arXiv:2603.08154v1 [cs.SD] 9 Mar 2026

频谱图中的声景：南亚声音多标签分类的开创性研究 *

Sudip Chakrabarty[1], Pappu Bishwas[1], Rajdeep Chatterjee[1], Tathagata Bandyopadhyay[2], DigontoBiswas[1], Bibek Howlader[3]

[1]计算机工程学院，KIIT大学，布巴内斯瓦尔，印度[2]计算、信息与技术学院，慕尼黑工业大学，德国[3]计算机工程学院，美国国际大学，达卡，孟加拉国

电子邮箱：{sudipchakrabarty6、cse.rajdeep、gata.tatha14}@gmail.com

2026年3月10日

摘要

环境声音分类对于城市监测和文化声景分析日益重要，尤其是在南亚声学环境丰富的地区。这些地区面临着独特的挑战：多种自然、人类和文化声音往往相互重叠，使传统方法难以应对，而传统方法通常依赖梅尔频率倒谱系数（MFCC）。本研究提出了一种新颖的基于频谱图的方法，能够更有效地捕捉这些复杂的听觉模式。研究采用卷积神经网络（CNN）架构，在 SAS-KIIT 数据集上解决具有挑战性的多标签、多类别分类问题。为验证该方法的稳健性和可比性，研究还使用著名的 UrbanSound8K 数据集进行了验证。结果证实，所提出的基于频谱图的方法显著优于现有的基于 MFCC 的技术，在两个数据集上均取得了更高的分类准确率。这一提升为现实应用中更稳健、准确的音频分类系统奠定了基础。

关键词：多标签分类、卷积神经网络（CNN）、SAS-KIIT、UrbanSound8K、频谱图分析、环境声音分类（ESC）。

1 引言

环境声音的分析对于城市监控、公共安全和保护文化声景至关重要。在南亚，由于其丰富多变的听觉环境，这项任务更具挑战性；自然、人为和文化声音往往交织在一起。准确识别并分类这些声音，对于开发智能城市系统、提升环境意识以及保护该地区的声学遗产至关重要。此外，环境音频承载着重要的背景信息，可用于监测实时变化、检测异常，并在资源有限的环境中支持决策。然而，南亚声景错综复杂，多个声源彼此重叠，这给传统分类方法带来了重大挑战。

用于分类混合环境声音的传统方法通常依赖盲源分离（BSS）技术，例如独立成分分析（ICA）和主成分分析（PCA） [1, 2], ，这些方法需要预先知道声源数量。另一种条件式音频分离方法 [3] 或许也可采用，但这类方法需要每种声音类别的干净预录样本。在真实且动态的环境中，当重叠信号

*Amygdala AI 是一个国际志愿者运营的研究团体，倡导人工智能创造更美好的明天，网址为 http: //amygdalaai.org/

未知或数量众多时，这些技术的表现不佳。此外，现有的大多数解决方案都是基于范围狭窄的数据集训练的，限制了它们适应多样且难以预测的声学环境的能力。

为克服这些挑战，本文提出一种以频谱图驱动的深度学习方法，用于处理混合音频样本的多标签与多类别分类。与基于梅尔频率倒谱系数的方法 [4], 不同，后者往往难以捕捉细粒度的时频变化，而所提出的基于频谱图的卷积神经网络模型能够直接学习复杂的音频模式，无需显式的声源分离。这使得该模型非常适用于南亚地区常见的多样且相互重叠的声景。

该方法在扩展版 SAS-KIIT 数据集 [5], 上进行评估，该数据集现包含 21 个南亚声音类别；同时，它也在广泛使用的 UrbanSound8K 数据集 [6]上进行测试以作比较。基于频谱图的方法表现出明显优于传统基于梅尔频率倒谱系数方法的性能，在两个数据集上均观察到更高的分类准确率。这表明频谱图分析有助于实现更准确的环境声音识别，并支持对南亚多样声学环境的记录。

本文结构如下：第3节详细介绍所使用的数据集以及所应用的预处理技术。第4节随后介绍所提出的方法。第5节呈现并讨论实验结果，并着重强调主要发现。本文在第6节作结，总结本研究工作并提出未来研究方向的建议。

2 文献综述与相关工作

许多论文已将梅尔频率倒谱系数和基于频谱图的方法应用于各类音频分类任务。在 [7],中，提出了一种基于注意力机制的卷积神经网络模型用于环境声音分类，并利用梅尔频率倒谱系数、伽马通倒谱系数、常数Q变换和色度图等多种特征类型。在 [8], 中，采用带高斯核的支持向量机以及包括短时能量、过零率和梅尔频率倒谱系数在内的特征提取技术进行音频分类。多项研究将卷积神经网络（CNN） [5]与频谱图和梅尔频率倒谱系数（MFCC）结合用于声音分类，在各类音频识别任务中显著提升了模型性能。在多标签分类领域 [9], , 许多技术已在各类数据集上得到探索与评估，以衡量其性能和效率。Tasnim Akter等人 [10] 提出了一种多标签声音分类方法，使用包括卷积神经网络、长短期记忆网络 [11], 和门控循环单元在内的深度学习模型，从音频信号中进行乐器识别。本文 [12] 提出了使用梅尔频谱图和名为Mel-Graph-GCN的深度卷积神经网络 [13]框架，根据现场录音进行鸟类物种分类的多标签分类方法。尽管环境声音分类已取得显著进展，但大多数研究依赖文化上受限的数据集，并主要聚焦于单标签任务，忽视了多标签场景的复杂性。针对这些局限，本研究致力于开发一种基于频谱图的卷积神经网络框架，专门用于南亚丰富而嘈杂的声景中的多标签分类，并同时使用区域和全球数据集，以确保更广泛的适用性和稳健性。

3 数据集与预处理

本研究使用两个数据集来评估我们的方法。SAS-KIIT[2] 收录了多样的南亚声音，包括自然、城市和人为事件。UrbanSound8K[3],是一个基准数据集，涵盖多种城市噪声类型，非常适合评估复杂环境中的分类准确率。数据集详情和预处理步骤如下所述。

3.1 SAS-KIIT 数据集

SAS-KIIT 数据集是经过整理的南亚声景集合，收录了来自印度、孟加拉国、尼泊尔、不丹和阿富汗等国家的听觉事件，用于多类别声音分类任务。

3.1.1 数据集概述

该数据集全面收集了来自不同环境的听觉事件，例如城市、乡村和工业环境，广泛呈现了真实世界声学条件。为了使

[2]SAS-KIIT：https://sas-kiit.netlify.app/ [3]UrbanSound8K：https://urbansounddataset.weebly.com/urbansound8k.html

数据集更加完整，我们加入了一些罕见的环境声音，提升了其多样性。这个扩展版数据集包含以下声音类别：

表 1：SAS-KIIT 数据集中 21 种声音类别列表

1. 坦布拉琴	12. 发射引擎
2. 传统歌曲	13. 笛子
3. 铁路发动机	14. 佛教祈祷
4. 儿童课堂噪声	15. 鱼市场
5. 簧风琴	16. 老虎
6. 达克鼓	17. 大象
7. 塔布拉鼓	18. 卡尔拜沙基风暴
8. 宣礼	19. 萨纳坦宗教阿拉蒂
9. 教堂祈祷	20. 人力车喇叭
10. 灌溉发动机	21. 阿富汗普什图音乐
11. 单弦琴	

3.1.2 组织与特征

该数据集系统地划分为 10 个文件夹，确保音频样本在训练和评估中的分布均衡。每个声音类别包含 450 个时长为 4 秒的片段，共计 9,450 个音频片段。音频文件采用 WAV格式，采样率为 44.1 kHz。除音频文件外，数据集还附有元数据标注，详细记录了片段名称、切片起止时间、类别ID、类别名称和文件夹ID等重要信息。这种结构化组织方式有助于高效处理数据和训练模型。
为更好地了解数据集的结构，文中提供了 t-SNE [14]可视化图（图 1），展示类别区分及其重叠情况，为了解数据集的分布提供了有价值的参考，并有助于在训练前进行模型解释。

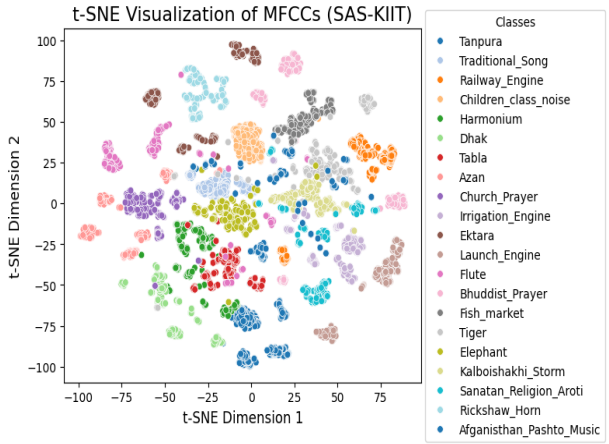


图 1：t-SNE：SAS-KIIT 数据集特征空间中的类别分布表示。

3.2 UrbanSound8K 数据集

UrbanSound8K 是环境声音识别领域中广受欢迎的数据集，提供了丰富多样的城市噪声样本。它包含从真实场景中采集的多种声音类别，适合用于评估复杂声学环境中的分类模型。

3.2.1 数据集概述

UrbanSound8K 数据集包含 1,302 条已标注的音频录音，全部采集自 freesound[4],一个用于音频共享的开源平台。每条录音均经过仔细标注，包含起止时间。该数据集保留了原始音频属性，包括编解码器、采样率和位深度，确保录音在未经修改的情况下保持其真实性与质量。该数据集包含 10 个不同的声音类别，概述如下：

表 2：UrbanSound8K 数据集中的 10 个声音类别列表

1. 空调	6. 引擎声
2. 车辆喇叭	7. 枪声
3. 儿童玩耍	8. 风镐
4. 狗吠	9. 警笛
5. 钻孔	10. 音乐

3.2.2 组织与特征

UrbanSound8K 数据集采用 10 折结构进行组织，通过交叉验证实现标准化评估。每段声音录音被分割为时长为 ≤ 4 秒的短片段，共产生 8,732 个带标签的片段。为确保音频处理的一致性，所有录音均已重采样至标准的 44.1 kHz 采样率。此外，数据集还提供了一个元数据文件，其中包含文件名、类别标签、片段起止时间以及折分配等关键信息。

为了进一步说明 数据集的结构，提供了 t-SNE（参见图 2）可视化结果，展示了类别 s
分布和关系 以及潜在的重叠情况，这些都为模型开发和评估提供了参考。

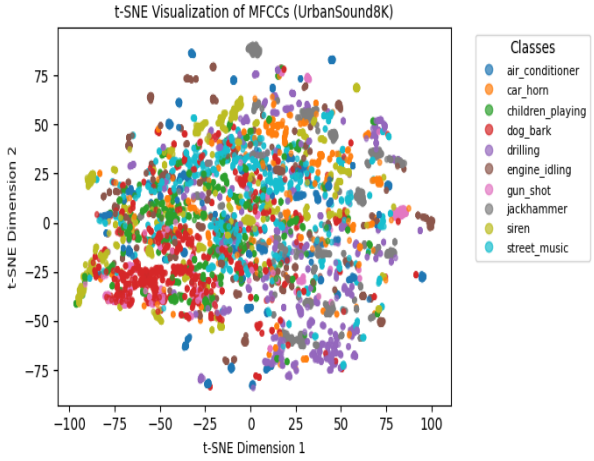
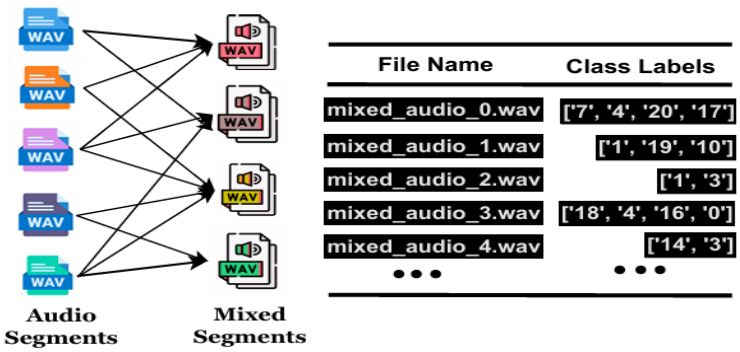


图 2：t-SNE：UrbanSound8K 数据集特征空间中的类别分布表示。

3.3 音频混合过程

在此阶段，通过组合数量固定或可变的片段来创建混合音频文件（参见图 3）。固定数量时，每个混合文件由三个片段构成；数量可变时，每个混合文件由一至四个独立音频片段合并而成。共生成 8,000 个混合音频样本。这种变化有助于模拟真实世界中的音频混合情况，使数据集更加稳健。混合文件被随机分配到 10 个文件夹中。可供参考的混合文件见此处[5]。此外，还维护了一个元数据文件，这对后续步骤至关重要，因为它便于将混合文件与对应的类别标签及文件夹分配信息关联起来。

[4]www.freesound.org [5]网址：
https://multilabelaudioclassification.netlify.app/



Fi图 3：音频混合过程与标签交互 作用。

4 提出的方法

4.1 梅尔频谱图生成

每个音频片段均以 44.1 kHz 的采样率进行采样，并使用 128 个梅尔滤波器组进行处理，最高频率为 8000 Hz。这些参数在频率分辨率与计算效率之间取得了良好平衡，使模型能够捕捉重要的音频特征，同时保持训练速度。总共生成了 8000 张梅尔频谱图，对应 8000 个混合音频样本。这些频谱图以 PNG 图像的形式存储，并分布在 10 个文件夹中；同时维护了相应的元数据文件，用于跟踪文件属性和标签。
为将每个音频信号转换为时频表示，本文使用短时傅里叶变换（STFT）将每个音频信号转换为时频表示，从而能够分析随时间变化的局部频率变化。数学上，STFT 定义如下：

$$\text{STFT}\{s[q]\}(p,\theta) = \sum_{q=-\infty}^{\infty} s[q] \cdot h[q-p] \cdot e^{-j\theta q} \tag{1}$$

其中， $s[q]$ 表示离散时间音频信号， $h[q-p]$ 表示以帧 p 为中心的分析窗函数， θ 表示角频率分量。为生成频谱图 S，将短时傅里叶变换（STFT）输出的幅度平方，具体如下：

$$S(p,\theta) = |\text{STFT}\{s[q]\}(p,\theta)|^2 \tag{2}$$

为了使频率表示与人类听觉感知相匹配，我们使用梅尔尺度滤波器组处理频谱图，将频率轴映射到符合感知规律的尺度，并着重强调人耳较为敏感的频段。由此得到梅尔频谱图，并将其用作基于卷积神经网络的分类模型的输入。

4.2 梅尔频率倒谱系数特征的计算

从每个采样率为 44.1 kHz 的音频片段中提取梅尔频率倒谱系数（MFCC）。每帧计算 40 个系数，以表示音频的频谱特征。为保持输入尺寸统一，将每个梅尔频率倒谱系数矩阵填充或截断至 400 帧。应用标准化对特征进行归一化，以提高一致性和训练性能。随后，根据元数据中的定义，将这些特征向量与对应的多标签标注配对。
梅尔频率倒谱系数通过对梅尔尺度滤波器组的对数压缩输出执行离散余弦变换（DCT）得到，如下所示：

$$\text{MFCC}(k) = \sum_{u=1}^{U} \log(R(u)) \cdot \cos\left[\frac{\pi k}{U}(u-0.5)\right] \tag{3}$$

其中， $R(u)$表示第 u个梅尔滤波器的能量， U 是所使用滤波器的总数 d.

4.3 输入准备

梅尔频谱图数据集按照 70-20-10 的比例分别划分为训练集、验证集和测试集。训练集用于模型优化，验证集用于监控中间性能并降低过拟合风险。测试集作为无偏基准，用于最终评估。频谱图图像被调整为 128×128 像素，并转换为张量，以确保与分类模型的输入要求兼容。

4.4 模型架构

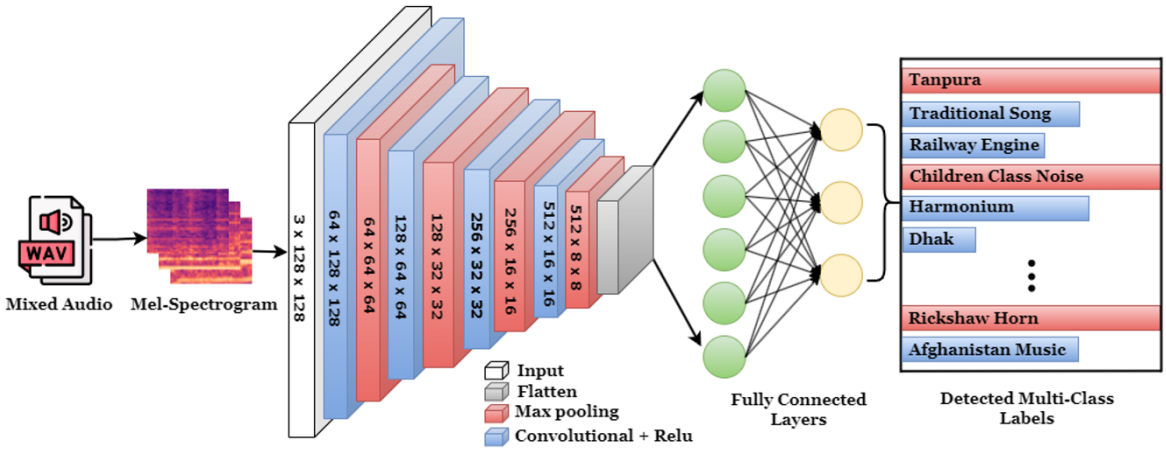


图 4：混合音频被转换为梅尔频谱图并输入带有 ReLU 和最大池化的卷积神经网络，以提取特征；全连接层输出多标签预测，示例录音中检测到的声音类别以灰粉色突出显示。

采用基于卷积神经网络的框架来解读频谱图图像，从而识别多个同时出现的声音类别。该网络结合卷积层提取相关特征、池化层缩减空间维度，并在末端设置全连接层进行最终类别预测，如图 4 所示。该架构包含以下组件：

- **卷积层：** 该架构以 64 个滤波器开头，后续层分别使用 128、256 和 512 个滤波器，均采用 3×3 卷积核和 ReLU 激活函数，以提取层次逐渐加深的频谱图特征。
- **池化层：** 最大池化层应用于卷积块之后，将空间维度从 128×128 缩减至第一次池化后的 64×64，再缩减至最后一次池化后的 32×32，以此类推，在保留关键特征的同时尽量减少过拟合。
- **全连接层：** 将展平后的特征图传入一个包含 128 个神经元（ReLU）的全连接层，以及一个包含 C 个神经元的最终输出层（其中 C 是类别数），从而生成原始逻辑值。
- **激活函数：** 中间层使用 ReLU，而输出层生成逻辑值，并在计算损失时通过 Sigmoid 函数进行处理。
- **BCEWithLogitsLoss：** 将 Sigmoid 激活函数和二元交叉熵损失合并为一个操作，无需在输出层显式使用 Sigmoid 激活函数。它能确保妥善处理多标签分类任务。下面给出了 Sigmoid 函数和 BCEWithLogitsLoss 的公式：

$$\sigma(\hat{p}_i) = \frac{1}{1 + e^{-\hat{p}_i}} \tag{4}$$

如下所示，Sigmoid 输出用于计算二元交叉熵损失：

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{L} \sum_{j=1}^{L} [p_j \cdot \log(\sigma(\tilde{p}_j)) + (1 - p_j) \cdot \log(1 - \sigma(\tilde{p}_j))] \tag{5}$$

在此公式中，$\tilde{p}_j$ 表示第 j 个标签的预测逻辑值，p_j 表示其对应的真实值。Sigmoid 函数 $\sigma(\cdot)$ 将逻辑值映射为概率。变量 L 表示输出标签的总数。这种公式非常适用于多标签分类任务，因为在训练过程中，每个标签都被单独处理。

该模型经过优化，能够从频谱图中高效学习有意义的表征，从而确保稳健的分类性能。

4.5 训练过程

模型训练进行了 100 个训练轮次，小批量大小为 16。优化采用 Adam 算法，初始学习率设为 0.001。这一设置兼顾了计算效率与稳定的梯度更新。验证过程用于评估泛化性能，其中对频谱图图像进行处理以生成原始逻辑值。随后应用阈值，获得用于多标签分类的二元标签。

5 结果分析

所提出的模型在两个数据集上分别使用梅尔频率倒谱系数和梅尔频谱特征进行评估。针对每个数据集，采用不同的数据划分策略进行测试，以评估其对分类性能的影响。为评估性能，使用精确率（P）、召回率（R）和 F1 分数，确保对各类别分类结果进行详细评估。此外，还根据阈值化预测计算准确率，以衡量模型性能。

5.1 固定混音音频数据集的结果

如前所述，该子集包含由三种不同声源混合而成的混合音频样本。在 SAS-KIIT 和 UrbanSound8K 上，基于频谱图的模型分别取得了**95.42%** 和**86.36%**的准确率。表 3 总结了这些结果。

表 3：性能　　固定数量混合音频样本的数据集上，梅尔频率倒谱系数与频谱图特征的性能比较　　.

数据集	特征	P	R	F1	Acc.
SAS-KIIT	MFCC	0.83	0.73	0.76	93.91
	频谱图	0.87	0.78	0.81	**95.42**
UrbanSound8K	MFCC	0.71	0.61	0.62	84.16
	频谱图	0.72	0.67	0.64	**86.36**

5.2 可变混合音频数据集的结果

在可变混合音频数据集中，将一至四个音频样本组合成一个混合文件。结果表明，使用频谱图特征的模型在 SAS-KIIT 数据集上的准确率达到**96.37%**，在 UrbanSound8K 数据集上的准确率达到**85.26%** 。如表4所示，在各种混合样本配置下，模型始终表现出色，凸显了其有效处理复杂真实世界音频场景的能力。

表4：不同混合音频样本数量的数据集上，梅尔频率倒谱系数和频谱图特征的性能比较。

数据集	特征	P	R	F1	Acc.
SAS-KIIT	MFCC	0.84	0.76	0.77	94.63
	频谱图	0.89	0.83	0.84	**96.37**
UrbanSound8K	MFCC	0.68	0.65	0.62	83.94
	频谱图	0.69	0.63	0.61	**85.26**

t-SNE 可视化结果（参见图 2）清楚地显示，与 SAS-KIIT 数据集相比，UrbanSound8K 数据集的类别分布更加复杂且重叠更多。这种更高的复杂性使分类变得更加

困难，导致模型在 UrbanSound8K 上的性能得分较低。图 5 展示了模型如何在混合音频输入中准确检测并区分多个声音类别，体现了其在现实场景中的实用性。

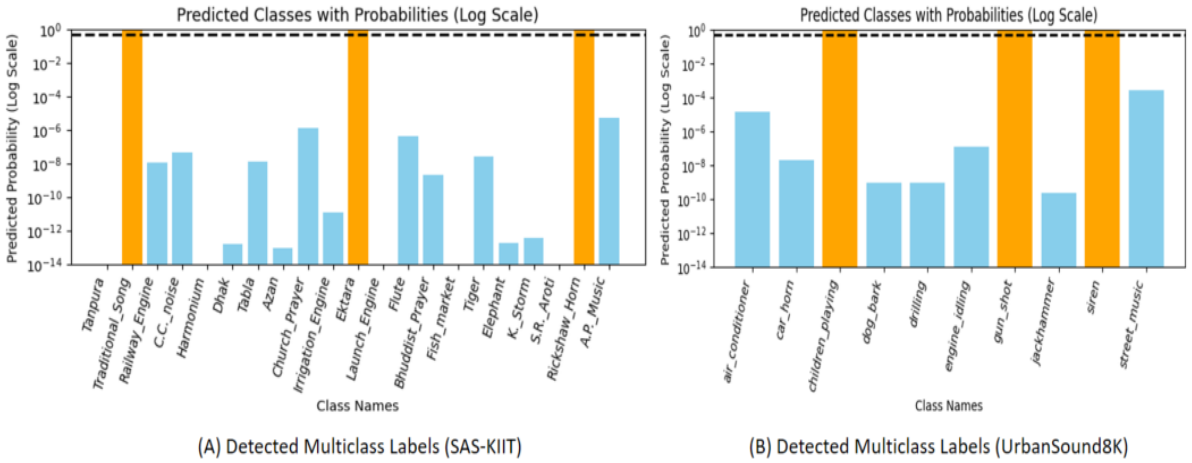


图 5：预测音频类别的对数尺度表示：橙色柱表示在样本音频中检测到的类别；天蓝色柱表示其他预测类别，其概率分数较低，并以对数尺度显示；黑色虚线（- - -）表示决策阈值，高于该阈值的类别将被视为存在。

5.3 与领先技术的比较分析

为评估所提出模型的性能，我们将其与最先进（SOTA）方法进行了比较，采用了 FACE [15] 和预训练音频神经网络（PANNs） [16] 模型，并使用从可变混合音频数据集中提取的频谱图和梅尔频率倒谱系数特征。PANNs 在通用音频分类任务中表现稳健，因而备受认可；它们基于规模庞大且多样化的数据集进行训练。FACE 是一种较新的复杂架构，专为环境声音识别而设计。如表 5 所示，所提出的模型在分类准确率方面优于 FACE，同时保持了显著更简单的架构。与 PANNs 相比，该模型在降低计算需求的同时，也展现出具有竞争力的性能。

表5：所提模型、FACE 和 PANNs 在使用频谱图特征且混合音频样本数量可变的数据集上的性能比较。

数据集	模型	P	R	F1	Acc.
SAS-KIIT	FACE [15]	0.91	0.79	0.80	95.22
	PANNs [16]	0.85	0.76	0.78	92.51
	所提出的模型	0.89	0.83	0.84	**96.37**
Urban-Sound8K	FACE [15]	0.71	0.57	0.59	84.54
	PANNs [16]	0.63	0.61	0.66	83.26
	所提出的模型	0.69	0.63	0.61	**85.26**

表 5 的结果显示，尽管 FACE 的设计复杂，但我们的模型在准确率方面优于它，同时速度更快、复杂度更低。该模型在保持计算效率的同时，实现了更高的分类准确率，因此非常适合部署在实际场景和资源有限的环境中。这一比较凸显了我们模型的高性能和实际优势。

6 结论

所提出的卷积神经网络在复杂混合环境声音的多标签分类方面表现出色。通过将频谱图用作输入特征，该网络能够有效捕捉关键的时频特征，即使单个音频样本中存在多个声源重叠，也能准确识别。在文化多样化的数据集和通用数据集上，该方法均保持较高的准确率，体现出良好的通用性和可靠性。其简洁高效的特点使其非常适合用于城市监测、文化声音景观保护，以及其他环境声音分类至关重要的现实应用。展望未来，可通过加入注意力机制或时间序列建模等先进组件来增强模型，从而更好地捕捉上下文依赖关系。该模型的适应性也使其有望部署于资源受限设备，以进行实时现场分析。未来的工作还可进一步扩展这一框架，纳入多模态数据输入，以提升其在复杂听觉场景中的表现。

参考文献

[1] M. G. Lopez P.、H. Molina Lozano、L. P. Sanchez F与L. N. Oliva Moreno，“使用独立成分分析和小波的音频信号盲源分离”，载于*CONIELECOMP 2011*，第*21*届电气通信与计算机国际会议，2011年，第152–157页。[2] M. S. Reza与J. Ma，“面向分类的独立成分分析与主成分分析集成特征提取”，载于*2016*年*IEEE*第*13*届信号处理国际会议（*ICSP*），2016年，第1083–1088页。[3] T. Bandyopadhyay，“Spectron：使用带对抗性优化的条件Transformer进行目标说话人提取”，*arXiv*预印本*arXiv:2409.01352*，2024年。[4] B. Vimal、M. Surya、Darshan、V. Sridhar与A. Ashok，“基于梅尔频率倒谱系数的机器学习音频分类”，载于 *2021*年第*12*届计算、通信与网络技术国际会议（*ICCCNT*），2021年，第1–4页。[5] R. Chatterjee、P. Bishwas、S. Chakrabarty与T. Bandyopadhyay，“南亚声音：音频分类”，载于 *2024*年第*4*届计算机、通信、控制与信息技术国际会议（*C3IT*），2024年，第1–6页。[6] J. Salamon、C. Jacoby与J. P. Bello，“面向城市声音研究的数据集与分类体系”，载于 第*22*届*ACM*多媒体国际会议（*ACM-MM’14*），美国佛罗里达州Orlando，2014年11月，第1041–1044页。[7] J. Sharma、O.-C. Granmo与M. Goodwin，“使用多特征通道和基于注意力的深度卷积神经网络进行环境声音分类”，载于 *Interspeech 2020*，2020年，第1186–1190页。[8] F. Rong，“基于机器学习的音频分类方法”，载于 *2016*年智能交通、大数据与智慧城市国际会议（*ICITBS*），2016年，第81–84页。[9] G. Tsoumakas与I. Katakis，“多标签分类：综述”，《国际数据仓库与挖掘期刊》，第3卷，第1–13页，2009年9月。[10] T. A. Onisha、J. Kim与J. Seol，“使用深度学习模型的多标签声音分类”，载于 *2024*年*IEEE/ACIS*第*22*届软件工程研究、管理与应用国际会议（*SERA*），2024年，第129–134页。[11] M. Bubashait与N. Hewahi，“使用深度神经网络、卷积神经网络与长短期记忆网络的城市声音分类：一种比较方法”，2021年9月，第46–50页。[12] N. A与R. Rajan，“使用梅尔图-GCN框架对现场录音进行多标签鸟类物种分类”，载于 *Interspeech2024*，2024年，第4793–4797页。[13] W.-C. Ye与J.-C. Wang，“基于图神经网络的多标签分类”，载于 《数据挖掘》，C. Thomas编，里耶卡：IntechOpen，2021年，第4章。 [在线]。可获取：https://doi.org/10.5772/intechopen.99681[14] L. van der Maaten与G. Hinton，“使用t-SNE可视化数据”，《机器学习研究期刊》，第9卷，第2579–2605页，2008年11月。[15] M. M. Morsali、H. -Mohammadzade与S. B. Shouraki，“FACE：快速、准确且具有上下文感知的音频标注与分类”，*arXiv*预印本 *arXiv:2303.03666*，2023年。[16] Q. Kong、Y. Cao、T. Iqbal、Y. Wang、W. Wang与M. D. Plumbley，“PANNs：用于音频模式识别的大规模预训练音频神经网络”，《*IEEE/ACM*音频、语音和语言处理汇刊》，第28卷，第2880–2894页，2020年。