训练单一模型检测单点听诊中的心音和肺音事件

Leander Melms[1,7], Robert R. Ilesan 2*、Ulrich Koehler[3], Olaf Hildebrandt[3], Regina Conradt[3], Jens Eckstein[4], Cihan Atila[4], Sami Matrood 5、Bernhard Schieffer[6], Jürgen R. Schaefer[6], Tobias Müller[6], Julius Obergassel[7], Nadine Schlicker[1] 和 Martin C. Hirsch[1]1马尔堡菲利普大学人工智能研究所；[2]巴塞尔大学医院口腔颌面外科诊所；[3]马尔堡菲利普大学呼吸与重症医学科；4 巴塞尔大学医院内科学系，瑞士巴塞尔；[5]马尔堡菲利普大学胃肠病学、内分泌学、代谢病学与感染病学系；[6]吉森和马尔堡大学医院心脏病学科；内科学系；[7]汉堡大学心脏与血管中心心脏病学系，德国汉堡 * 通讯作者：robert.ilesan@unibas.ch

摘要

目的：本研究提出一种半监督训练方法，仅使用一个训练好的模型，即可同步检测肺音和心音，且不受听诊点影响。**方法：**我们使用来自2016年PhysioNet/CinC挑战赛、2022年George Moody挑战赛以及肺音数据库HF_V1的开放获取数据。首先，我们利用不同听诊数据库中的前景真实标签（GT），训练专门的单任务模型，以识别相应肺音和心脏听诊数据库中的背景声音事件。随后，将由此生成的伪标签与真实标签结合，开展新一轮训练，从而训练出能够检测前景和背景信号的新模型。基准测试确保新训练的模型能够在不同听诊部位检测肺音和心音事件，同时不降低原有任务的性能。我们还为心音和肺音听诊中的相应背景信号建立了人工验证标签，以评估这些模型。**结果：**本研究首次报告了以下结果：i）对肺音事件进行多类别预测；ii）同步检测心音和肺音事件，并且仅用一个模型便取得了有竞争力的结果。组合多任务模型的心音检测性能略有下降，但肺音检测准确率显著提高；其在六个类别上的总体宏平均F1分数为39.2%，较单任务基线模型提高了6.7%。**结论/意义：**据我们所知，这是迄今为止首个能够测量心音和肺音事件，且同时不受听诊部位和采集设备影响的方法。因此，我们的模型能够在任意听诊位置进行肺音和心音检测。

索引词 — 声音事件检测、听诊、分割、卷积神经网络、肺、心脏、心音图、音频标注、人工智能、决策支持系统

I. 引言

通过听诊可以听到呼吸和心脏活动产生的特定声音，临床上会根据这些声音的特征及其出现情况进行分类。 基本声音源于呼吸节律（吸气和呼气）以及心脏活动（第一心音和第二心音），也可能源于病理性器官病变，例如肺部听诊中的哮鸣音和爆裂音。除了确认这些听诊事件是否存在之外，更重要的是，其出现的时间以及频率和时长数据能够提供有关患者状况的重要信息：吸气次数和第一心音（作为心动周期的起始）可用于直接确定呼吸和心率等生命体征，以及相应的节律。时长数据也能发挥重要作用：例如，呼气延长可能提示气道阻塞。此外，心脏和肺部事件出现的时间可用于推断呼吸性心律失常等情况。近期研究还表明，相对于呼吸活动而言，病理性肺音事件出现的时间可能有助于判断各种疾病病因（1）。

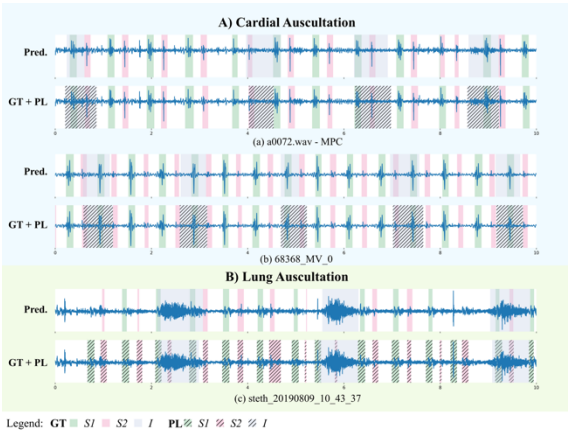

图 1。*PhysioNet/CinC 2016* 挑战赛数据集中自动分割的***A***）心音和***B***）肺音听诊文件（*10* 秒）示例，显示真实标签（*GT*）和预测标签（卷积循环神经网络）。背景事件的伪标签以条纹标示。*GT* = 真实标注，*PL* = 伪标签，*S1* =第一心音，*S2* = 第二心音，*I* = 吸气

肺音和心音在频率特征、响度、节律、事件发生情况及时长方面存在根本差异。由心脏活动引起的心音 S_1/S_2 等节律性事件（S_1 标志着收缩期开始，S_2 标志着收缩期结束和舒张期开始，见图 1）始终可以独立于肺音事件发生。在肺部听诊中，也存在遵循一定模式的倾向性节律事件，例如吸气和呼气。肺音事件——爆裂音、哮鸣音、鼾音和喘鸣——只能发生在吸气和呼气阶段。只要能听到吸气，生理上其后总会跟着呼气；同样，只要能听到第二心音，且它未被其他声音掩盖，其前面总会有第一心音。

其他所有声音事件（爆裂音、哮鸣音、鼾音、喘鸣）都不遵循这样的先后顺序。主要心音和肺音事件的具体特征可总结如下：

- **S_1/S_2：** 短促但重复出现的事件，与肺音事件相比，其声学特征变化较小
- **吸气/呼气**：时长不一，声音连续性各异
- **爆裂音**：时间短促、瞬时、不连续且爆发性强；在吸气期间，有时也在呼气期间听到；时长通常少于 20 ms (1)；受难以区分的摩擦声噪声影响；不连续，音调低或高；可发生在吸气期、呼气期或两个时期

- **哮鸣音**：时长从短到长，连续 (1)，音调高；通常比底层呼吸音更响；容易与鼾声混淆，在声谱图上通常表现为高强度的水平条纹；可发生在吸气期、呼气期或两个时期，连续且音调高

- **鼾音**：音调低、连续且具有音乐性；主导频率为 200 Hz 或更低；具有鼾声特征 (1)
- **喘鸣**：响亮且音调高；由气流湍流通过上呼吸道阻塞部位引起；主要发生在吸气期间 (1)

对音轨中的音频事件进行时间检测，并标明特定数量目标声音事件的起始和结束时间，这称为声音事件检测（SED）。在声音事件检测任务中，还会根据复杂程度作进一步的基本区分：在较简单的情形下，不同声音事件之间没有时间重叠；而在较复杂的任务中，多个不同的音频事件会同时发生，因此彼此重叠。前一种变体也称为单声道声音事件检测，后一种由于事件重叠而称为多声道声音事件检测。图 2 展示了一段多声道声音事件检测标签，其中发生了三个事件——吸气、S_1，和呼气。在时间步 t，吸气与 S_1 重叠。心音声音事件检测中的四个事件——S_1, 收缩期、S_2, 和舒张期——属于单声道声音事件检测任务，因为根据定义，这些事件不会重叠。不过，肺部声音事件检测始终属于多声部声音事件检测任务，因为事件会同时发生。

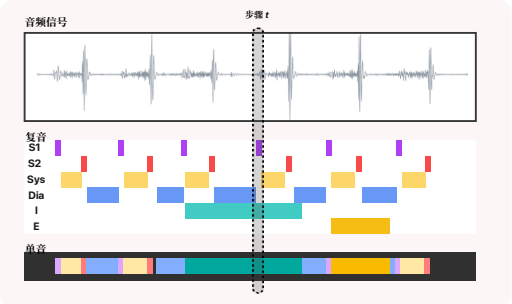

图2：三个事件类别——吸气、呼气和 S_1 ——在不同时间步的声音事件。在多声道声音事件检测中，声音事件可能重叠，如时间步 t 处所示（黑色边框矩形）。

体格检查时，心脏和肺部听诊按顺序分别进行。例如，首先在患者胸部心脏瓣膜的不同投影区进行心脏听诊（见图 3A）。随后听诊患者背部，以便更好地评估肺部。正如心肺听诊按顺序分开进行一样，现有的开源基准数据集也分别针对心脏听诊和肺部听诊（见图 3B）。在这些数据集中，标注的要么是心脏事件，要么是肺部事件，不过在许多情况下，相应的另一种信号也能在背景中隐约听到。

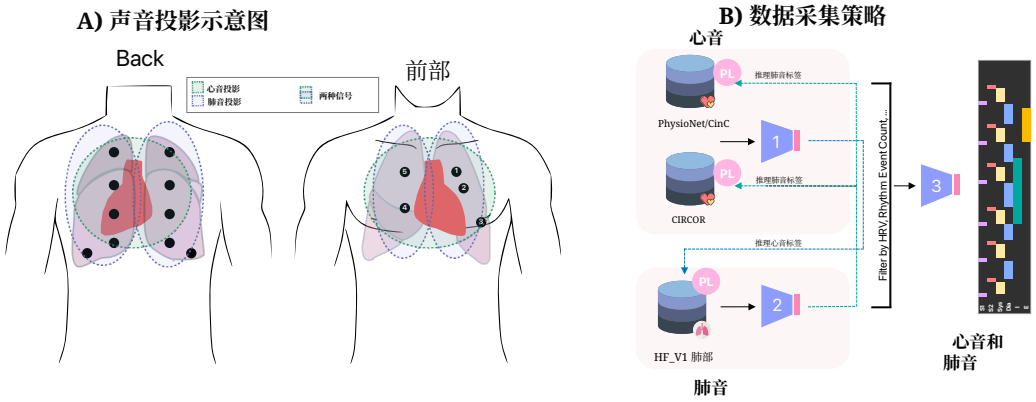

图3：**A**：肺和心脏在背部和前部的听诊点。图中显示了肺部和心脏听诊的首选听诊点。声音投影部分重叠。彩色点状投影区域示意性地表示了这种情况。 1 = 肺动脉瓣，2 = Erb点，3 = 二尖瓣，4 = 三尖瓣，5 = 主动脉瓣。 **B**：背景信号标签生成的策略。首先，使用真实标签（*GT*）训练两个独立的专用模型，分别检测心音（*1*）和肺音（*2*）事件。接下来，使用这些模型为相应的另一个数据库生成伪标签（*PL*）。然后，将真实标签和伪标签合并，用于训练能够检测心音和肺音事件的模型（*3*）。

在这项工作中，我们提出使用一个通用模型检测两种信号，不受听诊位置影响。图 3 展示了所提出的心肺声音事件检测新型训练方案。该模型基于半监督学习的相关见解：教师模型生成的输出中所包含的丰富信息可用于训练新的鲁棒模型。这样一来，所有可用的心肺听诊数据集都能用于训练。更大的数据差异性应该能够：i）提升模型的鲁棒性；ii）帮助模型更好地理解待检测事件，无论该事件是在背景还是前景中听到。此外，由于只需使用一个模型，近实时地在智能手机等边缘设备上进行推理成为可能，因而适用于连续监测应用。据我们所知，这是首个旨在训练听诊部位不变的端到端声音事件检测模型，以同时检测心音和肺音的方法。

本研究的主要贡献如下：
1. 我们提出一种受半监督学习启发的方法，为背景信号生成困难伪标签（PL）。为此，我们分别针对心肺声音事件检测训练专用模型，随后将其用于推理（见图3）。2. 我们将半监督训练应用于声音事件检测，并通过加入大量伪标记额外数据来研究性能提升。3. 我们表明，新训练的组合多任务模型在心音检测方面的性能仅略有下降，而肺音检测准确率则显著提高。

II. 相关工作

A. 声音事件检测

1) 声音事件检测模型架构。在竞赛中，卷积循环神经网络（CRNN）是声音事件检测的主流架构：其卷积神经网络（CNN）层作为特征提取器，位于时序循环层之前（2）。这种架构的优点在于能够通过循环层将丰富的局部信息关联起来，形成时序上下文。多项成功的研究都采用了这类架构（3–5）。此外，独立的卷积神经网络（6）、循环神经网络（RNN）（7），以及近期出现的基于变换器的网络（8,9）也用于声音事件检测。声音事件模型通常会针对每个类别输出其在给定时间步*t*处处于活动状态的概率。输入通常采用（梅尔）对数频谱图；为保留时间信息，后续处理中通常不进行池化，或者例如仅沿时间轴进行二分之一池化并随后插值。对于单音声音事件检测，模型的输出层采用Softmax激活函数；对于复音声音事件检测，则采用Sigmoid激活函数。之后，通常通过阈值将模型输出二值化，以获得二值活动矩阵。

2）声音事件检测不平衡： 声音事件检测中的数据不平衡涉及事件类别之间、活动与非活动时间帧之间以及时长方面的分布。为解决罕见事件与常见事件之间的不平衡，研究人员考察了数据增强（10–12）、时序子采样（12）和少样本学习（11）等方法。Imoto等人（13）证明，事件类别中活动帧与非活动帧分布的差异，对声音事件检测造成的负面影响可能大于类别不平衡。作者提出了一种新的损失函数——非对称焦点损失（AFL），通过重新加权损失，同时处理类别不平衡和活动/非活动帧不平衡。该方法使用两个参数：γ 用于平衡类别，ζ 用于平衡活动帧与非活动帧。在他们的研究中，使用 AFL（$\gamma = 0.0625$，$\zeta = 1.0$）同时处理声音事件类别之间的数据不平衡（γ）和活动帧与非活动帧之间的数据不平衡（ζ），其效果优于所有其他方法。

B 心音检测

1) 状态序列：与声音事件检测不同，心音按照定义遵循固定的状态序列（S_1-收缩期-S_2-舒张期）。因此，不可能出现重叠事件，心音检测中的事件检测始终是单音的。为保留生理状态顺序，心音检测中常采用互补方法。一种策略是强制模型的输出序列仅允许心音事件之间的有效转移（14）。然而，这种方法可能受到噪声影响，导致 S_1 和 S_2的心音事件发生偏移。其他先前研究（14–16）使用了引入状态持续时间概率的隐半马尔可夫模型（HSMM）。在 HSMM 中，标注通过扩展的维特比算法完成，该算法结合整个信号的信息来标注声音事件。然而，Messner 等人（17）发现，当心脏活动出现异常（例如心律失常）时，扩展的维特比算法容易出错。

2）模型架构：与大多数声音事件检测挑战一样，心音声音事件检测主要采用卷积神经网络、循环神经网络、时间卷积网络

C 肺部事件检测

1）模型架构： 大多数关于心肺听诊诊断支持的研究，重点是将音频文件分类为健康或病理性（20–26）、对正常呼吸音进行分类，以及对各种

附加音进行分类（26–39）。只有少数研究尝试使用强标签，精确确定声音事件的起止时间。 在肺音声音事件检测领域，近期研究对循环神经网络、卷积神经网络和卷积循环神经网络进行了基准测试（40）。Hsu 等人（40）针对每种声音事件分别训练了八种不同的模型架构，并表明卷积循环双向门控循环单元模型在检测吸气、呼气、连续附加音（CAS）和断续附加音（DAS）事件方面性能最佳（40,41）。Messner 等人也在（42）中使用双向门控循环单元模型，在一个非公开的内部数据集上同时预测呼吸阶段（吸气/呼气）以及爆裂音是否存在，爆裂音的平均基于事件的 F1 分数为 ≈ 72%，呼吸阶段事件的平均基于事件的 F1 分数为 ≈ 86%。在（43）中，Hsiao 等人使用基于注意力机制的编码器—解码器 模型进行呼吸阶段检测，并在一个非公开的内部数据集上以 100 毫秒时间窗口进行检测，吸气检测的 F1 分数为 ≈ 90%，呼气检测的 F1 分数为 ≈ 93%。 Jácome 等人（44）在一个非公开的内部数据集（3212 个吸气事件、2842 个呼气事件）上训练了更快区域卷积神经网络（Faster R-CNN）模型。该模型的吸气检测灵敏度为 97.5%、特异度为 85.0%，呼气检测灵敏度为 95.5%、特异度为 82.5%。

截至本文撰写时，尚无已知研究使用肺部或心脏听诊数据库识别复音（即在时间上同时发生）的事件。基于已发表的 HF_V1 数据集（40）开展的肺音事件识别研究中，主要的已知工作在实验中针对每种声音事件分别训练并评估了独立模型。因此，目前没有可用于比较单一模型肺部声音事件检测测试准确率的数据。诸如心肺实时监测之类的临床实用场景，对延迟有严格要求 ，因此无法并行使用多个模型进行信号处理。本研究首次探讨如何训练单一模型，使其达到有竞争力的准确率，从而支持未来的此类应用。

III. 方法

A. 肺音数据集

1) HF_肺_V1 数据库（40）：该数据集包含来自 279 名受试者的 9765 个 15 秒音频文件，是目前规模最大的开放获取肺音数据库。数据集中包含 34095 个吸气、18349 个呼气、8457 个哮鸣音、4740 个鼾音、686 个喘鸣和 15606 个爆裂音标签。音频录音采集自两个不同的研究队列，并使用了两种不同的听诊设备，即 Littmann 3200（3M，美国明尼苏达州圣保罗）和 HF_Type-1（Heroic Faith Medical Science，中国台湾台北）。与作者在 HF_肺_V1 中的实验不同（该实验针对每种声音事件分别训练模型），我们使用现有的训练-测试划分（7809 段录音用于训练，1956 段用于测试），训练了一个用于复音声音事件检测的单一模型，其输出为（*n* 个类别 *x* 个时间步）。验证集划分使用了 10% 的训练数据。

2）LEOSound 数据集：LEOSound 数据集包含 482704 段录音，使用商用呼吸音监测设备 LEOSound（Löwenstein Medical GmbH & Co. KG，德国巴特埃姆斯）录制肺音。听诊音于夜间 22:00 至 06:00 从不同年龄的不同亚组中采集。夜间睡眠监测听诊音被分割为 30 秒的录音，并通过三个粘贴式麦克风连续采集；麦克风分别贴在患者右肺和左肺背侧，以及气管侧面。该数据集除患者临床信息外，不包含其他信息或标签。我们使用两个专用的心音和肺音单任务声音事件检测模型（见图 3），生成了 18 191 736 个 S_1, 18205784 第二心音伪标签、4 095 807 个吸气伪标签、3 987 145 个呼气伪标签、536 956 个爆裂音伪标签（DAS）、298 884 个哮鸣音伪标签和 181 385 个鼾音伪标签，共计 45 497 697 个伪标签。

B. 心音数据集

1) Physionet / CinC 2016（挑战赛训练集）：
该数据集由多个较小的心脏听诊数据库合并而成，这些数据库在录音环境、录音传感器、录音时长和录音质量方面差异很大。数据集由六个数据库（a 至 f）构成，共包含来自 764 名受试者的 3153 段录音。所有录音通常会重采样至 2 kHz，时长介于 5 至 120 秒之间。该数据集为四种心动周期声音事件提供强标签：S_1-收缩期-S_2-舒张期。这些标签最初由基于 LR-HSMM 的分割算法生成，之后经过人工校正。我们仅纳入 S_1 和 S_2 事件，因为收缩期和舒张期均由这两种事件推导而来。噪声片段（以“(N”和“N)”标记）已被排除。

2) CirCor DigiScope 心音图数据集：
CirCor 数据集是规模最大的心脏听诊数据集，共包含 5272 段音频录音。听诊音采集自以儿科为主的研究人群，受试者年龄介于 0 至 21 岁；录音时长为 4.8 至 80.4 秒。与 PhysioNet/CinC 2016 数据集类似，心脏事件标注最初由三个声音事件检测模型生成，随后经过审核和校正。该数据集在 George B. Moody PhysioNet 挑战赛 2022（45）期间发布。在本研究开展时，版本 1.0.3 中[illegible]

C. 人工测试集标注

为了量化背景信号标注的表现，以及前景信号得分，我们为两个主要数据集 CirCor (46) 和 HF_V1 (40) 创建了验证集真实标签。一位受过培训的医生和一位经验丰富的研究人员通过音视频分析，判断音频录音中是否存在心音和肺音。每段听诊音录音仅由一位研究人员标注。研究人员定期召开共识会议，评估声音事件的标注，

并在存在分歧时达成共识。在心音听诊中，我们标注了吸气和呼气这两种肺音事件；在肺音听诊中，我们标注了 S_1 和 S_2这两种心音事件。

D. 预处理、特征提取

所有声音事件检测模型的输入特征都是对数梅尔频谱图，包含 64 个梅尔频带，直接从采样率为 4 kHz、时长为 10 秒的音频数据窗口中提取，未经过任何滤波，也未进行归一化。强标签信息（起始和结束时间）已转换为一个 $\{0,1\}^{N \times M}$ 多热向量，其中 N 和 M 分别表示声音片段中的时间帧数和声音事件类别数。如果某个事件类别（S_1, S_2, 吸气、呼气、哮鸣音、爆裂音、鼾音、喘鸣）处于活动状态，则将活动时间步标记为 1，非活动时间步标记为 0。

E. 训练策略

我们的半监督学习方法分为三个步骤（见图 3）：i）首先，使用心脏和肺部听诊数据库中前景信号的真实标签训练基线 CNN-BiGRU 模型；ii）使用训练好的模型为背景信号生成相应的伪标签；随后 iii）结合真实标签和带伪标签的听诊文件，在全部数据上训练新模型（见图 3）。我们针对基准数据库分别评估相应模型，并比较训练好的模型在各自类别中的测试准确率。训练期间，我们在波形域（增益、高通滤波、低通滤波、白噪声、音高移调、混响、时间丢弃、噪声注入）和频谱图域（高度／频率拉伸、频谱增强（47）、滤波器增强（48））进行了噪声注入和数据增强。最后，我们使用第三步（iii）训练好的模型，为更大规模语料库中的前景和背景信号生成共 45497697 个伪标签；该语料库包含 482704 条未标注音频文件，并将这些数据用于新一轮训练迭代。我们还创建了一个独立测试数据集，其中包含 CirCor 心脏听诊数据库（47）的肺音标签，以及 HF_V1 肺部听诊数据库（40）的心音标签，以量化对相应背景信号的检测测试准确率。

F. 训练挑战与筛选

使用伪标签和真实标签训练统一心肺声音事件检测模型时，会出现几个问题：i）域不变性（听诊器／传感器、听诊部位、环境、患者）；ii）类别不平衡；iii）不同事件类别的活动帧与非活动帧之间存在严重的数据不平衡。为确保在生成伪标签期间确实存在实际信号，我们仅在包含音频录音的情况下整合伪标签，且该录音根据生理指标能够产生可能的生命体征。对于心脏事件检测，我们仅在所得心率介于40至240之间时整合伪标签；对于肺部听诊，我们选择呼吸频率介于0至35之间的范围。在实验中，我们进一步采用了Imoto等人（13）提出的非对称焦点损失函数，以解决类别不平衡和活动/非活动帧不平衡问题。

G. 模型架构

1) CNN-BiGRU：对于基线网络架构，我们采用了 CNN-BiGRU，它被广泛用作基线声音事件检测系统，例如 DCASE2018 挑战赛任务（49）。图 4 显示，该模型由三个卷积块组成。每个卷积块包含两个卷积层（卷积核大小为 3×3）。每个卷积层之后均使用批归一化（50）和 ReLU 非线性（51）。2×1 平均池化在每个卷积块之后应用，以在频域进行下采样。图 4 中，“@”后面的数字表示特征图的数量。最后一个卷积块的三维输出沿频率轴取平均，得到二维音频特征表示，随后将其作为第一个 256 维双向 GRU 层的输入。接着，在循环层之后，应用带有 sigmoid 非线性的时间分布式全连接层，以预测每个时间步存在声音事件的概率。输出层为每个类别设置一个神经元。此外，我们还使用 BYOL-A（52）对卷积特征提取层进行无监督预训练，并将权重转换到用于声音事件检测任务的 CNN 层块中。

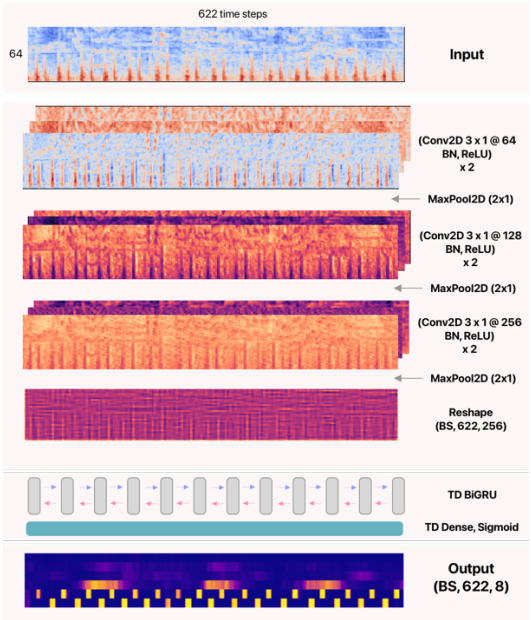

图 4： 用于多声道声音事件检测的卷积双向循环神经网络（*CRNN*）架构。为便于可视化，从模型中提取图像并放大。*CNN* 层改编自 *PANN (53)*，并使用 *BYOL- A (52)* 进行预训练。*TD* = 时间分布，*BS* = 批大小，*BN* = 批归一化。

2）时间卷积

作为第二个模型，我们采用了 Guirguis 等人（54）提出的计算高效的 SELD-TCN 框架。该框架在声音事件检测任务中的表现优于卷积循环神经网络架构，而且在 GPU 上的模型推理速度最高可快 40 倍。我们根据指定的输入调整了模型架构，并将卷积滤波器数量增加到 256。

H. 模型训练

模型使用 TensorFlow（版本 2.10.0）和 CUDA（11.2）实现。训练使用两块 NVIDIA RTX A6000 GPU。训练期间，我们从一段音频录音中随机选取一个 10 秒的窗口（622 帧）。模型输出是一个形状为（批大小、时间步、n 个类别）的三维张量。模型使用二元交叉熵（BCE）损失和非对称焦点损失（AFL）（13,55,56）训练 25 个周期，以 $\gamma = 0.0625$ 和 $\zeta = 1.0$ 作为目标，并采用 Adam (57) 优化器，迷你批次大小为 64，学习率为 $10{-}4$。使用早停法来减少过拟合。每个 10 秒音频窗口的预测输出是一个大小为 622（时间步）× 8 （类别）的序列预测矩阵。测试时，我们通过将模型预测值设为 > 0.5，将其阈值化为 1（激活），其余值设为 0（未激活）。

VI. 实验

A. 评估指标

我们使用 *声音事件检测_评估软件包*（58），在录音级别计算基于事件和基于片段的指标，以评估实验。除阈值二值化外，评估前未进行任何后处理。我们使用 F1 分数评估性能，因为它更适用于类别不平衡的数据集评估。

1) 基于事件/时间容差带的指标

当声音事件的时间位置与同一类别的真实标注事件在时间容差范围（时间容差带）内重叠时，我们便认为该声音事件被正确检测（见图 5D）。遵循既有文献中的方案，我们将心音声音事件检测的容差值设为± 60毫秒，将肺部声音事件检测的容差值设为± 500毫秒（42），因为心音和肺音事件的持续时间差异很大。我们在录音级别定义如下：

- **真正例：**已确定位置（预测）与目标位置（真实标签）之间的距离等于或小于容差窗口、因而彼此重叠的事件。
- **真阴性：**真实标签和预测标签均表明事件处于非活动状态。
- **假阳性：**预测中的事件并不存在于真实标签中。
- **假阴性：**真实标签中存在某个事件，但预测中未出现或未被识别。

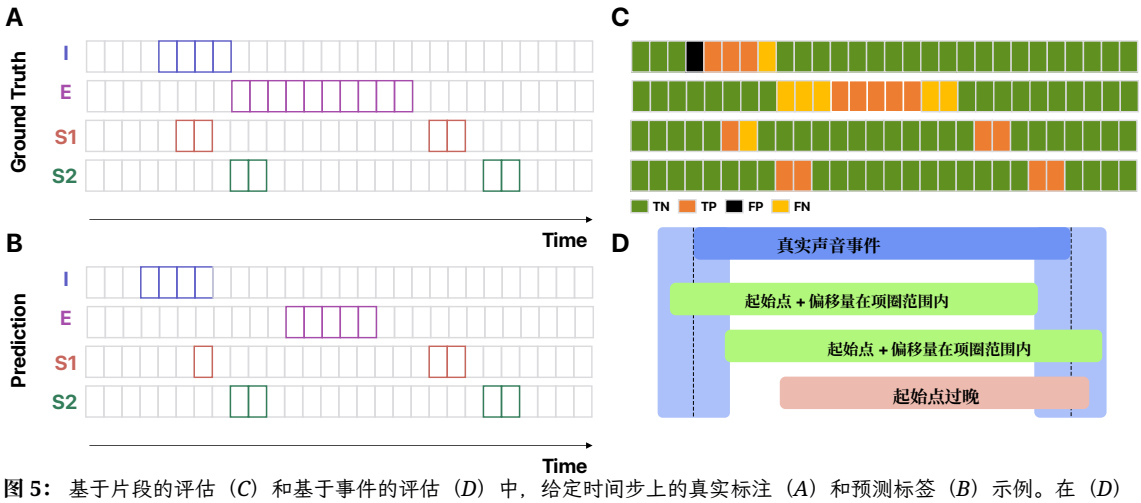

图 5： 基于片段的评估（*C*）和基于事件的评估（*D*）中，给定时间步上的真实标注（*A*）和预测标签（*B*）示例。在（*D*）中，*t* 容差时间窗以比真实声音事件更浅的蓝色显示，虚线分别标示起始和结束时间点处 *t* 容差时间窗的中心/中点。未激活的时间步以灰色显示，激活的声音事件在（*A*）和（*B*）中以彩色显示。图示改编自（*58*）。*I* = 吸气，*E* = 呼气，*S1* = 第一心音，*S2* = 第二心音，*TN* = 真阴性，*TP* = 真正例，*FP* = 假正例，*FN* = 假阴性

2）基于片段

我们首先使用真实事件标签的起始和结束时间，创建真实标注时间片段，并构建 $\{0, 1\}^{N \times M}$向量，其中 N是一次录音中的时间帧数，M是声音事件类别数（见图 5 A 和 B）。接下来，通过比较真实标注时间步（图 5 A 中，激活条以彩色显示，未激活条以灰色显示）与结果预测（图 5 B 中的条形），我们按时间步定义了：

- **真正例**（图 5 C 中的橙色条），
- **真阴性**（图 5 C 中的绿色条），
- **假正例**（图 5 C 中的黑色条），
- **假阴性**（图 5 C 中的黄色条）

随后，确定了每段音频录音中各个时间步分类的 F1分数。

3) 肺部声音事件检测指标

此外，我们采用了 HF_V1 肺数据库作者提出的肺音事件评估方案。他们采用了一种基于事件与杰卡德指数（JI）相似度的扩展评估协议（40）。按照作者的定义（40），使用 JI 评估检测到的事件与真实标注事件的重叠程度：

- **真正例**：JI 值大于 0.5，
- **假阴性**：JI 值介于 0 和 0.5 之间
- **假正例**：杰卡德指数值为 0

4) 心音检测指标

S_1 (第一心音)和S_2 (第二心音)事件用于确定心音定位的 F1分数。根据定义，收缩期介于前两个心音之间；舒张期始于第二心音结束处，并止于下一个第一心音的起始处。以往发表的研究在使用基于事件的指标评估起始和结束时间的有效性时，采用了不同的容差窗口（时间容差带）值。Springer 使用 100 ms 的时间容差带值 (16)，Schmidt 使用 60 ms (15)，Messner 使用 40 ms (17)。为了与以往研究进行公平比较，我们选择了 60 ms 的时间容差带值。

B. 结果

1) 与基线进行比较的性能评估

据我们所知，截至本文撰写时，还有两项研究（19,40）使用 HF_V1 肺部听诊数据库训练并评估了声音事件检测模型。不过，作者针对每种声音事件类别分别训练了模型。HF_V1 数据库的作者进一步将喘鸣、鼾音和哮鸣音合并为 CAS，因此总共分别评估了四个模型。据我们所知，尚无研究使用 HF_V1 数据集创建能够检测多种（重叠）肺音事件类别的模型。因此，我们首先针对六种声音事件类别（吸气、呼气、哮鸣音、爆裂音、喘鸣、鼾音）训练了一个多类别基础模型，随后训练了用于同步检测心音和肺音事件的模型，目标是在基础模型的基础上不出现性能退步。由于 Hsu 等人将肺音事件中的哮鸣音、喘鸣和鼾音合并为 CAS，我们在评估时也将它们视为同一种事件。所有针对 HF Lung V1 和 Physionet 2016 数据集训练的模型的结果见表 I。

2) 基线模型得分

采用（40）中的评估方案，我们表现最佳的基础模型在吸气、呼气、爆裂音和 CAS 上分别取得了 61.2%、34.8%、22.6% 和 23.6% 的 F1 分数（见表 I）。与基线结果相比，我们的心肺声音事件检测模型在心音检测方面略有退步，S_1 的 F1分数为 94.2%，S_2 的 F1分数为 92.7%；而在呼吸相位检测方面则有显著提升：吸气相对提升 11.6%，呼气相对提升 5.1%，但爆裂音下降 8.0%，CAS 下降 2.5%。

表 I:

基准数据库 HF_V1 和 Physionet 的实验结果。由于不同工作组对这两个基准数据集采用了不同的评估指标，因此，

	HF V1 数据集 (n = 1873，24320.94 秒，t_collar 500 毫秒)												Physionet (n = 568，12734.14 秒，t_collar 60 毫秒)			
	I			E			DAS			CAS			S_1		S_2	
模型	F1 分数			F1 分数			F1 分数			F1 分数			F1 分数		F1 分数	
	TBE	Seg.	JIE	TBE	Seg	JIE	TBE	Seg	JIE	TBE	Seg	JIE	TBE	JIE	TBE	JIE
基线 LS																
- 卷积循环神经网络 / 二元交叉熵	66.4	69.1	61.2	38.9	46.1	34.8	**28.3**	32.3	22.6	**31.1**	31.3	23.6	/	/	/	/
- 时间卷积网络 / 二	64.0	72.3	59.0	35.6	42.0	28.0	11.6	10.4	05.1	18.8	17.1	12.2	/	/	/	/
基线 HS																
- 卷积循环神经网络 / 二元交叉	/	/	/	/	/	/	/	/	/	/	/	/	**95.0**	96.9	**93.6**	94.3
- 时间卷积网络 / 二	/	/	/	/	/	/	/	/	/	/	/	/	94.6	96.5	93.2	94.1
拟议的多任务模型																
- 使用二元交叉熵的卷积循环神经网络	72.5	76.8	68.7	40.8	48.4	35.8	25.8	29.6	18.8	15.2	15.0	08.8	94.0	96.0	92.2	93.0
- 时间卷积网络，使用二元交叉熵	71.2	77.7	66.7	39.0	45.2	29.2	5.4	05.1	02.1	07.1	07.2	04.1	94.0	96.0	92.4	93.0
- **卷积循环神经网络（CRNN）+ AFL**	**72.8**	77.8	69.2	**43.3**	**49.6**	**39.9**	18.8	24.5	14.6	26.7	27.0	21.1	94.2	**96.1**	92.7	93.4
- 时间卷积网络（TCN）+ AFL	69.3	77.1	64.1	37.8	47.9	30.4	13.2	15.2	6.6	18.9	18.6	12.3	93.6	95.6	92.2	93.1
† 采用二元交叉熵的卷积循环神经网络	71.9	77.6	68.2	39.8	45.7	35.8	26.8	33.2	22.7	24.2	28.4	17.8	94.0	96.0	92.2	93.0
† 时间卷积网络（TCN）与二元交叉熵（BCE）	72.1	77.7	67.3	37.0	45.8	29.5	17.8	17.8	09.2	03.9	05.5	03.2	94.0	96.0	92.4	93.0
† **卷积循环神经网络（采用 AFL）**	72.6	**78.0**	69.2	42.5	48.8	38.2	28.0	**32.8**	**24.6**	30.5	**32.9**	**23.7**	94.2	96.0	92.7	93.4
† 时间卷积网络（采用 AFL）	67.9	74.7	62.7	34.3	39.2	25.5	12.6	07.7	04.3	01.4	01.8	0.05	93.6	96.0	92.2	93.1

† 额外数据：来自 482.704 个音频录音的 4500 万个伪标注事件

3) 评估额外数据的影响

如表 I 所示，所有使用额外数据（LEOSound 数据集，表 I 中以符号 † 标记）训练的模型，在肺音事件 DAS 和 CAS 上取得的 F1分数均高于未使用额外数据训练的模型，而吸气和呼气事件的 F1分数则基本保持不变。我们提出的采用 AFL 的 CRNN 模型在 DAS 上相对提升了 5.0%，在 CAS 上相对提升了 0.1%。与此同时，心脏事件检测的准确率并未下降，S_1 and 和 S_2 的准确率分别稳定在 94.2% 和 92.7% 左右。

4) 分类别性能评估

不同肺音事件类别的 F1分数存在显著差异。为了更详细地考察各个声音事件类别，可以比较阈值在 0 到 1 之间变化时的精确率-召回率曲线。图 6 展示了我们提出的性能最佳 CRNN 模型对所有声音类别的按声音事件分类的精确率-召回率曲线。蓝色和红色曲线分别表示测试集和验证集的精确率-召回率曲线。从图 6 可以看出，一些声音事件类别（如吸气）在多种阈值下都具有较高的精确率。还可以看出，验证集和评估集的精确率-召回率曲线趋势相似，但并不重叠，这表明验证数据与评估数据的分布存在差异，尤其是在呼气和鼾音事件方面。图 6 还表明，不同声音类别在达到最优指标（如 F1 分数）时对应的阈值各不相同。

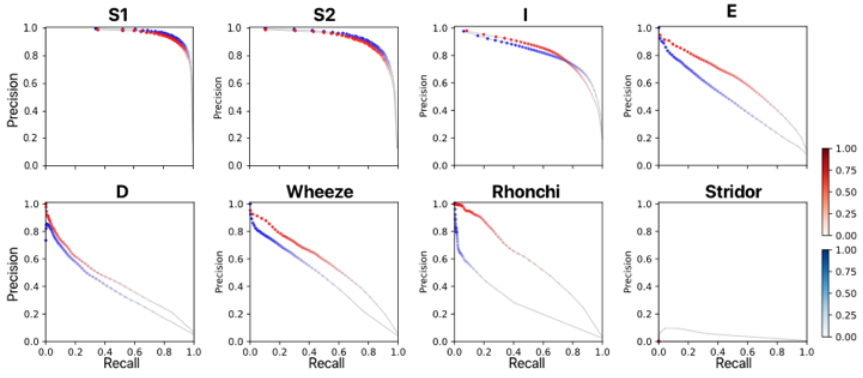

图 6： *S* 的声音事件在不同阈值下的精确率-召回率曲线*1, S2,* 吸气、呼气、*D*（非连续附加音、= 爆裂音），以及 *CAS*（= 哮鸣音+ 鼾音+ 喘鸣）检测。蓝色和红色曲线分别表示测试集和验证集的精确率-召回率曲线。该图改编自文献（*59*）。

图7也证实了这一观察结果，该图展示了阈值在0到1之间时的平均绝对百分比误差（MAPE）值。蓝色和红色曲线分别表示测试和验证的精确率-召回率曲线。可以看出，测试准确率较高的声音事件（S_1, S_2 和 I）的最优阈值约为0.5。在这些情况下，MAPE曲线呈现出相似的平缓趋势。对于测试准确率较低的声音事件，例如D、哮鸣音和鼾音，MAPE曲线则不太

对称。在此情况下，阈值低于0.5时可达到更高的准确率。对于喘鸣声音事件，当阈值大于0.5时，MAPE值达到1。

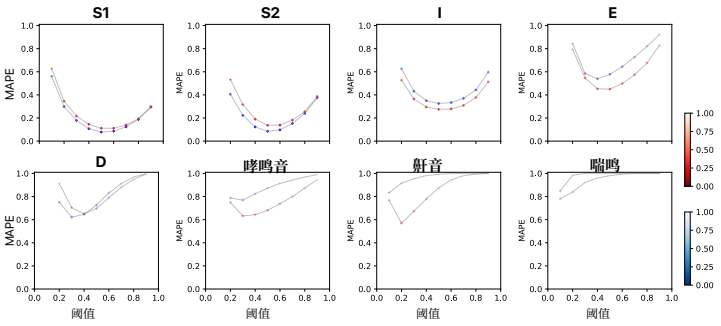

图7： 第一心音、第二心音、吸气、呼气、*D*（不连续附加音，= 爆裂音）、哮鸣音、鼾音和喘鸣事件检测的平均绝对百分比误差（*MAPE*）曲线。数值越低表示性能越好。蓝色和红色曲线分别表示测试和验证精确率-召回率曲线。该图改编自（*59*）。

5）评估多类别多任务模型与单类别模型

表II展示了 HF_V1 单类别声音事件检测的 F1 分数当前最先进成果，并与我们提出的多类别声音事件检测模型进行了比较。将采用多模型推理设置（每个声音事件类别使用一个模型）获得的结果与仅使用一个模型获得的结果进行比较，可以明显看出 F1 测试分数有所下降，尤其是在代表性不足的类别呼气、DAS 和 CAS 中。

表 IIHF Lung V1 数据集中吸气（I）、呼气（E）、连续附加音（CAS）和不连续附加音（DAS）事件检测的分段级和基于 JI 的 F1 分数。

	I		E		DAS		CAS	
	Seg	JIE	Seg	JIE	Seg	JIE	Seg	JIE
CNN-BiGRU (40)	**80.1**	85.9	62.4	68.4	69.9	**69.5**	52.6	**51.6**
时间卷积网络 (19)	n/a	**94.2**	n/a	**83.5**	n/a	**73.7**	n/a	**88.0**
OURS	78.0	69.2	49.6	39.9	32.8	24.6	32.9	23.7

6) 与最先进的心音检测模型进行比较

以往的心音声音事件检测研究采用了略有不同的评估方案。与以往的声音事件检测挑战赛一致，我们报告了通过基于事件的指标计算得出的两个事件 S_1 和 S_2 的 F1 分数（时间容差带 = 60 毫秒）。结果比较如表 III 所示。 无论与我们的基线结果相比，还是与表现最佳的心音和肺部事件检测模型相比，我们都取得了有竞争力的结果。 尤其是，与专用心音检测模型相比，多任务肺音和心音事件检测模型在 S_1 和 S_2 上的相对性能损失仅为 0.8% 和 0.9%。

表 III Physionet/CinC 2016 数据集对 S_1 和 S_2 事件进行检测的基于事件的 F1 分数。

作者	模型	S_1	S_2
		F1	F1
Messner (17)	双向循环神经网络	97.5	95.0
OURS	卷积循环神经网络（专用）	95.0	93.6
OURS	卷积循环神经网络（多任务）	94.2	92.7

7) 背景信号性能

为证明联合训练心音和肺音事件标签能够提升对背景声音事件的识别能力，我们首先在 CirCor 中另行创建的一个子集上进行了评估。在这个子集中，我们对 212 个文件中的吸气（n = 799）和呼气（n = 87）肺音进行了人工标注。采用 500ms 时间容差带进行基于事件的评估后，模型在吸气和呼气上的 F1 分数分别为 70.5% 和 42.5%，与 HF_V1（肺音前景）评估中获得的 F1 分数相当。同样，为了评估训练好的模型在肺部听诊中检测（背景）心音的能力，我们在 H1_V1（测试集划分）的随机子集中标注了 S_1 和 S_2 心音。与仅使用前景心音信号训练的模型相比，使用前景真实标签和背景伪标签训练的模型在 S_1 上的表现相对提升了 3.6%，在第二心音（时间容差带为 60 ms）上的表现提升了 3.0%。

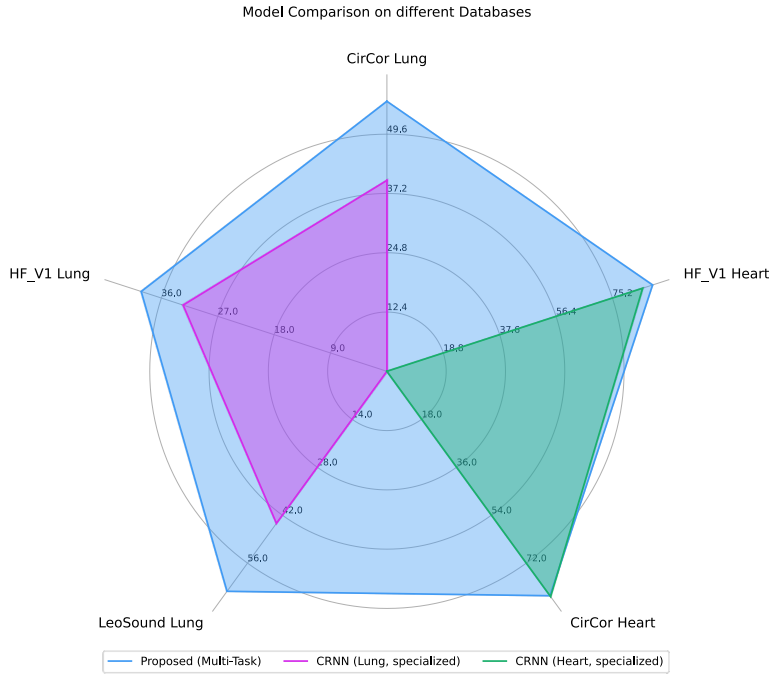

图8： 比较了我们提出的多任务声音事件检测模型与其他基线单任务专用模型在不同数据库上的表现：*CirCor* 肺部（背景心音信号）、*HF_V1* 心脏（背景心音信号）、*CirCor* 心脏（前景心音信号）、*LEOSound* 肺部（前景肺音信号）、*HF_V1* 肺部（前景肺音信号）。图中展示了各任务和数据集的分类宏平均 *F1* 分数。

图 8 总结了我们提出的多任务多类别模型在肺音和心音 SED 关键基准上的表现，并将其与单任务（专用）基线模型进行了比较。在心音 SED 方面，基线模型的测试准确率更高（前景心音信号，CirCor 数据集）；但在所有其他类别中，基线模型的表现均不如我们的模型。心音和肺部听诊数据库的联合训练方案对 HF_V1 心脏子集中的背景心音检测略有积极影响。相比之下，在 CirCor 肺部、HF_V1 肺部和 LEOSound 肺部数据集上，肺部听诊的测试准确率明显更高。

七、讨论

本研究探讨了一种新的半监督训练策略：首次合并此前彼此独立的数据集，训练声音事件检测模型同步检测心音和肺音事件。我们首先使用用于心音和肺部声音事件检测的公开挑战赛数据集，分别训练声音事件检测模型；随后利用这些模型为数据库补充背景声音事件的伪标签。尽管伪标签很可能包含大量噪声和错误，但通过合并真实标注与伪标签，我们得以训练多任务声音事件检测模型。与单任务基线模型相比，这些模型在肺部声音事件检测方面的性能没有下降，在心音声音事件检测方面也仅略有下降。我们还利用人工制作的标签证明，多任务声音事件检测模型在检测背景声音事件时取得了较高的测试 F1 分数；也就是说，在背侧肺部听诊时检测心音，或在胸部心脏听诊时检测肺音。我们观察到，在类别严重不平衡的肺音事件类别中，各类别的表现并不一致，这可能是训练期间类别不平衡以及活动/非活动帧不平衡所致。文献（40）和（19）中的先进单类别模型不受声音事件类别间类别不平衡的影响，因此取得了更均衡的测试准确率。然而，由于延迟和计算需求，在实时推理场景中，为八种不同声音事件类别分别训练声音事件检测模型并不实用。同样必须强调，出于解剖和生理方面的原因，在许多情况下，仅凭一个听诊点无法采集到所有心音和肺音。尤其对于严格局限于单侧或局部的病变，只有采用经过验证的按解剖位置设定的听诊点策略，才能确保诊断有效，例如大叶性肺炎或气胸。这意味着，模型优化永远无法解决信号缺失或不足的问题。

VIII. 结论

本研究提出了一种通用模型，无论听诊点位于何处，均可检测心肺事件。借助所提出的训练策略，我们训练了多类别多任务声音事件检测模型，通过单阶段推理获取心音和肺音事件。在检测前两个心音和吸气方面，我们取得了有竞争力的结果。结果表明，在一次训练迭代中合并不同数据库，可以训练出能够同步检测心音和肺音的声音事件检测模型。据我们所知，这是首篇关于多任务挑战的报告，探讨如何在不受听诊部位限制的情况下检测心音和肺音事件。为更好地理解这些结果的意义，并实现声音事件类别间更均衡的准确率，除了 AFL 之外，还应研究新的策略。从更广泛的角度看，我们的研究结果有望推动：i）患者，尤其是偏远地区的患者，能够以低成本进行自助听诊；ii）建立决策支持系统，帮助从业人员应对日常工作中的挑战

参考文献

1. Sarkar M, Madabhavi I, Niranjan N, Dogra M. 呼吸系统听诊. Ann Thorac Med. 2015;10(3):158–68. 2. Turpault N, Serizel R, Shah AP, Salamon J. 利用弱标注数据和声景合成检测家庭环境中的声音事件. 见 2019 [引自 2022年8月 27]. 获取地址：https://hal.inria.fr/hal-02160855 3. Miyazaki K, Komatsu T, Hayashi T, Watanabe S, Toda T, Takeda K. 用于半监督声音事件检测的卷积增强型Transformer技术报告. 2020年. 4. Jiakai L, Shanghai P. 用于DCASE 2018任务4的Mean Teacher卷积系统. 2018年. 5. Delphin-Poulat L, Plapous C. 用于DCASE 2019任务4的Mean Teacher与数据增强技术报告. 2019年. 6. Hershey S, Chaudhuri S, Ellis DPW, Gemmeke JF, Jansen A, Moore C, 等. 面向大规模音频分类的CNN架构. 见：国际声学、语音与信号处理会议（ICASSP）[网络]. 2017 [引自 2022年8月 27]. 获取地址：https://arxiv.org/abs/1609.09430 7. Hayashi T, Watanabe S, Toda T, Hori T, Le Roux J, Takeda K. 用于多声部声音事件检测的时长控制LSTM. IEEEACM音频、语音与语言处理汇刊. 2017年8月13日;25. 8. Miyazaki K, Komatsu T, Hayashi T, Watanabe S, Toda T, Takeda K. 基于自注意力的弱监督声音事件检测. 见：ICASSP 2020 - 2020年IEEE声学、语音与信号处理国际会议（ICASSP）. 2020年. 第66–70页. 9. Kong Q, Xu Y, Plumbley M. 基于CNN-Transformer和自动阈值优化的弱标注数据声音事件检测. IEEEACM音频、语音与语言处理汇刊. 2020年8月12日;PP:1–1. 10. Chen Y, Jin H. 利用深度学习和数据增强检测罕见声音事件. 见：Interspeech 2019 [网络]. ISCA；2019 [引自 2022年8月 31]. 第619–23页. 获取地址：https://www.isca-speech.org/archive/interspeech_2019/chen19_interspeech.html 11. Wang Y, Salamon J, Bryan NJ, Pablo Bello J. 少样本声音事件检测. 见：ICASSP 2020 - 2020年IEEE声学、语音与信号处理国际会议（ICASSP）. 2020年. 第81–5页. 12. Dinkel H, Yu K. 时长鲁棒弱监督声音事件检测. 见：ICASSP 2020 - 2020年IEEE声学、语音与信号处理国际会议（ICASSP）. 2020年. 第311–5页.

13. Imoto K, Mishima S, Arai Y, Kondo R. 非活动帧导致的数据不平衡以及声音时长差异对声音事件检测性能的影响. 应用声学. 2022年7月1日;196:108882. 14. Renna F, Oliveira J, Coimbra M. 用于心音分割的深度卷积神经网络. IEEE 生物医学与健康信息学期刊. 2019年1月21日;PP:1–1. 15. Schmidt SE, Holst-Hansen C, Graff C, Toft E, Struijk JJ. 基于时长依赖隐马尔可夫模型的心音录音分割. 生理测量. 2010年4月;31(4):513–29. 16. Springer DB, Tarassenko L, Clifford GD. 基于逻辑回归-HSMM的心音分割. IEEE 生物医学工程汇刊. 2016年4月;63(4):822–32. 17. Messner E, Zöhrer M, Pernkopf F. 心音分割——一种采用深度循环神经网络的事件检测方法. IEEE 生物医学工程汇刊. 2018年6月1日;PP:1–1. 18. Yin Y, Ma K, Liu M. 与抗心律失常隐藏

n

用于心音分割的半马尔可夫模型。应用科学。2020年10月11日；10：7049。19. Fernando T、Sridharan S、Denman S、Ghaemmaghami H、Fookes C。用于肺音录音事件检测的鲁棒且可解释的时间卷积网络。IEEE 生物医学与健康信息学期刊。2022年7月；26(7)：2898–908。20. Demir F、Sengur A、Bajaj V。基于卷积神经网络的肺部疾病分类高效方法。健康信息科学与系统。2020年12月；8(1)：4。21. Hosseini M、Ren H、Rashid HA、Mazumder A、Prakash B、Mohsenin T。利用听觉与人口统计信息进行肺部疾病诊断的神经网络。2020。22. Perna D、Tagarelli A。深度听诊：通过循环神经网络预测呼吸异常与疾病。2019。23. Pham L、Mcloughlin I、Phan H、Tran M、Nguyen T、Palaniappan R。用于预测呼吸异常与疾病的鲁棒深度学习框架。第2020卷。2020。164页。24. Tariq Z、Shah SK、Lee Y。使用深度卷积神经网络进行肺部疾病分类。载于：2019年IEEE生物信息学与生物医学国际会议（BIBM）。2019。第732–5页。25. Fraiwan M、Fraiwan L、Alkhodari M、Hassanin O。使用卷积神经网络和长短期记忆网络从肺音中识别肺部疾病。《环境智能与人性化计算期刊》[互联网]。2021年4月3 [引用于2022年8月27]；网址：https://doi.org/10.1007/s12652

-

021-03184-y 26. Pham L, Phan H, Schindler A, King R, Mertins A, McLoughlin I. 基于 Inception 的网络和多频谱图集成用于预测呼吸异常和肺部疾病 . In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Societ

y

(EMBC). 2021. 第 253–6 页。27. Bardou D、Zhang K、Ahmad SM。使用卷积神经网络进行肺音分类。Artif Intell Med. 2018 年 6 月 1 日；88：58–69。28. 使用优化 S 变换和深度残差网络对呼吸音进行三分类 | IEEE 期刊与杂志 | IEEE Xplore [互联网]。[引用于 2022 年 8 月 27]。可从以下网址获取：https://ieeexplore.ieee.org/document/8663379 29. Grzywalski T、Piecuch M、Szajek M、Bręborowicz A、Hafke-Dys H、Kociński J 等。人工智能算法在肺部听诊检查中的实际应用。Eur J Pediatr. 2019；178(6)：883–90。

30. Kochetov K、Putin E、Balashov M、Filchenkov A、Shalyto A。用于呼吸音分类的噪声掩蔽循环神经网络。载于：Kůrková V、Manolopoulos Y、Hammer B、Iliadis L、Maglogiannis I，编者。人工神经网络与机器学习——ICANN 2018。Cham：Springer International Publishing；2018。第 208–17 页。（计算机科学讲义）。31. Li L、Xu W、Hong Q、Tong F、Wu J。使用深度神经网络对正常肺音与附加肺音进行分类。载于：2016 年第 10 届国际中文口语处理研讨会（ISCSLP）。2016。第 1–5 页。32. Rocha BM、Pessoa D、Marques A、Carvalho P、Paiva RP。附加呼吸音自动分类：一个（未）解决的问题? Sensors. 2021 年 1 月；21(1)：57。33. Chamberlain D、Kodgule R、Ganelin D、Miglani V、Fletcher RR。半监督深度学习在肺音分析中的应用。载于：2016 年第 38 届 IEEE 医学与生物工程学会国际年会（EMBC）。2016。第 804–7 页。34. Ma Y、Xu X、Li Y。LungRN+NL：一种改进的附加肺音分类方法，使用噪声

n-

采用Mixup数据增强的局部块ResNet神经网络。2020。35. Ma Y, Xu X, Yu Q, Zhang Y, Li Y, Zhao J, 等。LungBRN：一种采用双ResNet深度学习算法检测呼吸系统疾病的智能数字听诊器。2019。36. Ngo D, Pham L, Nguyen A, Phan B, Tran K, Nguyen T. 用于预测呼吸音异常的深度学习框架。载于：2021年国际电气与电子工程研讨会（ISEE）。2021。第42–7页。37. Nguyen T, Pernkopf F. 使用卷积神经网络快照集成进行肺音分类。第2020卷。2020。38. Nguyen T, Pernkopf F. 使用迁移学习和多输入卷积神经网络进行肺音爆裂音检测。

ut

卷积神经网络。2020。39. Zhao X, Shao Y, Mai J, Yin A, Xu S. 基于BiGRU注意力网络与XGBoost的呼吸音分类。载于：2020年IEEE生物信息学与生物医学国际会议（BIBM）。2020。第915–20页。40. 在自主开发的开放获取肺音数据库上，对用于呼吸相位和附加音检测的八种循环神经网络变体进行基准测试——HF_Lung_V1 | PLOS ONE [互联网]。[引自2022年8月27]。获取地址：https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0254134 41. Hsu FS, Huang SR, Huang CW, Cheng YR, Chen CC, Hsiao J, 等。自动化肺音分析数据库的持续扩展与更新。2021。42. Messner E, Fediuk M, Swatek P, Scheidl S, Smolle-Juttner FM, Olschewski H, 等。使用深度双向门控循环神经网络进行肺音爆裂音与呼吸相位检测。载于：2018年第40届IEEE医学与生物学工程国际年会

d

生物学会（EMBC）。2018。第356–9页。43. Hsiao CH、Lin TW、Lin CW、Hsu FS、Lin F、Chen CW 等。利用基于注意力的编码器-解码器架构和迁移学习技术进行呼吸音分割与检测。第2020卷。2020。第754页。44. Jácome C、Ravn J、Holsbø E、Aviles-Solis JC、Melbye H、Ailo Bongo L。用于肺音呼吸相位检测的卷积神经网络。Sensors。2019年4月15日；19(8)：1798。45. George B. Moody PhysioNet 挑战赛 [互联网]。George B. Moody PhysioNet 挑战赛。[引用于2022年8月27]。网址：https://moody-challenge.physionet.org/2022/

46. Oliveira J、Renna F、Costa PD、Nogueira M、Oliveira C、Ferreira C 等。CirCor DigiScope 数据集：从杂音检测到杂音分类。IEEE 生物医学与健康信息学期刊。2022年6月；26(6)：2524–35。

47. Park D, Chan W, Zhang Y, Chiu CC, Zoph B, Cubuk E 等. SpecAugment：一种用于自动语音识别的简单数据增强方法。2019. 2613 页。

48. Nam H, Kim SH, Park YH. Filteraugment：一种声学环境数据增强方法。载于：ICASSP 2022 - 2022 IEEE声学、语音与信号处理国际会议 (ICASSP)。2022. 第 4308–12 页。49. Serizel R, Turpault N, Eghbal-zadeh H, Shah AP. 家庭环境中的大规模弱标签半监督声音事件检测。载于：DCASE。2018.

50. Ioffe S, Szegedy C. 批归一化：通过减少内部协变量偏移加速深度网络训练 [互联网]。arXiv；2015 [引用于 2022 年 9 月 20]。可获取自：http://arxiv.org/abs/1502.03167

51. Nair V, Hinton GE. 修正线性单元改进受限玻尔兹曼机。载于：第27届国际机器学习会议论文集。美国威斯康星州麦迪逊：Omnipress；2010年。第807–14页。（ICML'10）。

52. Niizumi D, Takeuchi D, Ohishi Y, Harada N, Kashino K. 音频BYOL：探索预训练通用音频表示。2022年。

53. Kong Q, Cao Y, Iqbal T, Wang Y, Plumbley M. PANNs：面向音频模式识别的大规模预训练音频神经网络。2019年。

54. Guirguis K, Schorn C, Guntoro A, Abdulatif S, Yang B. SELD-TCN：声音事件定位 & 通过时间卷积网络进行检测。载于：2020年第28届欧洲信号处理会议（EUSIPCO）[互联网]。2021 [引用于2022年8月27]。第16–20页。可查阅：http://arxiv.org/abs/2003.01609

55. Çakır E, Parascandolo G, Heittola T, Huttunen H, Virtanen T. 用于多声部声音事件检测的卷积循环神经网络。IEEEACM Trans Audio Speech Lang Process. 2017 Jun;25(6):1291–303.

56. Imoto K. 使用部分缺失通道的多通道观测进行声学场景分类。载于：2021年第29届欧洲信号处理会议（EUSIPCO）。2021. 第875–9页。

57. Kingma DP, Ba J. Adam：一种随机优化方法 [互联网]。arXiv；2017 [引用于2022年9月20]。可从以下网址获取：http://arxiv.org/abs/1412.6980

58. Mesaros A, Heittola T, Virtanen T. 多声部声音事件检测的指标。应用科学。2016 Jun;6(6):162.

59. Kong Q、Xu Y、Wang W、Plumbley MD. 使用 CNN-Transformer 和自动阈值优化对弱标注数据进行声音事件检测 [Internet]. arXiv；2020 [引用于 2022 年 12 月 30]. 获取地址：http://arxiv.org/abs/1912.04761