量$\boldsymbol{S}(\boldsymbol{t}_r)$与$C$点间的欧几里得距离$d_r$，选择最大值$d_{\max}$与最小值$d_{\min}$构成选择范围$[d_{\min},d_{\max}]$；

$$d_r=\sqrt{\sum_{h=1}^{D}(t_{rh}-C_{rh})^2} \tag{5}$$

式中，下标h表示频谱向量的第h维。

3)用同样的方法计算所有生成特征图的频谱向量$\boldsymbol{S}'(\boldsymbol{t}_s)$与其中点间的欧几里得距离$d_s'$，当$d_s'\in[d_{\min},d_{\max}]$时保留，否则舍弃。

通过计算频谱向量与中心点间的欧几里得距离，不仅在一定程度上保留了频谱向量的方向信息，同时还限制了距离范围。

本文提出的两个评价指标分别从空间特征和频域特征两个角度对生成特征图进行了筛选，当生成特征图同时满足这两个评价指标时保留。

3　分类–回归网络

MRVN将分类与回归相结合，用于抑郁分数预测，使样本在更小、更均匀的尺度上进行预测，改善以往研究中使用单一回归网络进行抑郁分数预测时偏差大的问题。将由DP–DCGAN生成、经两种评价指标筛选后的数据样本与原始样本共同组成新的数据集作为MRVN的输入。扩充后的数据集首先经由MSC–ResNet进行四分类，并按类别对样本进行归类；然后，输入对应类别的抑郁回归网络进行抑郁分数预测，得到预测值。图4为MRVN框架，以Conv1–1:128@5×5为例，其表示第1层的第1组卷积，卷积核个数为128，卷积核大小为5×5。

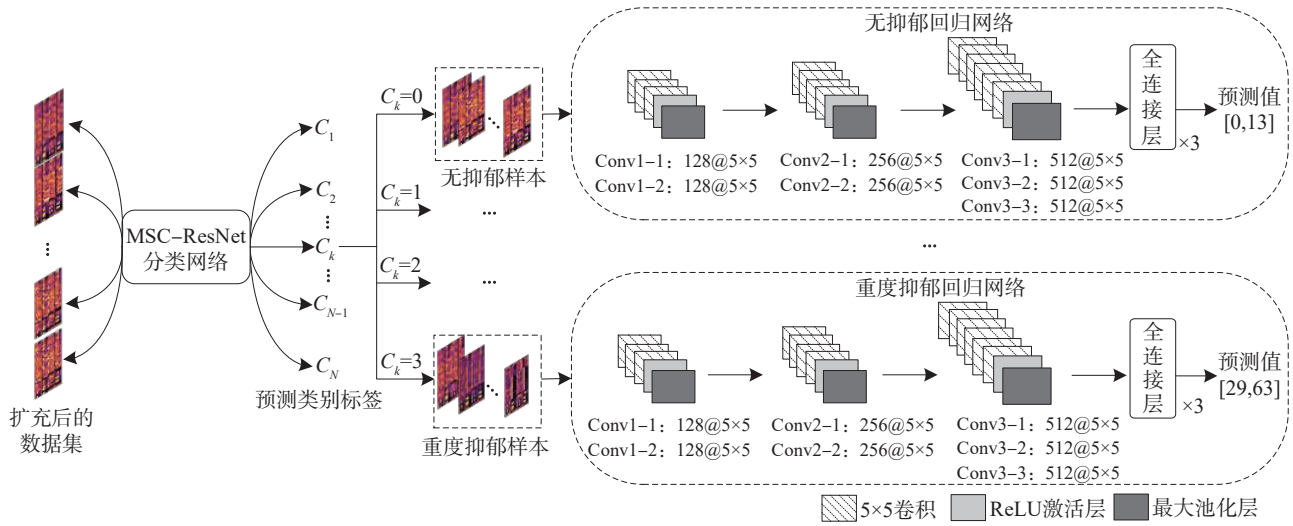


图4　MRVN框架

Fig. 4　Frame of MRVN

3.1　MSC–ResNet分类网络

由于传统的残差网络(ResNet)[32]结构不易改变，且每个卷积层通常具有固定大小的卷积核，单一尺度的感受野使网络在进行特征提取时具有较大的局限性，因此对于输入数据中不同尺度的特征，无法进行有效处理。

本文在提取特征的过程中，对每维特征均使用14个统计函数，因此，每维特征均包含14个HSF值。呈现在特征图上：每一行包含两个维度特征的各14个HSF值；每一列则对应着不同特征的同一HSF值。

针对ResNet中特征提取的局限性问题，结合本文特征图独有的特点，提出了MSC–ResNet。图5为MSC–ResNet分类网络框架。

首先，对输入特征图进行多尺度卷积(MSC)提取多尺度特征信息，3路卷积操作如下：

第1路卷积操作：卷积核大小为1×7，横、纵向移动步长分别为7、1，对同一特征的7个不同的HSF值进行信息整合。

第2路卷积操作：卷积核大小为7×1，横、纵向移动步长均为1，对不同特征的同一HSF值进行信息整合。

第3路卷积操作：采用空洞率(dilation rate)为14的2×2空洞卷积提取更加广泛的上下文信息。

然后，将经3路卷积操作得到的子特征图进行融合，并将得到的新特征图送入残差模块进行深层次特征的表示学习。共使用3个残差模块，3个残差模块分别包含2、3、2个残差块，每个残差块的中间卷积层的卷积核大小为3×3，其余为1×1。

最后，经平均池化层和全连接层将网络学习到的特征信息映射到各自的类别中，以此得出分类结果。

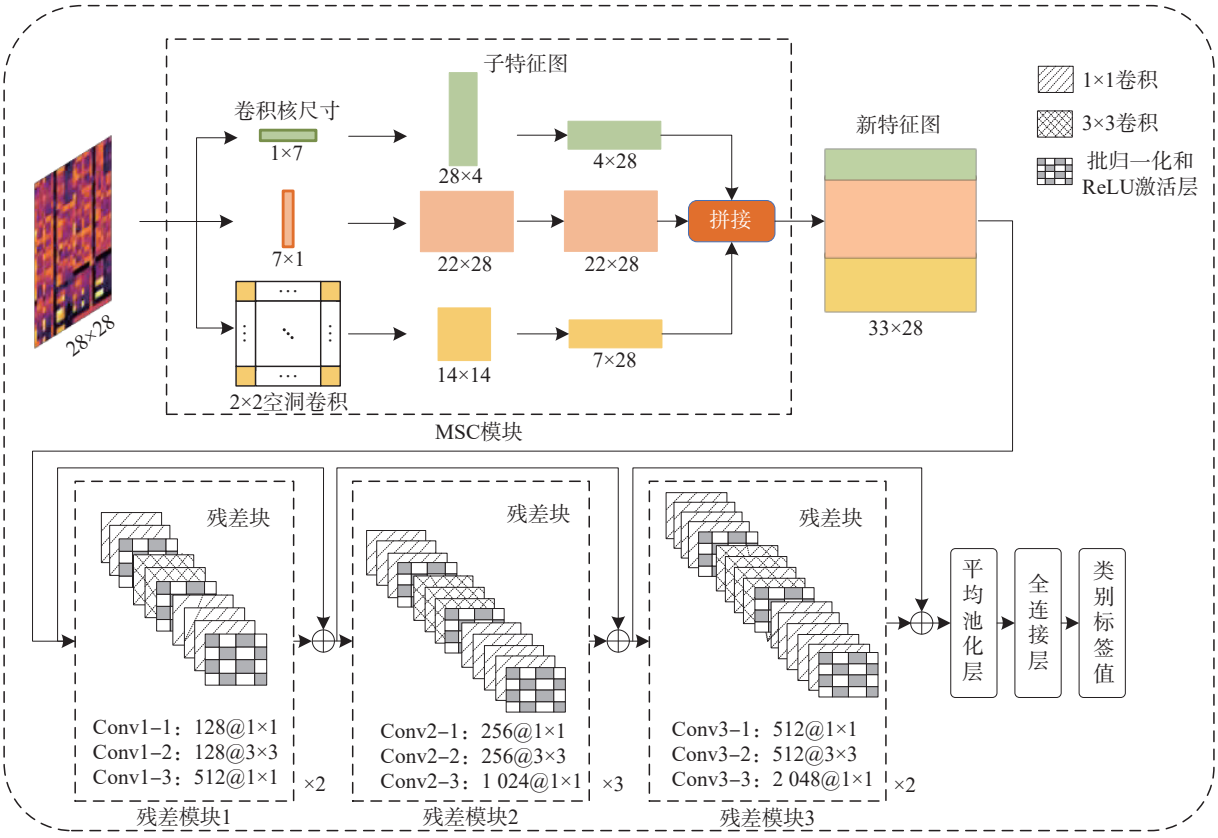


图5 MSC–ResNet分类网络框架

Fig. 5 MSC–ResNet classification network framework

3.2 回归网络

基于VGG16[33]网络进行抑郁分数的回归预测，网络架构如图4所示。首先由对应类别的训练样本分别训练出4个对应的回归网络，然后将MSC–ResNet分类网络完成分类预测的测试集样本输入对应类别的回归网络进行抑郁分数预测。

在任一回归网络中，以2层卷积层为一组，对于输入数据，首先经过两组卷积（两组卷积中卷积核数量分别为128和256）；之后，经过以3层卷积层为一组的卷积操作（每层卷积核数量为512），每组卷积均接1个ReLU激活层和最大池化层；最后，通过3层全连接层输出预测值。此外，每个ReLU激活层后均会加入一个dropout率为0.2的Dropout层以防止模型过拟合。

4 实验部分

4.1 实验环境

所有测试实验的硬件环境为Silver处理器，操作系统为64位Ubuntu系统，运行内存为32 GB，GPU为NVIDIA RTX A4000，软件环境为Python 3.7.4，Pytorch版本为1.13.0+cu117。

4.2 数据集及预处理

使用AVEC2014数据集[5]进行实验，该数据集是抑郁症研究方面的一个公开数据集，采用贝克抑郁量表Ⅱ（BDI–Ⅱ）评估抑郁严重程度，依据BDI–Ⅱ值可划分为4类：无抑郁（0～13）、轻微抑郁（14～19）、中度抑郁（20～28）和重度抑郁（29～63）。

在预处理阶段，去除每个语音文件中的长时间静音段后，将其切割为时长1 s的等长度语音段，相邻语音段之间有0.2 s的间隔，将同一抑郁分数的语音段放在同一文件夹中保存。

采用DP–DCGAN进行数据增强，在网络训练过程中，批大小（batch size）设置为256，当某一抑郁分数的语音段数量低于256时，采用波形位移、波形拉伸、音高修正的数据增强方式对该抑郁分数的语音段进行数据预扩充，以达到网络的训练要求。

4.3 实验参数设置

4.3.1 DP–DCGAN

在实验中，batchsize为256，损失函数为交叉熵损失函数；选择适应性矩估计（Adam）算法作为网络优化器，生成器和判别器的学习率分别设置为0.002和0.000 02。

针对每个抑郁分数的特征图，分别采用DP–DCGAN进行训练。在每次训练中，网络训练2 000个epoch，从第1 800个epoch开始，每间隔20个epoch将生成器生成的特征图保存下来，采用本文提出的两种评价指标进行筛选。筛选后的特征图与真实特征图构

成新数据集，共包含77 161张特征图，将每个抑郁分数中的特征图按8∶2的比例划分为训练集和测试集，作为后续MRVN的输入。

4.3.2　MRVN分类–回归网络

表2为MSC–ResNet分类网络与VGG16回归网络的实验参数设置。

表2　MRVN实验参数设置

Tab. 2　MRVN experiment parameter settings

网络	批大小	损失函数	优化器	学习率
MSC–ResNet	128	交叉熵损失函数	Adam算法	0.001
VGG16				0.000 1

在回归网络部分，采用常见的均方误差（MSE）和MAE作为回归损失函数，网络对测试集样本输出的预测分数分布如图6所示，其中红框为预测值范围。由图6可见，每个回归网络输出的预测值均分布在所在类别标签范围的中间值附近，与使预测分数能够较为均匀地分布在标签范围内的期望不符。基于此，采用交叉熵损失函数作为回归网络的损失函数。

交叉熵损失函数Loss(·)的计算公式如下：

$$\mathrm{Loss}(M,Z)=-\frac{1}{Z}\sum_{e=1}^{Z}\sum_{g=1}^{M}y_{eg}\ln P_{eg} \tag{6}$$

式中：Z为参与分类的样本总数；M为类别总数；y_{eg}用于样本类别判定，当样本e属于类别g时，$y_{eg}=1$，否则为0；P_{eg}为样本e被预测为类别g时的概率值；ln函数也可以为任意底数的log函数。

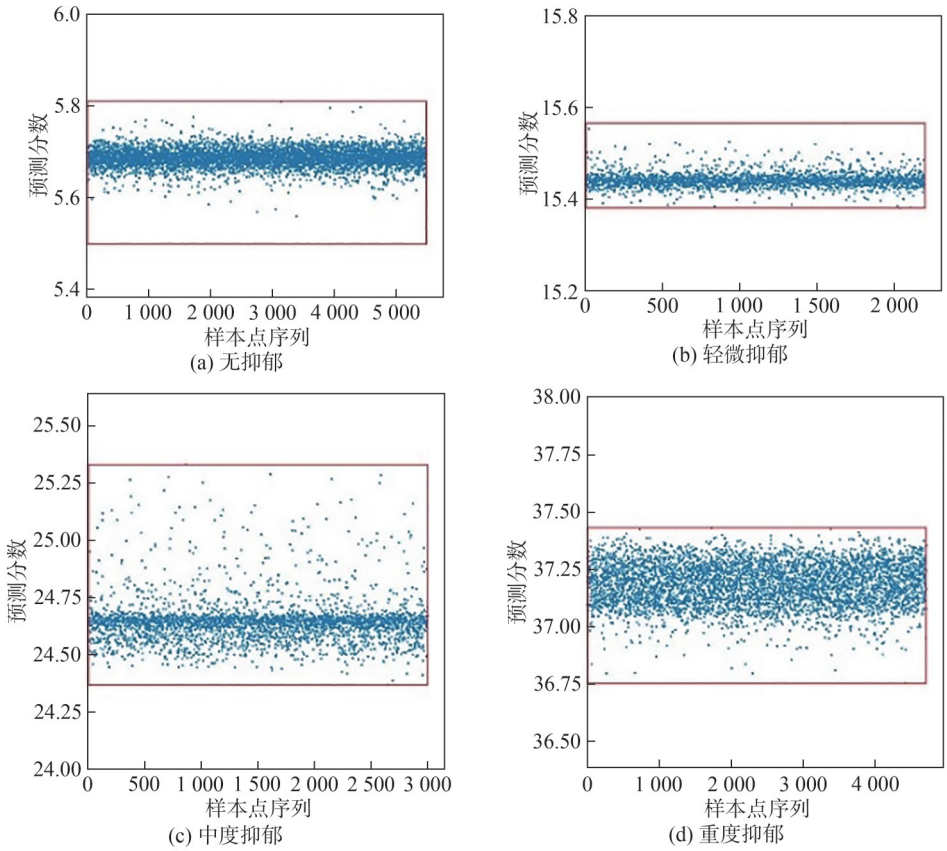


图6　预测分数分布

Fig. 6　Distribution of predicted score

4.4　评价指标

本文采用准确率（accuracy，记为A）、RMSE（记为E_{RMSE}）和MAE（记为E_{MAE}）作为MSC–ResNet的评价指标，计算公式如下：

$$A=\frac{T_{\mathrm{TP}}+T_{\mathrm{TN}}}{T_{\mathrm{TP}}+T_{\mathrm{TN}}+F_{\mathrm{FP}}+F_{\mathrm{FN}}}\times 100\% \tag{7}$$

$$E_{\mathrm{RMSE}}=\sqrt{\frac{1}{Y}\sum_{l=1}^{Y}(y_l-\hat{y}_l)^2} \tag{8}$$

$$E_{\mathrm{MAE}}=\frac{1}{Y}\sum_{l=1}^{Y}\left|y_l-\hat{y}_l\right| \tag{9}$$

式（7）～（9）中，T_{TP}、T_{TN}为被正确分类为正、负样本的数量，F_{FP}、F_{FN}为被错误分类为正、负样本的数量，Y为样本数，y_l为真实标签值，$\hat{y}_l$为预测值，下标l为样本序号。

4.5　实验结果分析

4.5.1　不同特征对比实验

为了验证所选取特征的有效性，在短时能量、过

零率、声强的基础上，依次加入MFCC、MFCC–TEO、LPCC和Jitter特征，分别通过DP–DCGAN网络进行数据增强、MRVN网络进行抑郁分数预测，计算不同输入特征下的准确率、RMSE及MAE。表3为不同特征输入下的实验结果。

表3 不同特征输入下的实验结果

Tab. 3 Experimental results under different input characteristics

输入特征	准确率/%	RMSE	MAE
MFCC	89.76	6.17	2.08
MFCC–TEO	92.07	5.49	1.58
MFCC–TEO+LPCC	93.41	5.09	1.39
MFCC–TEO+LPCC+Jitter	94.73	4.55	1.11

由表3可见：使用MFCC–TEO作为输入特征比使用MFCC时准确率提升了2.31个百分点，RMSE和MAE分别下降了0.68和0.50，说明通过将MFCC特征与TEO相结合，增大了MFCC中能量分布的差异性，使结合后的MFCC–TEO系数较MFCC具有更优的抑郁症表征能力；LPCC、Jitter特征加入后，通过弥补线性信息的缺失及对声带周期性振动变化信息的捕捉，丰富了所提取的特征信息，模型检测结果达到最优。

4.5.2 DP–DCGAN中单、双路卷积及不同卷积核大小对检测性能的影响

为了研究使用DP–DCGAN进行数据增强的效果，将原始数据集与增强后的数据样本分别输入MRVN，计算MRVN中MSC–ResNet的准确率和回归网络的RMSE、MAE。此外，为了验证在DP–DCGAN中当判别器使用双路卷积时对模型检测性能的影响，在判别器中分别采用单路和双路卷积进行数据增强实验，为了选出使模型性能达到最优的卷积核大小，分别使用大小为3×3、5×5、7×7的3种卷积核及其组合进行对比实验，计算MRVN的准确率、RMSE、MAE。表4为DP–DCGAN中单、双路卷积实验结果。

表4 DP–DCGAN中单、双路卷积实验结果

Tab. 4 Experimental results of single and double convolutions in DP–DCGAN

数据集	卷积通道数	卷积核大小	准确率/%	RMSE	MAE
原始数据集			80.51	8.47	3.94
数据增强后的数据集	单路卷积	3×3	94.01	4.91	1.27
		5×5	94.29	4.83	1.23
		7×7	93.65	5.09	1.34
	双路卷积	3×3和5×5	94.73	4.55	1.11
		3×3和7×7	94.25	4.79	1.23
		5×5和7×7	94.06	4.85	1.25

由表4可见，使用数据增强后的数据集进行实验的效果较原始数据集有大幅提升，准确率提升了14.22个百分点，RMSE和MAE分别降低了3.92和2.83，证明使用DP–DCGAN进行数据增强能够有效达到扩充数据集的目的。此外，当判别器采用双路卷积，且两路卷积核大小分别为3×3和5×5时，MSC–ResNet的准确率最高，达到94.73%，同时回归网络的RMSE和MAE最低，分别为4.55和1.11；MSC–ResNet的准确率比单路卷积中最优者提升了0.44个百分点，回归网络的RMSE和MAE比单路卷积中最优者分别降低了0.28和0.12，说明当判别器使用双路卷积时，生成特征图更接近真实特征图，对后续分类、回归网络的识别性能有一定提升作用。

4.5.3 DP–DCGAN中使用不同筛选指标进行筛选的实验结果

为了验证本文提出的两种评价指标（二维熵、FFT+欧氏距离）的有效性，与Yang等[22]采用的两种评价指标（一维熵、FFT+余弦距离）进行对比实验。表5为不同筛选指标下的实验结果。

表5 不同筛选指标下的实验结果

Tab. 5 Experimental results under different screening indexes

筛选指标	准确率/%	RMSE	MAE
一维熵、FFT+余弦距离[22]	92.96	5.28	1.45
二维熵、FFT+欧氏距离（本文评价指标）	94.73	4.55	1.11

由表5可见，生成特征图经本文评价指标筛选后在后续进行抑郁分数预测的实验中具有明显优势，相较于经一维熵、FFT+余弦距离[22]的评价指标进行筛选，准确率提升了1.77个百分点，RMSE和MAE分别降低了0.73和0.34，证明了本文评价指标的有效性。

4.5.4 MSC–ResNet中MSC模块对分类性能的影响

在DP–DCGAN中，当判别器使用3×3和5×5双路卷积时，为了研究分类网络中MSC模块对分类结果的影响，对MSC–ResNet与原始的ResNet50分类网络进行对比实验。此外，在残差块的中间卷积层分别使用大小为3×3、5×5、7×7的3种卷积核进行实验，观察不同尺度卷积核对模型性能的影响。表6为优化网络与原始网络分类结果。

从表6可见：在ResNet50分类网络中，当残差块的中间卷积层使用大小为7×7的卷积核时准确率最高，为93.86%，与大小为3×3和5×5的卷积核相比准确率分别提升了0.58和1.07个百分点，说明较大尺度的感受野中包含着不可忽略的全局信息；当使用相同大小的卷积核时，MSC–ResNet的准确率均比ResNet50分

表6　优化网络与原始网络分类结果

Tab. 6　Classification results of optimized network and original network

分类网络	残差块中间卷积层卷积核大小	准确率/%
ResNet50	3×3	93.28
	5×5	92.79
	7×7	93.86
MSC–ResNet	3×3	94.73
	5×5	94.23
	7×7	93.96

类网络高；在MSC–ResNet中，当残差块的中间卷积层使用大小为3×3的卷积核时准确率最高，为94.73%，与7×7的卷积核相比准确率提升了0.77个百分点，说明通过多尺度卷积生成的新特征图中包含较为丰富的全局信息以及更加广泛的上下文信息，并通过使用卷积核大小为3×3的卷积层提取新特征图中的细节信息，使得网络能够充分感知特征图中所包含的多尺度信息，达到了提升模型性能的目的。

4.5.5　本文模型与以往工作的对比

为了进一步验证本文模型在抑郁检测方面的有效性，将本文模型与以往工作进行对比，结果如表7所示。本文模型提取各数据样本的6类情感特征并构建其二维特征图，各抑郁分数的样本数据分别通过DP–DCGAN进行数据增强，同时采用本文提出的两种评价指标进行筛选，将筛选后的数据样本与原始样本组成新数据集，共同作为后续MRVN的输入样本进行实验。

表7　本文模型性能与以往工作的对比

Tab. 7　Comparison of proposed model performance with previous works

文献	RMSE	MAE
Valstar等[5]	12.56	10.03
He等[4]	10.00	8.19
Pan等[13]	9.44	7.33
Niu等[34]	9.13	7.65
Uddin等[35]	8.46	6.95
本文	4.55	1.11

由表7可见，本文模型性能达到最优。文献[4–5]均使用OpenSMILE提取2 268维音频特征，不同的是文献[4]使用深度卷积神经网络（DCNN）进行深度特征提取，性能较优。文献[13,34–35]同时考虑了时空特征进行抑郁分数预测，采用RMSE或Huber损失函数。上述5篇文献相同的是均进行单一回归预测，虽然文献[13]尝试使用交叉熵损失函数代替Huber损失函数进行分类与回归任务，但结果并不理想。而本文模型在特征提取过程中不仅考虑了音频信号中的时频特性及线性、非线性特性，同时将特征维数降至770，并通过将分类与回归相结合，使用交叉熵损失函数，对输入数据在更小、更均匀的尺度上进行回归预测，降低了预测偏差，使性能达到最优。

5　结　论

本文提出了一种结合深度卷积生成对抗网络的分类–回归抑郁症检测模型。将TEO引入MFCC得到MFCC–TEO特征，并提取包含其在内的6类特征构建具有时频特性及线性、非线性特性的二维特征图；通过构建DP–DCGAN对原始数据集中各个抑郁分数的特征图进行数据增强，增加了特征图的多样性，并从空间特性和频域特性两方面提出评价指标筛选出高质量的特征图，优质且多样化的特征图显著提升了模型性能；通过提出的MRVN进行抑郁分数预测，结合本文特征图独有的特点在ResNet中加入多尺度卷积模块，改善了ResNet中单一尺度感受野特征提取局限性的问题；通过将分类与回归相结合，使输入数据能够在更均匀的尺度上进行回归预测，改善了回归预测时预测偏差大的问题。

在后续工作中，可以考虑加入图像、文本或其他模态的数据弥补语音模态中缺失的信息，多模态融合或将进一步提升对抑郁严重程度诊断的准确性。

参考文献：

[1] Institute for Health Metrics and Evaluation. Global burden of disease(GBD)2021[DB/OL]. Seattle: IHME, 2024 [2026–03–04]. https://vizhub.healthdata.org/gbd-results/.

[2] Mathers C D,Loncar D.Projections of global mortality and burden of disease from 2002 to 2030[J]. PLoS Medicine, 2006,3(11):e442.

[3] Eyben F, Schuller B. openSMILE:): The Munich open-source large-scale multimedia feature extractor[J]. ACM SIGMultimedia Records,2015,6(4):4–13.

[4] He Lang,Cao Cui.Automated depression analysis using convolutional neural networks from speech[J].Journal of Biomedical Informatics,2018,83:103–111.

[5] Valstar M,Schuller B,Smith K, et al.AVEC 2014:3D dimensional affect and depression recognition challenge[C]// Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge.Orlando:ACM,2014:3–10.

[6] Wang Congcong,Liu Decheng,Tao Kemeng,et al.A multimodal feature layer fusion model for assessment of depression based on attention mechanisms[C]//Proceedings of the 2022 15th International Congress on Image and Sig-

nal Processing, BioMedical Engineering and Informatics (CISP–BMEI).Beijing:IEEE,2022:1–6.

[7] Yang Le,Jiang Dongmei,Han Wenjing,et al.DCNN and DNN based multi-modal depression recognition[C]//Proceedings of the 2017 Seventh International Conference on Affective Computing and Intelligent Interaction(ACII). San Antonio:IEEE,2017:484–489.

[8] Liu Zhenyu.Research on method and key technology for depression recognition based on speech[D]. Lanzhou: Lanzhou University,2017.[刘振宇 . 基于语音的抑郁识别方法及关键技术研究[D]. 兰州:兰州大学,2017.]

[9] Tian Han,Zhu Zhang,Jing Xu.Deep learning for depression recognition from speech[J].Mobile Networks and Applications,2024,29(4):1212–1227.

[10] Lu Xiaoyong, Shi Daimin, Liu Yang, et al.Speech depression recognition based on attentional residual network[J]. Frontiers in Bioscience,2021,26(12):1746–1759.

[11] Ishimaru M, Okada Y, Uchiyama R, et al.Classification of depression and its severity based on multiple audio features using a graphical convolutional neural network[J].International Journal of Environmental Research and Public Health,2023,20(2):1588.

[12] Sardari S, Nakisa B, Rastgoo M N, et al.Audio based depression detection using convolutional autoencoder[J]. Expert Systems with Applications,2022,189:116076.

[13] Pan Yuchen,Shang Yuanyuan,Wang Wei,et al.Multi-feature deep supervised voiceprint adversarial network for depression recognition from speech[J]. Biomedical Signal Processing and Control,2024,89:105704.

[14] Cummins N,Scherer S,Krajewski J,et al.A review of depression and suicide risk assessment using speech analysis[J]. Speech Communication,2015,71:10–49.

[15] Pampouchidou A,Simos P G,Marias K,et al.Automatic assessment of depression based on visual cues:A systematic review[J].IEEE Transactions on Affective Computing, 2019,10(4):445–470.

[16] Ma Xingchen,Huang Di,Wang Yunhong,et al.Cost-sensitive two-stage depression prediction using dynamic visual clues[C]//Computer Vision – ACCV 2016. Cham: Springer, 2017:338–351.

[17] Senoussaoui M,Sarria–Paja M,Santos J F,et al.Model fusion for multimodal depression classification and level detection[C]//Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge.Orlando:ACM,2014: 57–63.

[18] Dong Yizhuo,Yang Xinyu.A hierarchical depression detection model based on vocal and emotional cues[J]. Neurocomputing,2021,441:279–290.

[19] Muzammel M,Salam H,Othmani A.End-to-end multimodal clinical depression recognition using deep neural networks:A comparative analysis[J].Computer Methods and Programs in Biomedicine,2021,211:106433.

[20] Rejaibi E,Komaty A,Meriaudeau F,et al.MFCC-based Recurrent Neural Network for automatic clinical depression recognition and assessment from speech[J]. Biomedical Signal Processing and Control,2022,71:103107.

[21] Goodfellow I,Pouget–Abadie J,Mirza M,et al.Generative adversarial nets[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2. Cambridge:MIT Press,2014:2672–2680.

[22] Yang Le,Jiang Dongmei,Sahli H.Feature augmenting networks for improving depression severity estimation from speech signals[J].IEEE Access,2020,8:24033–24045.

[23] Valstar M,Gratch J,Schuller B,et al.AVEC 2016:Depression, mood, and emotion recognition workshop and challenge[C]//Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. Amsterdam: ACM, 2016:3–10.

[24] Wang Kunfeng,Gou Chao, Duan Yanjie,et al.Generative adversarial networks:Introduction and outlook[J].IEEE/CAA Journal of Automatica Sinica,2017,4(4):588–598.

[25] Kaiser J F.On a simple algorithm to calculate the ‘energy’ of a signal[C]//Proceedings of the International Conference on Acoustics,Speech,and Signal Processing.Albuquerque:IEEE,2002:381–384.

[26] Gao Hui,Chen Shanguang,Su Guangchuan.Emotion classification of mandarin speech based on TEO nonlinear features[C]//Proceedings of the Eighth ACIS International Conference on Software Engineering,Artificial Intelligence,Networking,and Parallel/Distributed Computing (SNPD 2007).Qingdao:IEEE,2007:394–398.

[27] Ozdas A,Shiavi R G,Silverman S E,et al.Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk[J]. IEEE Transactions on Biomedical Engineering,2004,51(9):1530–1540.

[28] Han Zhuojin,Shang Yuanyuan,Shao Zhuhong,et al.Spatial-temporal feature network for speech-based depression recognition[J].IEEE Transactions on Cognitive and Developmental Systems,2024,16(1):308–318.

[29] Salimans T,Goodfellow I,Zaremba W,et al.Improved techniques for training GANs[EB/OL].(2016–06–10)[2024–03–18].https://arxiv.org/abs/1606.03498v1.

[30] Heusel M,Ramsauer H,Unterthiner T,et al.GANs trained by a two time-scale update rule converge to a local Nash equilibrium[EB/OL].(2017–06–26)[2024–03–18].https://arxiv.org/abs/1706.08500v6.

[31] Brink A D.Thresholding of digital images using two-dimensional entropies[J].Pattern Recognition,1992,25(8):803–808.

[32] He Kaiming,Zhang Xiangyu,Ren Shaoqing,et al.Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR).Las Vegas:IEEE,2016:770–778.

[33] Simonyan K,Zisserman A.Very deep convolutional networks for large-scale image recognition[EB/OL].(2014–09–04)[2024–03–18].https://arxiv.org/abs/1409.1556.

[34] Niu Mingyue,Tao Jianhua,Liu Bin,et al.Multimodal spatiotemporal representation for automatic depression level detection[J]. IEEE Transactions on Affective Computing,2023,14(1):294–307.

[35] Uddin M A,Joolee J B,Sohn K A.Deep multi-modal network based automated depression severity estimation[J]. IEEE Transactions on Affective Computing,2023,14(3):2153–2167.

Depression Detection Based on Dual–Path DCGAN Data Generation and Classification–Regression Network

LU Jingxue, *LI Hongyan**, *ZHENG Ruichao*, *QIN Ruizhen*

(College of Electronic Information and Optical Engineering, Taiyuan University of Technology, Taiyuan 030024, China)

Abstract:

Objective Accurate evaluation of depression scores in patients with depression provides effective support for clinical auxiliary diagnosis and enables the development of personalized diagnosis and treatment plans, improving the overall accuracy of clinical diagnosis and intervention and contributing significantly to patient health outcomes. Existing research on voice-based depression detection exhibits several limitations, including complex feature extraction processes, single-mode data augmentation, and uncontrollable prediction bias in regression-based estimation. This study proposes a dual–path DCGAN for data generation and introduces a classification-regression network model for depression score prediction, enabling effective auxiliary diagnosis of depression severity.

Methods Firstly, based on the audio characteristics of depressed patients, six types of emotional features were selected from existing speech features, and corresponding two-dimensional feature maps were constructed for each audio signal sample. For MFCC features, the Teager energy operator was fused with MFCC to form MFCC–TEO features, which further highlighted differences in energy distribution. In addition, the dual–path deep convolutional generative adversarial network proposed in this study was utilized to enhance the two-dimensional feature maps of each depression level to expand the dataset, increase feature map diversity, and improve model robustness and generalization. Simultaneously, an evaluation index based on spatial and frequency domain characteristics was proposed to screen generated feature maps and retain high-quality samples. Finally, a classification regression network was introduced into the prediction framework to reduce prediction bias by narrowing the prediction confidence interval. For residual networks within the classification framework, multi-scale convolution was introduced to enhance information interaction among features, which enabled the residual network to fully perceive multi-level information contained in the feature maps.

Results and Discussions Feature validity tests were conducted for the six selected emotional features, in which MFCC, MFCC–TEO, LPCC, and Jitter features were sequentially added based on short-term energy, zero-crossing rate, and sound intensity, and accuracy (Acc), root mean square error (RMSE), and mean absolute error (MAE) under different input configurations were calculated. Experimental results showed that Acc, RMSE, and MAE were 89.76%, 6.17, and 2.08, respectively, when MFCC was added. When MFCC–TEO was added, Acc, RMSE, and MAE reached 92.07%, 5.49, and 1.58, respectively. When MFCC–TEO and LPCC were added, Acc, RMSE, and MAE further improved to 93.41%, 5.09, and 1.39, respectively. When MFCC–TEO, LPCC, and Jitter were added, Acc, RMSE, and MAE reached 94.73%, 4.55, and 1.11, respectively. These results demonstrated that when MFCC–TEO was used as an input feature, Acc increased by 2.31 percentage points, while RMSE and MAE decreased by 0.68 and 0.50, respectively, compared to using MFCC alone, which indicated that combining MFCC with TEO enhanced the representation of energy distribution differences. The MFCC–TEO coefficient exhibited stronger depression characterization capability than the MFCC coefficient. Subsequent incorporation of LPCC and Jitter features further improved prediction accuracy to a certain extent. In the data enhancement experiments, when the original dataset was utilized to predict depression scores, Acc, RMSE, and MAE were 80.51%, 8.47, and 3.94, respectively. After data enhancement using the dual deep convolutional generative adversarial network, Acc, RMSE, and MAE improved to 94.73%, 4.55, and 1.11, respectively. Compared to the original dataset, prediction accuracy significantly improved, with Acc increasing by 14.22 percentage points, and RMSE and MAE decreasing by 3.92 and 2.83, respectively, which demonstrated that DP–DCGAN-based data enhancement effectively expanded the dataset. In the prediction network, the classification accuracy of the original ResNet was 93.28%, while the MSC–ResNet achieved a classification accuracy of 94.73%, representing an improvement of 1.45 percentage points. These results confirmed that the multi-scale convolution strategy extracted richer global and contextual information, after which the residual network captured detailed informa-

tion, enabling the network to fully perceive multi-scale characteristics within the input feature maps and ultimately improve overall model performance.

Conclusions This study proposes a depression diagnosis model based on a deep generation network and a classification regression framework. The MFCC – TEO feature is obtained by introducing TEO into MFCC, and six features, including TEO, are extracted to construct a two-dimensional feature map incorporating time–frequency, linear, and nonlinear properties. Feature maps corresponding to each depression score in the original dataset are enhanced to increase feature diversity, and evaluation indicators are proposed to screen high-quality feature maps from both spatial and frequency domain perspectives by constructing a DP–DCGAN network. High-quality and diversified feature maps significantly improve the overall performance of the model. Finally, the proposed MRVN classification regression network is applied to predict depression scores. A multi-scale convolution module is added to the ResNet classification network to address the limitation of single-scale receptive fields in feature extraction by integrating the unique characteristics of the feature maps proposed in this study. In addition, the input data can be predicted on a more uniform scale by combining classification and regression strategies, reducing large prediction deviations commonly observed in regression tasks.

Key words: audio; depression detection; generative adversarial network; category–regression; residual network

（编辑　李轶楠）

引用格式: Lu Jingxue,Li Hongyan,Zheng Ruichao,et al.Depression detection based on dual–path DCGAN data generation and classification–regression network[J].Advanced Engineering Sciences,2026,58(2):57–68.[卢静雪,李鸿燕,郑睿超,等. 基于双路DCGAN数据生成和分类–回归网络的抑郁症检测[J]. 工程科学与技术,2026,58(2):57–68.]