arXiv:2603.11241v1 [eess.AS] 11 Mar 2026

咳嗽活　　动检测用于自动结核病筛查

Joshua Jansen van V¨uren[1], *Devendra Singh Parihar*[1],*Daphne Naidoo*[2],*Kimsey Zajac*[3],*Willy Ssengooba*[4], *Grant Theron*[2], *Thomas Niesler*[1]

[1]电气与电子工程系，斯泰伦博斯大学，南非斯泰伦博斯　　ca

[2]DSI-NRF生物医学结核病研究卓越中心、南非医学研究理事会结核病研究中心、分子生物学与人类遗传学系、医学与健康科学学院、斯泰伦博斯大学，南非开普敦

[3]德国哥廷根大学商学与经济学学院，德国哥廷根

[4]马凯雷雷大学健康科学学院医学微生物学系，生物医学研究中心（MAKBRC），乌干达坎帕拉

摘要

通过确定起止点，在音频中自动识别咳嗽片段，对于构建可扩展的健康技术筛查工具，以用于肺部相关疾病，至关重要。我们提出将两种当前的预训练架构应用于咳嗽活动检测任务。研究采用了来自南非和乌干达社区级护理中心的录音数据集，其中包含自行前来就诊且有结核病（TB）症状的患者咳嗽声。使用 XLS-R 自动确定起止点后，测试集的平均精度达到 0.96，受试者工作特征曲线下面积达到 0.99。我们表明，仅使用网络的前三层即可取得最佳平均精度；这还具有降低计算和内存需求的双重优势，对基于智能手机的应用至关重要。结果显示，这种 XLS-R 配置在测试集平均精度上分别绝对领先音频频谱图变换器（AST）和逻辑回归基线 9% 和 27%。此外，使用 XLS-R 自动分离出的咳嗽声训练的下游结核病分类模型，显著优于使用 AST 分离出的咳嗽声训练的模型；与使用真实标注咳嗽训练的分类器相比，性能也仅略逊一筹。我们得出结论，大型预训练 Transformer 模型可有效识别咳嗽的终止点，并且将此类模型集成到筛查工具中是可行的。

索引词— 咳嗽活动检测、结核病筛查、Transformer模型、自动咳嗽分割、XLS-R、AST

1. 引言

收集咳嗽声以供移动健康（移动健康）技术使用，是一个日益受到关注的领域 [1,2]。多项研究通过在线众包或在临床环境中收集了咳嗽声。这些研究已被用于对多种肺部相关疾病进行分类，例如肺炎 [3], 、冠状病毒病 [4, 5, 6],、结核病 [7, 8], 、慢性阻塞性肺疾病 [9, 10], 和哮喘 [11]。通常，包含咳嗽声的录音需人工标注。尽管先前的工作已利用基于机器学习的方法在音频信号中识别咳嗽声，但据我们所知，尚无研究考虑此类自动咳嗽检测对下游疾病分类的影响 [1]。

在临床环境中，人工标注咳嗽声可能并不可行，因为这需要耗费时间和精力，还涉及额外的卫生问题。因此，能够从录音中自动获取咳嗽声的起止点颇具吸引力：这既可避免人工介入，也可纳入基于音频的疾病筛查工具的处理流程。在这种情况下，从音频录音中准确分离咳嗽声至关重要，因为混入输入中的非咳嗽信号可能对后续处理产生不利影响。本研究明确比较了将自动提取的咳嗽声输入结核病分类模型，与向同一模型提供由人工标注的咳嗽声这两种方式所产生的效果。

已有多位作者研究了在音频信号中识别咳嗽声。在 [5]中，使用接收者操作特征曲线下面积（AUC）达到 $96.4 \pm 3.3\%$，其任务是判断一段音频录音中是否存在咳嗽。该方法采用基于 XG-Boost 的分类模型，输入为 68 种不同特征，例如梅尔频率倒谱系数和功率谱密度、

我们感谢欧盟支持的EDCTP2计划提供资金（RIA2020I-3305，CAGE-TB）。本项目还得到了全球健康EDCTP3及其成员的支持（101145817，4-CAGE-TB）。

以及过零率。然而，这些研究未进行活动检测实验。使用支持向量机（SVM）对咳嗽帧与非咳嗽帧进行分类时，逐帧AUC达到90.7%；输入特征为若干频谱特征（质心、平坦度、带宽等）的移动平均值和标准差 [12]。研究者提出了一种基于均方根误差的咳嗽分离策略，并用其从一个开源语料库中提取咳嗽声；但由于缺少真实值人工标注，无法确定该分类器的性能 [13]。在 [14] 中，研究人员比较了卷积神经网络（CNN）与编码器-解码器循环神经网络（RNN）在咳嗽声和语音判别方面的表现。两种模型的性能相当，逐帧AUC分别达到95%和96%。使用SincNET [15], , 在 [16] 中对来自COUGHVID的100段人工标注录音子集进行测试，逐帧AUC达到98.7 % [5]。

尚未有研究探讨将基于Transformer的模型应用于音频录音中的咳嗽活动检测任务。我们研究了 XLS-R [17], , 这是一种拥有 3 亿个参数的 Transformer，采用自监督方式在涵盖 128 种以上语言、时长超过 40 万小时的语音上训练。我们将 XLS-R 与一种成熟的基于频谱图的Transformer——AST [18]——进行性能比较。AST 预训练所用的语料库由一般声音构成， [19] 与 XLS-R 不同，它并非仅在语音上进行预训练。这是一个有趣的比较，因为两种模型不仅采用了不同的预训练数据，特征提取策略也有所不同。尽管 Transformer 通常比传统机器学习模型或特定领域的深度学习模型计算开销更大，但其显著的性能提升通常足以弥补这一点。为验证这一假设，并提供一个额外的基线，我们还研究了一种逻辑回归（LR）模型，其配置旨在模拟时延神经网络 [20]。

因此，本文的目标有两个方面。首先，我们旨在确定基于变换器的网络是否适用于咳嗽活动检测，并将其性能与基线模型进行比较。其次，我们研究与人工标注相比，使用自动分离出的咳嗽声是否会降低结核病分类性能。

本文其余部分的结构如下。第 2 节介绍本研究所用的数据集。第 3 节描述实验设置，第 4 节则给出咳嗽活动检测和结核病分类结果。最后，第 5 节总结全文。

2. 数据

本研究使用了来自南非和乌干达 1193 名患者的咳嗽录音，这些录音采集自一个更大型项目，该项目旨在识别有结核病风险、且可以随后仅凭咳嗽声转介接受确诊测试（CAGE-TB、4-CAGE-TB）。

该数据集包含来自南非患者的 464 段录音，采集自三个不同的社区级护理中心；还包含来自乌干达患者的 729 段录音，同样采集自三个社区级护理中心。录音流程首先由医护人员报出患者识别编号以及当前日期和时间，以便记录并在之后进行质量控制。随后，患者按照要求从一数到十。最后，患者按要求咳嗽数次，之后医护人员说明录音已完成。所有录音中的数数和咳嗽部分均使用 ELAN [21] 软件包进行了人工标注。该数据集总计包含 21,808 个端点标注咳嗽声，相当于 2.52 小时的咳嗽音频和 20.2 小时的其他音频。在整个数据集中，咳嗽时长的均值和标准差为 416 ± 207 毫秒。

需要注意的是，由于这些录音是在繁忙的初级卫生保健机构中采集的，且采集地点通常在室外，因此机动车辆、建筑施工或发电机等产生的背景噪声十分常见。标注人员只需在咳嗽声被背景噪声掩盖、无法辨别时将其忽略。我们将这些标注用作真实标注目标，训练神经网络检测与背景噪声明显区分的咳嗽声。此外，必须在这类环境中评估这些系统，因为这类环境能代表结核病分诊或筛查在实际中开展的场所。

图1展示了所有咳嗽音频和非咳嗽音频的平均功率随频率的变化。显然，在该数据集中，咳嗽声的平均功率高于其他信号成分。值得注意的是，大多数咳嗽声功率集中在8kHz以下。因此，采用采样频率固定为16kHz的预训练Transformer模型是可行的。

我们将数据划分为训练集、开发集和测试集，如表1所示。训练和开发数据全部来自乌干达子队列，其中约75%用于训练，其余用于开发。测试集完全由南非子队列的数据组成。这一选择的前提有两方面。首先，它构成了具有挑战性的测试条件，因为各集合的录音地点互不相同，因而带来了全新的环境条件。此外，录音中使用的语言也有所不同：乌干达录音中出现英语和卢干达语，而南非录音中出现南非荷兰语、英语和科萨语。选择的第二个原因是，咳嗽检测器可以在乌干达子队列上训练和优化，然后应用于南非测试子队列。自动起止点共同构成一个包含自动分离的咳嗽声的数据集。该数据集随后可用于训练结核病分类

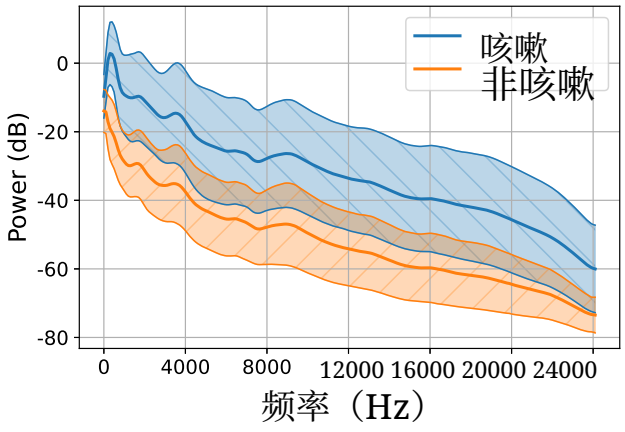


图1. 所有咳嗽（蓝色）与非咳嗽（橙色）音频的功率（dB）随频率（Hz）的变化，阴影区域表示语料库上的标准差。

表1. 数据集的训练、开发与测试划分统计。训练集和开发集仅包含在乌干达采集的数据，而测试数据仅包含在南非采集的录音。

Part	数量 录音	时长 录音	数量 咳嗽声	时长 咳嗽声
训练集	583	13.20	11890	1.37
Dev	146	3.31	3015	0.36
Test	464	7.20	6874	0.79
总计	1193	23.71	21791	2.52

模型。通过这种方式，可以评估不同咳嗽活动检测模型对结核病分类性能的影响。

3. 实验设置

咳嗽活动检测模型经过训练，可逐帧预测录音中是否存在咳嗽声。如图 2 所示，每个模型（AST、XLS-R 和逻辑回归）都会对输入音频进行特征提取，从而生成一系列特征向量 $\boldsymbol{x}^{(n)}$，每个向量均基于一个输入帧，该帧以时间步 $t^{(n)}$ 为中心，帧移为 T。分类器根据这些特征向量，为每一帧生成一个标量分数 $z^{(i)}$。这些输出分数经 sigmoid函数转换后，用于估计后验概率，记为 $\hat{} y^{(t)}$，表示由起始和结束时间边界 $[t^{(n)} - \frac{T}{2}$和 $t^{(n)} + \frac{T}{2}]$ 分别界定的音频片段中存在咳嗽的概率。模型训练所用的标签由相应时间步的人工标注确定；每帧的标签取决于占据该帧时长大部分（$> 50\%$）的类别帧时长。

通过应用决策阈值，将逐帧分数转换为二值指示，从而确定咳嗽的起止时间。随后，将连续且被判定为包含咳嗽的帧序列标记为一次咳嗽，如图 2 所示。在此示例中，咳嗽的起止时间分别为第一帧的起始时间（$t^{(1)} - \frac{T}{2}$）和最后一帧的结束时间（$t^{(3)} + \frac{T}{2}$）。在后续实验中，还会对该二值指示序列应用中值滤波。未进行其他后处理，例如最小咳嗽持续时间和最小咳嗽间隔处理。

由于 AST 和 XLS-R 均为预训练模型，帧长和帧移是固定的，无法自由选择。对于 AST，默认的特征提取策略采用 128 维梅尔间隔滤波器组。从该梅尔频谱图的 16×16 补丁中提取并拼接而成的一系列向量作为 AST 的输入。随后以时间和频率两个方向上均为 10 的步长提取补丁，这对应于 160ms 的帧长和 100ms 的帧移。XLS-R 直接接收输入波形，并将其送入卷积神经网络，该网络的卷积核大小和步长被设计为产生 25ms 的帧大小和 20ms 的帧移。对于逻辑回归，我们选择采用与 AST 相同的补丁划分方式。因此，AST 和逻辑回归架构每 100 ms 产生一次输出分类，而 XLS-R 每 20ms 产生一次输出。

3.1. 咳嗽活动检测模型

我们使用 AdamW [22]优化器，在训练集（表1）上对所考虑的每个变换器进行微调。学习率在前 10% 的训练步数中从零线性增加到设定的最大值，随后在剩余训练步数中线性衰减至零。所有模型均训练 16 个轮次。

对于这两种变换器网络，最后一个 Transformer 层之后均接入一个两层神经网络，以输出咳嗽后验概率。该网络以 Transformer模型的隐藏状态向量作为输入。对于 XLS-R，网络的第一层将维度减半（从 1024 降至 512），随后应用 GeLU [23]激活函数。最后一层接收由此得到的 512 维特征，并生成一个二元输出，用于指示该帧中是否检测到咳嗽。

对于 XLS-R，AST 计算出的 768 维输出向量由两层网络中的第一层转换为 384 维向量，随后应用 GeLU 激活函数。考虑到频率维度上的步幅，$\lfloor \frac{128-16}{10} \rfloor + 1 = 12$ 从梅尔频谱图中同一组 16 帧的部分重叠频率内容中提取输出片段。最终两层网络第一层输出的相应 12 个向量

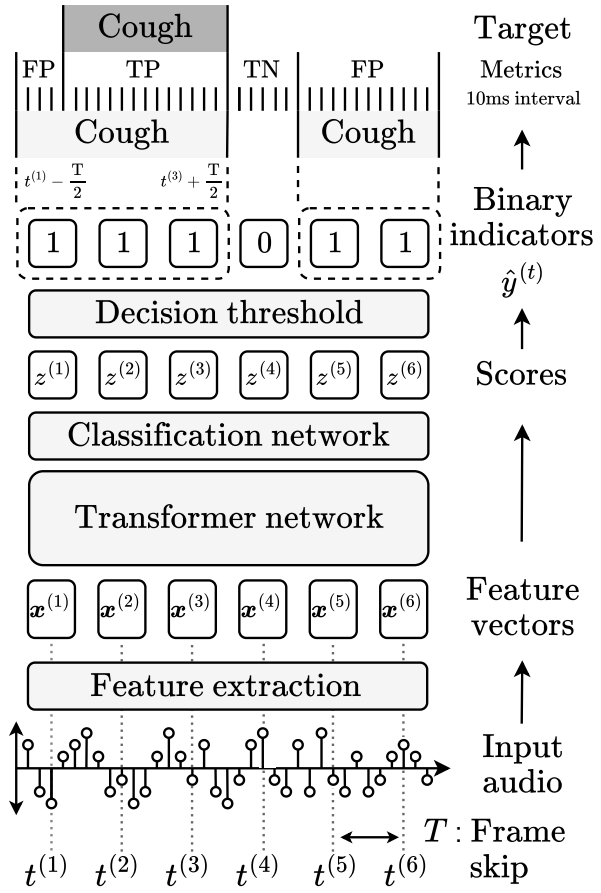


图 2。咳嗽活动检测模型架构，以及咳嗽起止点判定策略概览。TP：真阳性，FP：假阳性，TN：真阴性。

随后拼接成一个 4608 维向量，作为最终线性层的输入，从而再次生成单个二元输出。

最后，作为基线模型，我们纳入逻辑回归（LR）。该模型接收上述梅尔频谱图的一部分作为输入，并输出每帧出现咳嗽的后验概率估计。该分类器同样训练 16 轮，但不使用学习率调度器。

3.2. 评估指标

咳嗽活动检测模型使用逐帧曲线下面积、平均精度（AP），以及覆盖率（灵敏度）和纯度（精确率）进行评估。由于每个模型的输出速率不同，如图 2 所示，逐帧评估指标均按固定的 10 毫秒间隔计算。因此，对于 AST 和 LR，每个输出预测被划分为 16 个间距相等且互不重叠的区间，并分别进行评分；对于 XLS-R，则划分为两个区间。这样做是为了使模型的输出速率不会影响接受评分的分类总数，从而确保各项数值可以直接比较。选择较短的 10 毫秒间隔，是为了确保预测的咳嗽边界始终与评分区间的边界对齐，因为所有真实值的起止时间均精确到最接近的 10 毫秒。

覆盖率定义为 $\frac{\text{TP}}{\text{P}}$，其中 TP（真阳性）是被正确识别为咳嗽的帧，P（阳性）是真实咳嗽帧的总数。纯度定义为 $\frac{\text{TP}}{\text{PP}}$，其中 PP $=$ TP $+$FP 是被正确或错误识别为咳嗽的帧数，FP 是假阳性，PP 是预测为阳性的帧数。

平均精度是衡量咳嗽活动检测效果的实用指标，因为它直接反映了模型能够检出的真实咳嗽声比例（通过覆盖率 $\frac{\text{TP}}{\text{P}}$），以及在不同判定阈值下，被识别为咳嗽的音频中确实包含咳嗽声的比例（通过纯度 $\frac{\text{TP}}{\text{PP}}$）。性能良好的咳嗽活动检测器应能检出较大比例的真实值咳嗽帧，同时产生较少的假阳性。所研究的数据集存在严重的类别不平衡（$\approx$ 1:10 个阳性对应一个阴性）。因此，假阳性率（FPR $=$ $\frac{\text{FP}}{\text{N}}$）被纳入 ROC 曲线并进而用于计算 AUC 分数，但由于分母中的阴性数量（N）较大，其数值可能偏低。平均精度将纯度纳入考量，因此会着重减少相对于真阳性数量的假阳性。

下游结核病分类模型仅根据曲线下面积进行评估，因为这是诊断检测中最常用的指标。

3.3. 超参数搜索

XLS-R 和 AST 的超参数通过网格搜索在以下取值中进行优化：

- Batch size: 8, 16, 32
- Learning rate: $1 \cdot 10^{-4}$, $1 \cdot 10^{-5}$

对于 XLS-R 和 AST，还将用于提取特征的 Transformer 层视为超参数。研究了全部 24 个可用层，并且对于 XLS-R，也考虑了直接从 CNN 特征提取器获取特征。

线性分类器所考虑的超参数为：

- 批量大小为 4、8、16
- 学习率：$1 \cdot 10^{-2}$, $1 \cdot 10^{-3}$

所有超参数均根据开发集上的最佳平均精度选取。

3.4. 结核病分类模型

如图 3 所示，结核病分类模型使用通过咳嗽活动检测模型自动识别的孤立咳嗽进行训练，

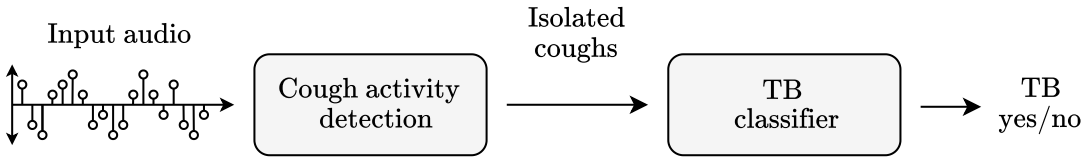


图3. T　从原始音频到下游结核病分类的自动咳嗽分离顶层示意图　　.

或者，为了进行比较，使用人工识别的咳嗽进行训练。

先前研究表明，双向长短期记忆（LSTM）网络应用于结核病咳嗽分类时效果显著 [24]。我们仅考虑这一架构，以确保观察到的结核病分类性能差异能够合理地归因于咳嗽活动检测的差异。LSTM 的优化超参数如下：

- 批量大小：8、16、32
- 学习率：$1 \cdot 10^{-4}, 1 \cdot 10^{-5}$
- 隐藏向量维度：32、64
- LSTM层数：1、2
- Dropout：0.2、0.6

对于结核病分类，南非咳嗽数据被划分为 10 个相互独立的折，每折均反映总体结核病患病率（约为 18.5%）。根据平均开发集损失，分别为这些折独立选择超参数。为进一步确保下游结核病分类器的一致性，超参数根据人工标注数据集进行选择。随后，在通过自动咳嗽检测获得的数据集上训练的分类器会直接使用这些超参数，不作任何修改。

4. 结果

在以下各节中，第 4.1 节首先报告跨队列咳嗽活动检测的结果。第 4.2 节随后讨论对每位患者可用的完整录音应用咳嗽活动检测后所获得数据集的特性，并将其与人工标签进行比较。第 4.3 节介绍应用中值滤波后的进一步结果。最后，第 4.4 节报告分别使用自动检测得到的孤立咳嗽或真实值孤立咳嗽训练下游结核病分类器时的性能。

4.1. 咳嗽活动检测

表 2 报告了使用第 3.3 节所述搜索流程得到的超参数值。该表显示，批量大小在不同模型间是一个稳定的超参数，且较小的值始终能带来更好的性能。

此外，对于两种基于Transformer的模型，较早的层（AST 为第 1 层，XLS-R 为第 3 层）在平均精度方面表现更好。这一结果令人鼓舞，因为这意味着可以舍弃网络后续的层，大幅缩小模型规模；考虑到模型未来可能部署在移动设备上，这是一项优势。对于 XLS-R，这使模型规模缩小至原来的六分之一，并通过 $3.82\ [3.82, 3.83]$ 提高了吞吐量（在 Nvidia A4000 RTX 上重复运行 10,000 次的平均值，置信区间为 95%）。

表 2。三个模型（LR、AST、XLS-R）的最佳超参数。

模型	批次	LR	层	参数量 Full	最优
LR	8	0.01	-	2k	-
AST	8	0.0001	1	87M	9.5M
XLS-R	8	0.0001	3	316M	51M

表3展示了这些优化模型在开发集和测试集上的曲线下面积及平均精度，图 4则展示了受试者工作特征曲线，以及覆盖率和纯度曲线。由表3可见，XLS-R在开发集和测试集上的曲线下面积和平均精度均始终表现最佳。在开发集上，其平均精度分别比 AST 高出 8 个百分点、比 LR 高出 11 个百分点；在测试集上，则分别比 AST 高出 9 个百分点、比 LR 高出 27 个百分点。尽管 LR 因计算效率高而颇具吸引力，但它似乎无法很好地泛化到测试条件，在开发集到测试集的过程中，平均精度下降了 10 个百分点。覆盖率和纯度曲线也反映了这一点。Transformer模型在未见测试数据上的性能则没有出现这种下降。

4.2. 自动隔离的咳嗽声

三个咳嗽活动检测模型各自采用三个阈值，这些阈值仅根据开发集确定。由此得到九个自动提取的咳嗽声数据集。选择这些阈值是为了实现：

覆盖率：覆盖率高（97%），因此牺牲了纯度。

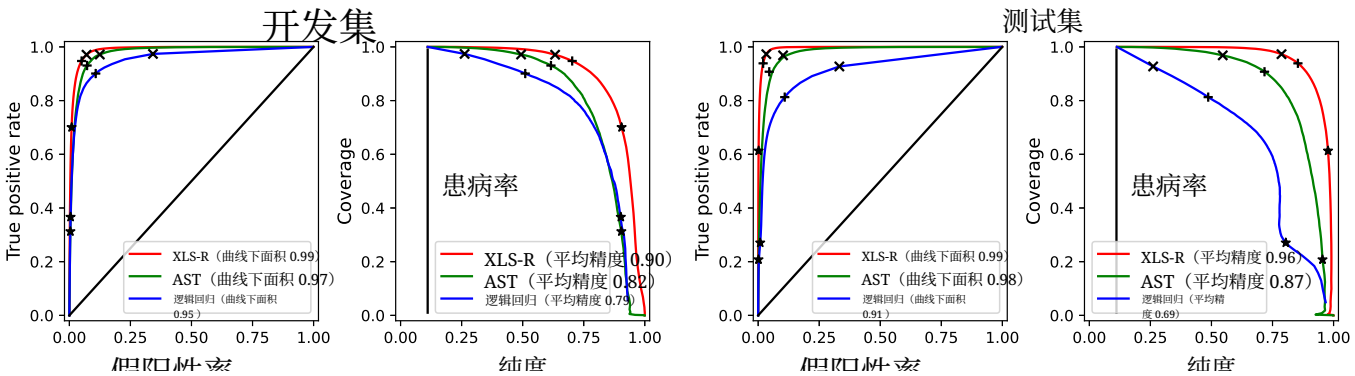


图4。三个咳嗽检测器（XLS-R、AST、逻辑回归）各自在开发集上的受试者工作特征曲线，以及覆盖率（敏感度）和纯度（精确率）曲线；并显示三个相应的工作点，×：覆盖率为 97%，⋆：纯度为 90%，以及 +：等错误率。

表3。所考虑的三个模型（逻辑回归、AST、XLS-R）的开发集和测试集平均表现（曲线下面积、平均精度）。

模型	开发集		Test	
	AUC	AP	AUC	AP
LR	0.95	0.79	0.91	0.69
AST	0.97	0.82	0.98	0.87
XLS-R	**0.99**	**0.90**	**0.99**	**0.96**

EE：ROC 曲线上的等错误率。

精确率：纯度高（90%），因此牺牲了覆盖率。

图 4 中标出了开发集和测试集对应的这些工作点。

图 5 展示了直方图，分别呈现使用三种咳嗽活动检测模型（LR、AST、XLS-R），在三个已定义工作点（C、EE、P）下分离咳嗽声后，得到的九个数据集各自的咳嗽持续时间。从图中可以清楚地看出，LR 无法在任何工作点保留咳嗽时长分布。这在针对纯度进行优化的工作点上尤为明显：提取出的咳嗽声太少，无法构成可用于结核病分类的数据集。此外，三种模型在所有三个工作点下似乎都会将咳嗽声切分成碎片。与真实值分布相比，持续 0.1 秒的咳嗽声数量异常偏高，足以说明这一点。表 4 列出了使用咳嗽检测模型所获得数据集的统计信息。

4.3. 中值滤波

对两种基于 Transformer 的咳嗽检测器计算出的逐帧分类序列应用了中值滤波器。由于逻辑回归模型未能

表 4。利用咳嗽检测模型获得的南非测试集各咳嗽数据集统计数据，以及人工标注的（真实值）统计数据。三个咳嗽检测模型分别在三个工作点（C：覆盖率，EE：等错误率，P：纯度）下用于分离咳嗽声†：应用中值滤波器后。

工作点	模型	数量 咳嗽声	平均咳嗽 时长（毫秒）	总时长。（分钟）
真实值		6886	413 ± 148	47.36
C	LR	22322	461 ± 1623	171.58
	AST	9431	561 ± 564	88.16
	XLS-R	6682	538 ± 417	59.97
C†	AST	5703	903 ± 738	85.82
	XLS-R	5144	697 ± 380	59.73
EE	LR	13585	356 ± 1905	80.61
	AST	6487	573 ± 407	61.98
	XLS-R	6401	498 ± 352	53.08
EE†	AST	4962	723 ± 405	59.83
	XLS-R	5289	602 ± 337	53.06
P	LR	3988	244 ± 198	16.19
	AST	4017	194 ± 122	13.00
	XLS-R	7544	241 ± 188	30.32

提取出合理的咳嗽起止点，因此未对其进行进一步研究。对于任何为提高纯度（精确率 P）而自动获取并优化的数据集，也未作进一步研究，因为这类模型会牺牲覆盖率。研究发现，应用中值滤波，尤其是使用较大的滤波器尺寸时，会进一步降低覆盖率，因此在这种情况下应用中值滤波适得其反。

我们研究了长度为 3、5、7 和 9 帧的滤波器。对于 AST，这对应于 300、500、700 和 900ms 的滤波器长度；而对于 XLS-R，这对应于

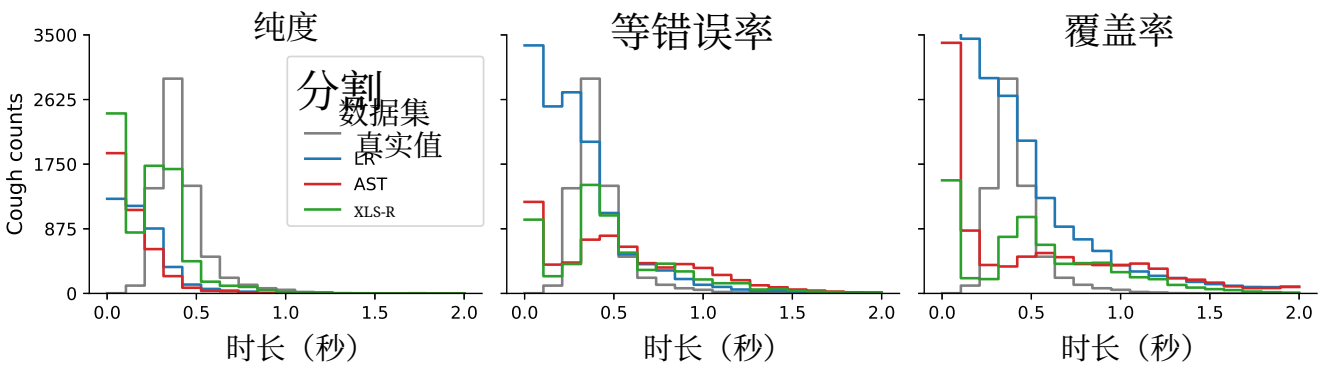


图 5。三个咳嗽活动检测模型（AST、XLS-R、逻辑回归）在三个不同工作点（纯度、覆盖率、等错误率）下的测试集咳嗽时长直方图。

60、100、140 和 180ms。我们还研究了适用于 XLS-R 的滤波器长度，对应于 300、500 和 700ms，以便与 AST 进行更直接的比较。表 5 报告了这些中值滤波器尺寸下的开发集曲线下面积和平均精确率。对于 AST，所研究的中值滤波器尺寸均未优于原始输出；而对于 XLS-R，140ms 的中值滤波器略有提升。

表 5。AST 和 XLS-R 在采用最佳超参数（表 2）并进一步应用不同尺寸的中值滤波器后的开发集平均精度。

模型	中值滤波器 size	近似值 时长（毫秒）	开发集 AUC	AP
AST	None	-	0.9732	**0.8234**
	3	300	0.9735	0.8223
	5	500	0.9716	0.8009
	7	700	0.9675	0.7635
	9	900	0.9612	0.7279
XLS-R	None	-	0.9870	0.9035
	3	60	0.9871	0.9043
	5	100	0.9872	0.9047
	7	140	0.9872	**0.9048**
	9	180	0.9872	0.9046
	15	300	0.9866	0.8986
	25	500	0.9837	0.8649
	35	700	0.9777	0.8061

4.4. 结核病分类

对于九个数据集中的一部分，我们确定了双向长短期记忆网络的下游结核病分类结果，并在表6中呈现。与前文所述相同，我们排除了使用逻辑回归提取的、以覆盖率为优化目标的数据集和以纯度为优化目标的数据集，因为这些配置未能生成可用的数据集。这一点可从以下情况看出：它们与表4和图5中的真实值分布存在较大差异。

在所有三个工作点（覆盖率、等错误率、精确率）上，使用 XLS-R 自动分离的咳嗽声训练的结核病分类器，性能均优于使用 AST 和逻辑回归自动分离的咳嗽声训练的结核病分类器。在各种咳嗽活动检测策略中，使用针对覆盖率优化的 XLS-R 所生成的数据集进行训练，可获得最佳的结核病分类性能。与使用 AST 提取的数据集训练的结核病分类模型相比，该策略在开发集曲线下面积上绝对提升了 6%，在测试集曲线下面积上绝对提升了 4%（XLS-R（覆盖率）与 AST（覆盖率）相比）。使用 XLS-R 自动分离的咳嗽声训练的结核病分类器仅略逊于使用人工标注的咳嗽声训练的分类器，在开发集和测试集曲线下面积上均低 2 个百分点。因此，将 XLS-R 应用于从连续音频中自动提取咳嗽声，似乎是实现自动咳嗽筛查的一种可行方法。

尚未发现应用中值滤波能够持续改善孤立咳嗽的质量，从而提升下游分类性能。对于 XLS-R 和 AST 咳嗽活动检测策略，下游结核病分类器的曲线下面积最多仅受到 1% 的影响。

5. 结论

已针对基于咳嗽音频的结核病筛查系统流程中的自动咳嗽活动检测任务研究了三种架构。结果表明，大型基于Transformer的网络XLS-R在测试集平均精度上分别以9%和27%的绝对优势优于AST和逻辑回归，尽管其仅在语音上进行了预训练。通过使用各咳嗽检测模型自动分离咳嗽声，并分别针对覆盖率、等错误率和纯度优化决策阈值，编制了九个数据集。当使用这些数据集进行训练

表6。使用九个自动提取的咳嗽数据集之一训练下游双向长短期记忆网络结核病分类模型时的十折嵌套交叉验证结果。C：覆盖率，EE：等错误率，P：纯度，†：应用中值滤波器之后。人工：在人工（真实值）咳嗽声上训练时的分类性能。

操作点	模型	开发 AUC	Test AUC
	手动	0.66 ± 0.01	0.65 ± 0.07
C	AST	0.58 ± 0.02	0.59 ± 0.09
	XLS-R	$\mathbf{0.64 \pm 0.02}$	$\mathbf{0.63 \pm 0.10}$
C†	AST	0.59 ± 0.02	0.57 ± 0.09
	XLS-R	0.63 ± 0.02	0.62 ± 0.09
EE	LR	0.59 ± 0.03	0.59 ± 0.10
	AST	0.60 ± 0.01	0.59 ± 0.06
	XLS-R	0.62 ± 0.01	0.61 ± 0.10
EE†	AST	0.61 ± 0.01	0.60 ± 0.05
	XLS-R	0.62 ± 0.02	0.62 ± 0.09
P	AST	0.57 ± 0.03	0.57 ± 0.12
	XLS-R	0.61 ± 0.02	0.61 ± 0.10

在下游结核病分类模型中，研究发现，利用XLS-R分离出的咳嗽声始终取得最佳测试集曲线下面积，测试集曲线下面积绝对值比AST高出4%，仅略逊于采用人工标注的咳嗽声训练的分类器（绝对值低2%）。

6. 参考文献

[1] Praveer Sharan，“音频录音中的咳嗽自动判别：一项范围综述”，*Frontiers inSignalProcessing*，第2卷，2022年。

[2] Antoine Serrurier、Christiane Neuschaefer-Rube和Rainer R¨ohrig，“咳嗽声采集、自动检测与自动分类的历史与趋势：一项比较综述”，*Sensors*，第22卷，第8期，2022年。

[3]Siqi Liao、Chao Song、Xiaoqin Wang和Yanyun Wang，“基于小规模咳嗽声音数据集识别儿童支气管炎与肺炎的分类框架”，*PLOS ONE*，第17卷，第10期，第1–17页，2022年。

[4] Debarpan Bhattacharya、Neeraj Kumar Sharma、Debot- tam Dutta、Srikanth Raj Chetupalli、Pravin Mote、Sriram Ganapathy、C. Chandrakiran、Sahiti Nori、K. K. Suhail、Sadhana Gonuguntla 和 Murali Alagesan，《Coswara：用于 SARS-CoV-2感染远程筛查的呼吸音与症状数据集》，*Scientific Data*，第 10 卷，第 1 期，2023 年。

[5] LaraOrlandic、Tomas Teijeiro 和 David Atienza，《COUGHVID 众包数据集：用于研究大规模咳嗽分析算法的语料库》，*Scientific Data*，第 8 卷，第 1 期，2021 年。

[6] D Trejo Pizzo和Santiago Esteban，“IATos：基于咳嗽音频样本的COVID-19人工智能预筛查工具”，*arXiv*预印本*arXiv:2104.13247*，2021年。

[7] Manuja Sharma、Videlis Nduba、Lilian N. Njagi、Wil- fred Murithi、Zipporah Mwongera、Thomas R. Hawn、Shwetak N. Patel 和 David J. Horne，《TBscreen：使用受控数据集进行结核病筛查的被动式咳嗽分类器》，*Science Advances*，第 10 卷，第 1 期，页码 eadi0282，2024 年。

[8] Sophie Huddart、Vijay Yadav、Solveig K. Sieberts、Lar- son Omberg、Mihaja Raberahona、Rivo Rakotoarivelo、Issa N. Lyimo、Omar Lweno、Devasahayam J. Christo- pher、Nguyen Viet Nhung、Grant Theron、William Worodria、Charles Y. Yu、Christine M. Bachman、Stephen Burkot、Puneet Dewan、Sourabh Kulhare、Pe- ter M. Small、Adithya Cattamanchi、Devan Jaganath 和 Simon Grandjean Lapierre，《用于结核病分诊检测的主动采集咳嗽声音数据集》，*ScientificData*，第 11 卷，第 1 期，页码 1149，2024 年。

[9] Anthony Windmon、Mona Minakshi、Sriram Chellappan、Ponrathi R. Athilingam、Marcia Johansson和Bradlee A. Jenkins，“利用智能手机录制的音频信号检测慢性阻塞性肺疾病（COPD）咳嗽”，载于第十一届生物医学工程系统与技术国际联合会议论文集（*BIOSTEC*），葡萄牙马德拉丰沙尔，2018年。

[10] Paul Porter、Scott Claxton、Javan Wood、Vesa Peltonen、Joanna Brisbane、Fiona Purdie、Claire Smith、Natasha Bear和Udantha Abeyratne，“使用基于智能手机的、以咳嗽为中心的算法诊断慢性阻塞性肺疾病（COPD）急性加重”，《欧洲呼吸杂志》，第54卷，增刊63，2019年。

[11]Yusuf Amrulloh、Udantha Abeyratne、Vinayak Swarnkar和Rina Triasih，“面向儿科人群的肺炎与哮喘分类的咳嗽声分析”，载于第六届智能系统、建模与仿真国际会议论文集，马来西亚吉隆坡，2015年。

[12]Jes´us Monge-Alvarez、Carlos Hoyos-Barcel´o、Luis Miguel San-Jos´e-Revuelta 和 Pablo Casaseca-de-la Higuera，《一种基于特定频带音频特征高层表示的稳健咳嗽检测机器听觉系统》，*IEEE*汇刊

生物医学工程，第66卷，第8期，第2319–2330页，2019年。

[13] Alice E Ashby、Julia A Meister、Goran Soldar和Khuong An Nguyen，“一种用于COVID-19检测的新型咳嗽音频分割框架”，载于《后疫情时代开放数据与知识研讨会（*ODAK22*）论文集》，英国布莱顿，2022年。

[14] Justice Amoh 和 Kofi Odame，《用于识别咳嗽声的深度神经网络》，*IEEE*生物医学电路与系统汇刊，第10卷，第5期，第1003–1011页，2016年。[15] Mirco Ravanelli 和 Yoshua Bengio，《使用 sincnet 从原始波形进行说话人识别》，载于*IEEE*口语语言技术研讨会（*SLT*）论文集，希腊雅典，2018年。[16]Roneel V. Sharan，《使用 sincnet 和双向GRU从原始波形检测咳嗽声》，生物医学信号处理与控制，第82卷，第104580页，2023年。[17] Arun Babu、Changhan Wang、Andros Tjandra、Kushal Lakhotia、Qiantong Xu、Naman Goyal、Kritika Singh、Patrick von Platen、Yatharth Saraf、Juan Pino、Alexei Baevski、Alexis Conneau 和 Michael Auli，《XLS-R：大规模自监督跨语言语音表征学习》，载于*Interspeech*论文集，韩国仁川，2022年。[18] Yuan Gong、Yu-An Chung 和 James Glass，《AST：音频频谱图变换器》，载于*Interspeech*论文集，捷克共和国布尔诺，2021年。[19] Jort F. Gemmeke、Daniel P. W. Ellis、Dylan Freedman、Aren Jansen、Wade Lawrence、R. Channing Moore、Manoj Plakal 和 Marvin Ritter，《Audio set：用于音频事件的本体与人工标注数据集》，载于*IEEE*声学、语音与信号处理国际会议（*ICASSP*）论文集，美国路易斯安那州新奥尔良，2017年。[20] A. Waibel、T. Hanazawa、G. Hinton、K. Shikano 和 K.J. Lang，《使用时延神经网络进行音素识别》，*IEEE*声学、语音与信号处理汇刊，第37卷，第3期，第328–339页，1989年。[21] Peter Wittenburg、Hennie Brugman、Albert Russel、Alex Klassmann 和 Han Sloetjes，《ELAN：多模态研究的专业框架》，载于第五届语言资源与评估国际会议（*LREC’06*）论文集，意大利热那亚，2006年。

[22] Ilya Loshchilov 和 Frank Hutter，《解耦权重衰减正则化》，载于国际学习表征会议（*ICLR*）论文集，美国路易斯安那州新奥尔良，2019年。[23] Dan Hendrycks 和 Kevin Gimpel，《高斯误差线性单元（GELU）》，*arXiv*预印本 *arXiv:1606.08415*，2016年。[24]Geoffrey T. Frost、Grant Theron 和 Thomas Niesler，《结核病还是不结核病？用于结核病分类的咳嗽声学分析》，载于*Interspeech*论文集，韩国仁川，2022年。