用于认知障碍检测的自监督声学表征中的偏差与公平性

Kashaf Gulzar，*IEEE*会员，Korbinian Riedhammer，*IEEE*高级会员，Elmar N¨oth，*IEEE*会员，Andreas K. Maier，Paula Andrea P´erez-Toro，*IEEE*会员

arXiv:2603.02937v1 [eess.AS] 3 Mar 2026

摘要—基于语音的认知障碍（CI）检测为早期诊断提供了一种前景良好的无创方法，但不同人口统计与临床亚组之间的性能差异仍未得到充分研究，引发了对公平性和泛化性的担忧。本研究使用 DementiaBank Pitt语料库，对基于声学特征的认知障碍和抑郁症分类进行系统的偏差分析。我们比较了传统声学特征（梅尔频率倒谱系数、eGeMAPS）与来自 Wav2Vec 2.0（W2V2）的上下文化语音嵌入，并评估了不同性别、年龄和抑郁状态亚组的分类性能。在认知障碍检测中，高层 W2V2嵌入优于基线特征（非加权平均召回率最高达 80.6%），但存在性能差异；具体而言，女性和较年轻参与者的判别能力较低（AUC：分别为 0.769 和 0.746），且特异性差异显著（Δ_{spec}最高分别达 18% 和 15%），因此相较于其他群体，他们面临更高的误分类风险。这些差异反映了表征偏差，即模型在不同人口统计或临床亚组间性能的系统性差异。对认知障碍患者进行抑郁症检测时，整体性能较低，而 W2V2 的低层和中层带来了小幅提升。认知障碍与抑郁症分类之间的跨任务泛化能力有限，表明每项任务都依赖于不同的表征。这些发现凸显了在临床语音应用中进行公平性感知模型评估和亚组特异性分析的必要性，尤其是考虑到真实应用场景中人口统计与临床异质性的存在。

索引词—**阿尔茨海默病、认知障碍、抑郁症、声学嵌入、偏差分析**

I. 引言

认知衰退（CD）是指认知能力逐渐且通常不可逆地退化，包括记忆、推理、注意力和语言能力，通常与衰老有关。然而，多种神经系统和心理障碍，尤其是阿尔茨海默病（AD）和抑郁症，是认知衰退的主要原因，在个人和社会层面都带来了重大挑战 [1]。在这一范围内，认知障碍（CI）是一个更广泛的临床类别，用于描述超出典型年龄相关变化的、可测量的能力缺损。轻度认知障碍（MCI）是认知障碍的一种特定早期阶段，表现为明显的认知困难，但尚未严重影响日常功能 [2]。MCI具有临床意义，因为它是正常衰老与痴呆（尤其是阿尔茨海默病）之间的过渡阶段，受影响者面临较高的进展风险 [1]。

阿尔茨海默病是一种进行性神经退行性疾病，主要表现为记忆丧失、推理能力受损和行为改变。早期检测，尤其是在轻度认知障碍阶段，对于及时干预和改善患者预后至关重要。然而，由于依赖临床生物标志物且早期症状不明显，早期检测仍然困难 [2]。此外，30–50% 的阿尔茨海默病患者会出现抑郁症状，尤其是在认知障碍早期和中期阶段 [3], [4]。睡眠紊乱、冷漠和注意力不集中等症状重叠，使鉴别诊断颇具挑战 [5]。由于抑郁症仍可治疗，且两者共病会增加进展为阿尔茨海默病的风险 [6], ，因此准确鉴别在临床上至关重要。

A. 相关工作

近年来，人们对利用机器学习（ML）技术和声学分析来检测认知障碍及其相关疾病（如抑郁症）表现出日益浓厚的兴趣 [7]。现有研究大致可分为三个方法方向：（1）基于人工设计声学特征的阿尔茨海默病检测方法，（2）针对认知障碍人群的基于声学的抑郁症分类，以及（3）基于嵌入表示和迁移学习的模型。

大量研究探讨了用于阿尔茨海默病检测的传统声学特征。[8] 的作者使用仅通过自发语音识别阿尔茨海默病痴呆（ADReSSo）挑战赛数据集，并将逻辑回归分类器应用于多种声学和语言特征。在各种声学配置中，扩展日内瓦简约声学参数集（eGeMAPS）特征集取得了最高准确率，为 74%。类似地，Moro-Vel´ázquez等人 [9] 在概率线性判别分析框架中结合了梅尔频率倒谱系数（MFCCs）、静音相关特征和神经嵌入，仅使用声学特征便取得了 73% 的准确率。在另一项研究中，Balagopalan 和 Novikova [10] 采用了梅尔频率倒谱系数、基频微扰、微扰和过零率，并结合支持向量机（SVMs）、神经网络和决策树等多种机器学习模型。他们的研究结果表明，使用支持向量机时，声学特征的准确率为 66%。

还有研究在认知障碍人群中探索了基于声学特征的抑郁症分类。Abdallah-Qasaimeh 和 Ratt´e [11] 评估了频谱、韵律以及基于梅尔频率倒谱系数的

K. Gulzar和K. Riedhammer就职于人工智能中心（KIZ）Zentrum f¨ur K¨unstliche Intelligenz，纽伦堡应用技术大学Technische Hochschule N¨urnberg，德国N¨urnberg（电子邮箱：kashaf.gulzar@th-nuernberg.de；korbinian.riedhammer@th-nuernberg.de）。E. N¨oth、A. K. Maier和P. A. P´erez-Toro就职于模式识别实验室，埃朗根-纽伦堡弗里德里希-亚历山大大学 Friedrich-Alexander Universit¨at Erlangen-N¨urnberg，德国埃朗根（电子邮箱：elmar.noeth@fau.de；andreas.maier@fau.de；paula.andrea.perez@fau.de）。P. A. P´erez-Toro就职于GITA实验室，安蒂奥基亚大学，哥伦比亚Medell´ın。

统计特征，这些特征来自 DementiaBank 的 Pitt语料库，并使用支持向量机和随机森林分类器。他们的最佳表现是随机森林模型达到 91% 的准确率。在此基础上，Abdallah 等人 [12] 将声学特征划分为递增的组，并应用多层感知机（MLP），最高准确率达到 77%。

近年来，基于嵌入和迁移学习的方法逐渐兴起。Perez-Toro 等人 [13] 提出了一种迁移学习框架，结合了 DB Pitt 语料库和交互式情感双人动作捕捉（IEMOCAP）语音数据集中的韵律特征与频谱特征。采用 ForestNet 模型后，其系统在阿尔茨海默病检测和抑郁症分类任务中的非加权平均召回率（UAR）分数分别达到 87% 和 82%。在一项相关研究中，Braun等人 [14] 利用 Wav2Vec 2.0（W2V2）嵌入和支持向量机，对两个独立的德语语音数据集进行了跨语料库分析。尽管取得了中等水平的性能（UAR=62%），他们的研究凸显了基于语音的痴呆检测在跨数据集泛化能力方面面临的挑战。

除了以性能为导向的研究，自监督学习（SSL）模型的公平性和人口统计偏差问题也日益受到关注。研究表明，Wav2Vec 2.0（W2V2）和 Whisper 等模型存在人口统计性能差异，尤其是在不同性别和年龄组之间 [15], [16], [17]。此外，Hu 等人 [18] 通过分析老年人和构音障碍语音以及多种语言中的语言表征，研究了 SSL 模型在自动语音识别（ASR）性能和下游二元阿尔茨海默病检测任务中的偏差与鲁棒性。尽管取得了令人瞩目的进展，目前尚无研究探讨人口统计因素（如年龄、性别）和临床因素对用于认知障碍检测的基于声学的机器学习模型的性能和公平性的潜在影响。这一研究空白限制了现有系统的临床可靠性和公平性，也凸显了该领域开展偏差感知评估的必要性。

B. 本研究的贡献

在本研究中，我们系统地考察人口统计学和临床因素对基于声学的机器学习模型在认知障碍检测中的性能的影响。我们使用 DementiaBank Pitt 语料库，评估传统基线特征（梅尔频率倒谱系数、eGeMAPS）以及 Wav2Vec 2.0 嵌入在认知障碍和抑郁症分类任务中的表现。我们的主要贡献如下：

- 我们对比评估了基线特征和 W2V2 嵌入在认知障碍及共病抑郁症分类中的表现。
- 我们系统地分析了人口统计学和临床因素对 W2V2 性能的影响，揭示了影响公平性和临床可靠性的偏差。
- 我们考察了认知障碍与抑郁症检测之间的表征重叠及跨任务泛化能力。

据我们所知，这是首次对基于语音的认知障碍检测中的人口学与临床偏差开展全面研究。除了分类性能之外，我们的工作还受到人们日益关注自监督模型在临床语音应用中的公平性和可靠性的推动。基于语音的人工智能工具日益被提议用于早期检测认知障碍，但模型在不同人口学或临床亚组中的性能不均，可能会加剧现有的健康差异。通过量化这些亚组差异，并指出即使经过数据平衡，表征偏差仍持续存在，我们旨在推动制定公平性感知评估协议，以供未来临床部署基于语音的机器学习系统。

II. 数据

该数据集包含来自 DementiaBank Pitt语料库子集的 229 名参与者的半自发语音录音 [19]，，其中包括 139 名伴有认知障碍的阿尔茨海默病患者和 90 名认知正常（无认知障碍）对照者。参与者完成了“CookieTheft”图片描述任务 [20]。根据提供的时间戳，预处理时移除了访谈者语音。

根据早发型和晚发型阿尔茨海默病的标准临床界值 [21]，参与者分为两个年龄组：0–65 岁（组1）和 66 岁及以上（组2）。按照既定临床指南，分别依据简易精神状态检查（MMSE）和汉密尔顿抑郁量表（HAM-D）分数判定认知障碍和抑郁症的临床标签。MMSE 分数范围为 0 至 30，低于 24 分表示存在认知障碍 [22]，；HAM-D 分数范围为 0 至 22，达到或超过 8 分则表示存在抑郁症 [23], [24]。在阿尔茨海默病患者中，非抑郁组（$n = 88$）的 HAM-D 均值为 4.41，而抑郁组（$n = 51$）的均值为 11.20。

该数据集存在类别不平衡。为缓解这一问题，我们在预处理过程中先按认知障碍状态、再按性别对数据进行平衡（见图 1）。我们从每个组中随机抽取说话者，并在额外实验中使用剩余样本作为替代测试集，以验证鲁棒性。

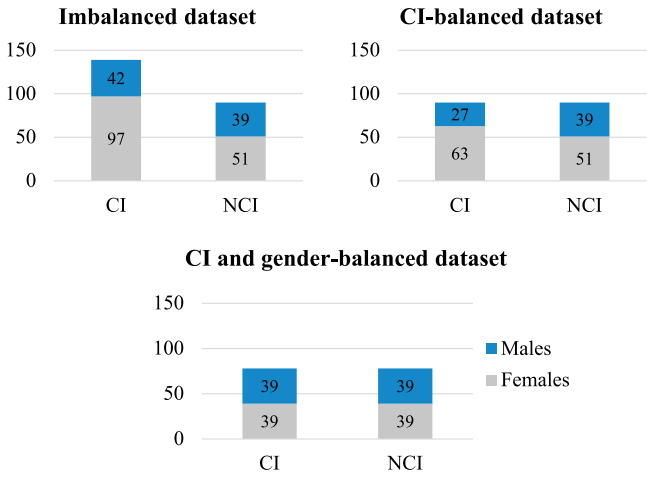


图 1. 基于认知障碍状态和性别的受试者分布。

按年龄组进行平衡并不可行，原因在于阿尔茨海默病在老年人（≥ 65 岁）中占主导地位 [4] 以及有限的

表I 不平衡、CI平衡和CI性别平衡数据集受试者的人口统计学和临床信息。CI：CI患者，NCI：NCI受试者，年龄组：年龄组，F：女性，M：男性。简易精神状态检查和汉密尔顿抑郁量表分数以均值（标准差）表示。

	不平衡		CI平衡		CI性别平衡	
	CI [女性/男性]	NCI [女性/男性]	CI [F/M]	NCI [F/M]	CI [F/M]	NCI [F/M]
性别	97/42	51/39	63/27	51/39	39/39	39/39
年龄组1	22/13	29/25	19/10	29/25	14/12	21/25
年龄组2	75/29	22/14	44/17	22/14	25/27	18/14
认知障碍-低	0/0	51/39	0/0	51/39	0/0	39/39
认知障碍-轻度	75/28	0/0	56/18	0/0	34/25	0/0
严重认知障碍	22/14	0/0	7/9	0/0	5/14	0/0
抑郁	33/18	3/7	24/11	3/7	13/16	3/7
MMSE	18.29 (4.12)/ 17.37 (4.29)	29.29 (0.99)/ 28.85 (1.05)	19.40 (3.73)/ 17.52 (4.53)	29.29 (0.99)/ 28.85 (1.05)	19.15 (4.05)/ 17.13 (4.31)	29.44 (0.84)/ 28.85 (1.05)
汉密尔顿抑郁量表	6.69 (3.04)/ 7.30 (4.69)	2.20 (3.11)/ 3.79 (4.26)	6.89 (3.91)/ 7.48 (5.24)	2.20 (3.11)/ 3.79 (4.26)	6.31 (3.71)/ 7.15 (4.63)	2.33 (3.35)/ 3.79 (4.26)

较年轻阿尔茨海默病参与者数量，这反映了数据集的固有特征（见表I）。

表I详细列出了原始不平衡数据、CI平衡数据和CI性别平衡子集的分布。MMSE分数高于24的参与者被归类为认知正常。由于无认知障碍组中抑郁个体数量较少，与抑郁症相关的分析仅限于CI参与者。

为尽量减少与数据采集相关的混杂因素的影响，所有音频录音均经过降噪和归一化预处理。这些步骤在所有参与者中一致实施，以确保录音条件的差异不会系统性地影响模型在不同人口统计学或临床亚组中的表现。

III. 方法

图2展示了整体工作流程。从不平衡、CI平衡和CI性别平衡数据集中提取语音录音的声学表征，包括梅尔频率倒谱系数、eGeMAPS和W2V2。

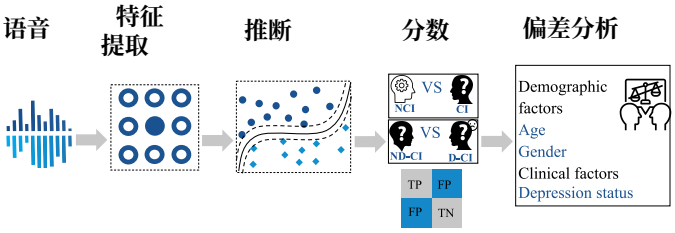


图2. 展示数据分析关键步骤的工作流程图。

针对每个特征集开展了两项分类任务：(1) 认知障碍 *vs.* 无认知障碍，以及 (2) 抑郁认知障碍 *(D-CI) vs.* 非抑郁认知障碍 *(ND-CI)*。由于数据集中抑郁的无认知障碍受试者数量有限，第二项分类任务未纳入这类受试者（参见表I）。为考察潜在的人口学与临床偏差，我们对每个预测数据类别进行了偏差分析，考察年龄、性别和抑郁状态（基于汉密尔顿抑郁量表）等因素。偏差分析未考虑认知障碍严重程度，因为轻度认知障碍受试者仅出现在 NCI 组，而轻度/重度病例仅出现在 CI组。严重程度的受试者仅出现在 NCI组，而轻度/重度病例仅出现在 CI组。

所有数据最初均由DementiaBank PittCorpus团队收集，并获得参与者的知情同意及机构审查委员会（IRB）的批准。本研究仅使用现有的去标识化数据集，未收集任何新数据，也未直接与参与者互动。

A. 声学表征

我们采用了三种复杂度各异的声学表征：基础频谱特征（梅尔频率倒谱系数）、进一步衍生的参数集（eGeMAPS），以及基于话语的神经嵌入（W2V2）。梅尔频率倒谱系数和eGeMAPS用作基线，而W2V2嵌入则用于详细评估和偏差分析。所有特征均独立归一化，以确保尺度一致。这些表征共同涵盖了传统方法和最先进的方法，用于建模与认知障碍及抑郁症分类相关的声学模式。

1) 梅尔频率倒谱系数：梅尔频率倒谱系数是广泛使用的语音特征，其灵感来源于人类听觉系统对声音的非线性感知。它们将频率映射到梅尔刻度上，使其更贴合人类对音高的感知 [25]。计算梅尔频率倒谱系数时，先对语音信号的短时窗应用快速傅里叶变换 (FFT)，将得到的频谱转换为梅尔频谱图并取其对数，然后应用离散余弦变换 (DCT)。这些系数以对说话者差异和录音差异具有鲁棒性而著称 [26]。

在本研究中，使用 2048 个采样点的窗口、512 个采样点的步长，以 16 kHz 采样率提取了 40 维梅尔频率倒谱系数 [27]。为捕捉说话人特定信息，针对每个系数沿时间维度对梅尔频率倒谱系数求均值，得到一维表示，同时保留全部 40 个梅尔频率倒谱系数特征 [28]。尽管可以加入动态系数（Δ 和 ΔΔ）及其他统计量，但内部交叉验证表明，均值聚合的梅尔频率倒谱系数为我们的分类任务提供了最有效且稳定的配置，这与此前认知障碍语音分析领域的研究一致 [10]。

2）扩展日内瓦简约声学参数集：eGeMAPS 是一套用于语音分析的标准化声学特征集，旨在捕捉与情感及临床应用相关的嗓音特征 [29]。该集合包含 88 个参数，在 18 个低层描述符（LLDs）的简约集合基础上，进一步加入了倒谱特征和动态特征 [30]。

- **频率相关参数：**包括音高、基频微扰，以及共振峰1、2和3的中心频率与带宽。
- **振幅相关参数：**包括微扰、响度和谐噪比。
- **频谱（平衡）参数：**包括α比率、Hammarberg指数、谱斜率（0–500 Hz、500–1500 Hz）、梅尔频率倒谱系数1–4、谱通量，以及谐波差（H1–H2、H1-H3）。

本研究采用了最新的“eGeMAPSv02”配置，并使用“统计泛函”特征层级，涵盖能够捕捉整体语音模式、反映认知衰退的特征 [29]。

3) Wav2Vec 2.0：W2V2是一种用于语音的自监督神经嵌入模型，它结合卷积前端层与Transformer模块，从原始音频中学习上下文语音表征 [31]。该模型在每个Transformer层每20毫秒输出一个512维嵌入，可通过聚合（例如取平均值）用于下游任务。W2V2采用对比学习框架进行预训练，根据周围上下文预测大型无标注语音数据集中的掩蔽音频片段。此过程会最小化对比损失，促使相邻语音片段获得相似的表征 [32]。其分层架构能够提取低层和高层声学模式。

鉴于其在基于语音的分类任务中的出色表现，我们采用了 Wav2Vec2-Base-960h 模型。该模型在 960 小时、16kHz 的 LibriSpeech 数据上进行了预训练和微调 [33]。分别对来自潜在层和隐藏层的嵌入进行分类评估。在此背景下，潜在层对应于卷积特征编码器的输出，该编码器将原始音频压缩为声学表征 [31],；而隐藏层则指能够捕获上下文信息和更高层次语音特征的 Transformer编码器状态 [34]。

B. 优化与分类

分类使用了来自不同算法类别的三种常用分类器，以全面覆盖偏差分析场景：(1)径向基函数支持向量机 *(RBF-SVM)* [10],(2)随机森林 [11],以及 (3)多层感知机 [12]。

对于 RBF-SVM，超参数通过分层抽样5折交叉验证网格搜索确定，其中$C \in \{10^{-2}, 10^{-1}, \ldots, 10^{3}\}$ 和 $\gamma \in \{10^{0}, 10^{-1}, \ldots, 10^{-4}\}$。针对每种实验设置，均使用该具体任务所采用的数据集单独进行调参。在模型选择过程中，以准确率作为主要优化指标。

随机森林分类器由 100 棵决策树组成，每棵树都会生长至最大深度，除非某个节点包含的样本少于两个。每个节点用于划分的特征数量设置为输入特征总数的平方根。为确保结果可复现，所有分类实验均使用固定的随机种子（42）。随机森林的集成特性通常能够缓解过拟合，因此与支持向量机等模型相比，对特定超参数调整的敏感性较低。因此，随机森林的配置参考了阿尔茨海默病分类研究中的常见做法 [35], , 因为本研究的主要重点并非优化随机森林的超参数，而是考察声学表征中的偏差。

多层感知机包含三个隐藏层，神经元数量分别为150、100和50，这种设置旨在降低复杂度并提升泛化能力。隐藏层采用整流线性单元（ReLU）激活函数，并使用学习率为0.001的Adam优化器进行优化。训练轮次设为1,000，以实现最佳分类性能，并采用随机种子42以保证可复现性。多层感知机模型配置基于 [36] 中用于阿尔茨海默病分类的架构选定，因为本研究的主要重点并非优化多层感知机架构，而是比较三种不同算法类别的性能。

对于所有分类器，数据被划分为70%的训练集和30%的测试集。所有划分均根据分类标签进行分层抽样：在认知障碍与无认知障碍任务中依据认知障碍状态，在D-CI与ND-CI任务中依据抑郁状态。最终评估指标通过对五个固定随机数据划分（种子分别为0、50、100、150和200）的结果取平均得到。

为支持研究结果的可复现性，我们已在 GitHub 代码库中公开源代码，以及对应各实验条件（不平衡、CI平衡和CI性别平衡）的音频文件名列表：https://github.com/th-nuernberg/taslp-bias-in-dementia。由于数据共享限制，我们无法分发 DB Pitt语料库中的原始音频录音和转录文本。有意使用该数据集的研究人员可直接通过 DementiaBank 申请访问权限 [1]。

C. 公平性分析

为评估潜在的模型相关偏差，我们针对关键的人口统计和临床因素（包括年龄组、性别和抑郁状态）进行了亚组性能评估。对于每个亚组，我们计算了亚组特异性敏感度和亚组特异性特异度，以便在不受全局数据集不平衡影响的情况下进行公平比较。

$$\text{Subgroup-Specific Sensitivity} = \frac{\text{Correct CI predictions in subgroup}}{\text{CI samples in that subgroup}}$$

$$\text{Subgroup-Specific Specificity} = \frac{\text{Correct NCI predictions in subgroup}}{\text{NCI samples in that subgroup}}$$

对于给定亚组g，组内性能不平衡(δ)量化为亚组特异性敏感度与亚组特异性特异度之差：

$$\delta_g = \text{Subgroup-Specific Specificity}_g - \text{Subgroup-Specific Sensitivity}_g$$

[1]https://dementia.talkbank.org

表 II 不同声学特征类型、数据集配置和分类器下，认知障碍与无认知障碍分类的最佳结果。结果以准确率、非加权平均召回率、灵敏度和特异度表示，取 5 次固定随机数据划分的平均值，并以均值（标准差）% 表示。梅尔频率倒谱系数和 eGeMAPS 用作基线特征。每种数据集配置下最高的非加权平均召回率以高亮显示。星号 (*) 表示配对 t 检验得出$p<0.05$ 的比较。分类器：分类器；IMB：不平衡；CI-B：认知障碍平衡；CIG-B：认知障碍与性别平衡；平衡数据训练／剩余数据测试：使用平衡数据训练，并在剩余数据上测试；W2V-HL：Wav2Vec 2.0 隐藏层。

数据集	特征	Clf	准确率	UAR	灵敏度	特异度
IMB	梅尔频率倒谱系数	MLP	66.38 (2.81)	63.39 (3.36)	77.14 (4.42)	49.63 (7.98)
CI-B		MLP	63.33 (4.74)	63.33 (4.74)	65.18 (5.54)	61.48 (6.02)
CIG-B		MLP	66.81 (3.95)	66.75 (4.02)	64.35 (10.79)	69.17 (7.73)
IMB	eGeMAPS	SVM	59.42 (2.25)	49.73 (0.96)	94.29 (7.16)	5.19 (5.54)
CI-B		SVM	50.74 (5.31)	50.74 (5.31)	57.04 (6.87)	44.45 (13.25)
CIG-B		SVM	46.81 (6.31)	47.25 (6.31)	68.69 (19.90)	25.83 (19.61)
IMB	W2V2 HL 9	MLP	81.16 (3.55)	**80.56 (5.02)***	83.33 (3.98)	77.78 (12.61)
CI-B		MLP	80.00 (4.29)	**80.00 (4.29)***	77.04 (7.18)	82.96 (4.45)
CIG-B		SVM	76.60 (3.01)	76.54 (2.96)*	73.91 (4.76)	79.17 (6.97)
IMB	W2V2 HL 10	SVM	80.00 (3.35)	79.47 (2.99)*	81.90 (7.31)	77.04 (7.18)
CI-B		SVM	76.29 (1.81)	76.29 (1.81)*	74.81 (7.18)	77.78 (6.19)
CIG-B		MLP	76.17 (4.54)	76.18 (4.62)*	76.52 (8.95)	75.83 (3.11)
平衡数据训练／剩余数据测试						
CIG-B	W2V2隐藏层7	MLP	76.83 (4.81)	**78.53 (2.76)**	74.29 (10.77)	82.78 (11.44)

其中，正的 δ_g 表示模型较为保守，倾向于判定为NCI病例；负值则表明模型较为激进，倾向于判定为同一亚组中的认知障碍病例。

组间性能差异（Δ）计算为各亚组之间性能指标的差值，用于衡量表征偏差：

$$\Delta_{sens} = \text{Subgroup-Specific Sensitivity}_{groupA} - \text{Subgroup-Specific Sensitivity}_{groupB}$$

$$\Delta_{spec} = \text{Subgroup-Specific Specificity}_{groupA} - \text{Subgroup-Specific Specificity}_{groupB}$$

同时报告 δ 和 Δ，可确保分析既反映各亚组内部错误的性质，也反映亚组之间性能的公平性。通过对五个固定交叉验证折进行配对t检验（$p<0.05$），评估这些差异的统计显著性。

此外，为了找出性能差异的来源，我们通过叠加的密度直方图可视化模型在各亚组中的得分分布，其中类别之间的重叠区域代表模型不确定性最大的区域。类别重叠明显增加，表明存在算法偏差，也说明提取的特征对该特定亚组的代表性较差。为了独立于决策阈值量化这种区分能力，我们计算了亚组特异性的 ROC 曲线下面积（AUC），该指标衡量在某个人口统计群体中，随机选取的 CI 样本得分高于随机选取的 NCI 样本的概率：

$$AUC_g = P(Score(CI_g) > Score(NCI_g))$$

各组之间 AUC 存在显著差距，这证实了这种差异并非仅仅是校准或阈值设定造成的假象，而是模型针对这些人口群体表征和分类数据的能力存在根本差异。

IV. 实验与结果

本研究进行了两项不同的实验：(1)评估认知障碍（*CI*）与无认知障碍（*NCI*）以及 *D-CI* 与 *ND-CI* 分类中的特征性能，以及 (2) 针对不同人口统计和临床因素，对表现最佳的特征—分类器组合进行偏差分析。

A. 特征性能

在本实验中，我们评估了三种声学表征的分类性能。我们将梅尔频率倒谱系数和 eGeMAPS 作为基线声学表征，在全部三种数据集配置上进行评估，即不平衡、CI平衡以及 CI 性别平衡数据集。虽然我们考察了 W2V2 模型的潜在层嵌入和隐藏层嵌入（编号为 1 至 12），但报告中仅列出性能最高的隐藏层，因为潜在层嵌入的分类性能始终较低。

我们采用准确率、非加权平均召回率、灵敏度和特异度作为评估指标。灵敏度反映正确识别出的 CI（或抑郁）病例比例，而特异度反映正确识别出的 NCI（或非抑郁）病例比例 [37]。为评估性能差异的可靠性，我们对五个固定的分层交叉验证折所得的非加权平均召回率进行了配对 t 检验。当差异达到$p<0.05$的阈值时，即认为具有统计学显著性。

1) CI 与无认知障碍：表II列出了每种声学特征类型、数据集配置和分类器所取得的最佳分类结果。在基线特征中，梅尔频率倒谱系数表现最佳；使用多层感知机分类器时，其在CI平衡数据集和CI性别平衡数据集上的非加权平均召回率分别达到63.3%和66.75%。两个数据集上的灵敏度和特异度相对均衡：在CI平衡数据集中分别为65.2%和61.5%，在CI性别平衡数据集中分别为64.3%和69.1%。在不平衡数据集上，尽管非加权平均召回率相近，灵敏度却有所提高（77.1%），代价是特异度降低（49.6%），

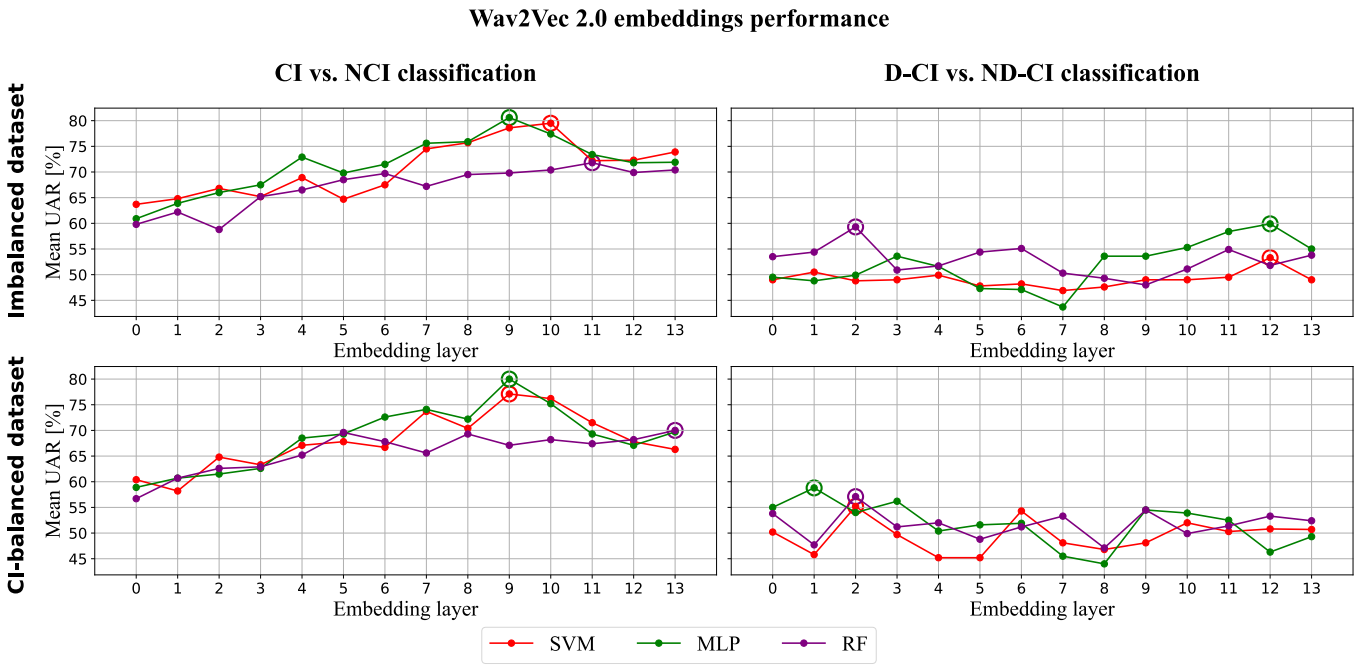


图3. 不平衡数据集和CI平衡数据集上，CI与NCI分类以及D-CI与ND-CI分类中各W2V2嵌入层的分类器性能比较。每种分类器的最佳结果均以突出显示。

这表明模型倾向于将样本过度分类为CI。此现象符合预期，因为不平衡设置使模型接触到的CI样本显著增多，促使模型优先检测CI，同时低估无认知障碍模式；而平衡数据则恢复了公平的类别表示，并相应改变了错误分布。相比之下，eGeMAPS的判别能力有限；使用支持向量机分类器时，其最高非加权平均召回率仅为50.7%，且出现在CI平衡数据集上。在不平衡数据集上，其表现尤其不佳，分类结果受到极低特异度（5.2%）的影响，表明模型难以正确识别NCI病例。

相比之下，在所有数据集配置中，W2V2第9层和第10层的表现均显著优于传统声学特征（所有$p<0.05$）。W2V2第9层结合多层感知机在不平衡数据集上取得了最高的非加权平均召回率（80.6%），灵敏度（83.3%）和特异度（77.8%）也较高。在每种数据集配置中，性能均相对稳定：在CI平衡数据集上的非加权平均召回率为=80.0%，在CI性别平衡数据集上的非加权平均召回率为=76.5%。类似地，对于W2V2第10层，使用支持向量机时，不平衡数据集上的准确率最高，为80.0%（非加权平均召回率为=79.5%），且在其他数据集配置中性能也保持稳定。

此外，在模型于CI和性别平衡数据集上训练并在剩余数据上评估的实验设置中，W2V2第7层在多层感知机下的表现优于所有其他已分析的声学表征（非加权平均召回率=78.5%），表明其在平衡训练条件之外具有较强的泛化能力。

图3展示了在不平衡和CI平衡数据集上，W2V2嵌入层在认知障碍与无认知障碍分类任务中的性能对比。

对于这两种数据集，嵌入层9和10的表现均优于其他所有层。此外，在嵌入层6及之前，各分类器的表现相近；此后，支持向量机和多层感知机的非加权平均召回率有所提升，但随机森林分类器没有。

2) D-CI与ND-CI：表III汇总了所有声学特征及数据集配置下抑郁症分类的最佳结果。梅尔频率倒谱系数和eGeMAPS在CI平衡数据集上均表现出较弱的分类性能，非加权平均召回率约为50%，灵敏度低至6%，表明在认知障碍人群中可靠检测抑郁症存在较大困难。

W2V2嵌入相较基线特征略有性能提升。在不平衡数据集上，第12层配合多层感知机取得了最高的非加权平均召回率59.9%，比梅尔频率倒谱系数高出约9%，灵敏度为中等水平（49.3%）。尽管这代表了相对改进，总体分类性能仍然不够理想。在CI平衡数据集和CI性别平衡数据集上，第2层配合随机森林分类器分别取得了最高的非加权平均召回率57.1%和61.3%。然而，在所有配置中，更高的特异度始终以灵敏度下降为代价，反映出在尽量减少假阳性的同时倾向于漏检抑郁症。

为评估泛化能力，我们在CI性别平衡数据集上训练模型，并在剩余数据上进行评估，结果并不一致。W2V2第2层结合多层感知机分类器后，非加权平均召回率下降了约3.4%；而第6层则提升了6.2%，这表明不同层的逐层鲁棒性存在差异。

在各 W2V2嵌入层中，对 D-CI 与 ND-CI 进行分类的性能比较。

表III 不同声学特征类型、数据集配置和分类器下 D-CI 与 ND-CI 的最佳分类结果。结果以准确率、非加权平均召回率、灵敏度和特异度表示，并对 5 次固定的随机数据划分取平均，以均值（标准差）% 的形式呈现。梅尔频率倒谱系数和 eGeMAPS 用作基线特征。每种数据集配置下最高的非加权平均召回率以突出显示。CLF：分类器，IMB：不平衡，CI-B：认知障碍平衡，CIG-B：认知障碍与性别平衡，TRAIN-BAL/TEST-REM：在平衡数据上训练／在剩余数据上测试，W2V-HL：Wav2Vec 2.0 隐藏层。

数据集	特征	Clf	准确率	UAR	灵敏度	特异度
IMB	梅尔频率倒谱系数	RF	58.57 (5.13)	51.78 (6.51)	28.00 (12.22)	75.56 (3.78)
CI-B		SVM	62.22 (1.48)	50.65 (1.29)	6.00 (12.00)	95.29 (9.41)
CIG-B		MLP	58.33 (6.97)	52.89 (6.76)	31.11 (8.31)	74.67 (8.84)
IMB	eGeMAPS	MLP	55.24 (4.10)	51.26 (3.74)	37.33 (11.62)	65.19 (9.83)
CI-B		SVM	61.48 (3.78)	53.76 (5.15)	24.00 (20.59)	83.53 (13.62)
CIG-B		MLP	61.67 (9.65)	59.11 (11.36)	48.89 (19.37)	69.33 (6.79)
IMB	W2V2 HL 11	MLP	63.33 (3.87)	58.44 (3.34)	41.33 (8.84)	75.56 (8.31)
IMB	W2V2 HL 12	MLP	62.86 (8.73)	**59.85 (8.86)**	49.33 (10.83)	70.37 (9.37)
CI-B	W2V2 HL 2	RF	61.48 (6.46)	57.06 (7.57)	40.00 (12.65)	74.12 (4.71)
CIG-B		RF	66.67 (7.45)	**61.33 (7.84)**	40.00 (11.33)	82.67 (8.00)
CI-B	W2V2 HL 6	SVM	62.22 (2.77)	**54.35 (2.70)**	24.00 (16.25)	84.71 (12.67)
CIG-B		RF	60.00 (4.25)	50.22 (3.40)	11.11 (0.00)	89.33 (6.80)
			平衡数据训练／剩余数据测试			
CIG-B	W2V2 HL 2	MLP	64.47 (5.54)	57.89 (2.88)	35.55 (13.93)	82.22 (15.65)
CIG-B	W2V2 HL 6	MLP	65.41 (7.46)	56.43 (4.85)	23.23 (8.01)	89.63 (15.30)

表 IV

使用 Wav2Vec 2.0 隐藏层 9 和支持向量机分类器进行 CI 与 NCI 分类的亚组性能和差异指标。 δ 表示组内性能不平衡（SP - SE）；正值反映偏向 NCI（特异度较高），负值则表示偏向 CI（灵敏度较高）。 Δ 量化组间性能差异。星号（*）表示配对t检验得出$p<0.05$的比较结果。所有数值均以百分比表示，并突出显示存在偏差的结果。SE：亚组特异性敏感度，SP：亚组特异性特异度，IMB：不平衡，CI-B：认知障碍平衡，CIG-B：认知障碍与性别平衡，N-DEPR.：非抑郁，DEPR.：抑郁。

数据集	年龄组								性别								临床状态							
	组1			组2			年龄 Δ		Male			女性			性别 Δ		非抑郁			抑郁			临床 Δ	
	Sp	灵敏度	δ	特异	Se	δ	Δ_{Sp}	Δ_{Se}	Sp	Se	δ	Sp	Se	δ	Δ_{Sp}	Δ_{Se}	Sp	Se	δ	Sp	Se	δ	Δ_{Sp}	Δ_{Se}
IMB	80	76	4	71	83	**-12**	**9**	-7	86	76	**10***	68	83	**-15***	**18***	**-7**	73	81	-8	95	81	**14***	**-22***	0
CIB	84	78	6	69	75	**-6**	**15***	3	85	69	**16***	72	80	**-8***	**13***	**-11***	74	77	-3	100	74	**26***	**-26***	3
CIG-B	79	74	5	80	74	6	-1	0	86	74	**12***	73	74	-1	**13***	0	76	77	-1	100	70	**30***	**-24***	7

不平衡数据集和CI平衡数据集上，使用不同W2V2嵌入层对D-CI与ND-CI进行分类的比较性能。不同分类器的性能趋势大体一致。在不平衡和CI平衡设置下，嵌入层2都使随机森林分类器的非加权平均召回率略有提升。此外，在不平衡数据集上，嵌入层11和12取得了相对较高的非加权平均召回率。一种可能的解释是，不平衡数据集中的类别分布差异较大（51例D-CI和88例ND-CI，见表I），因此整体非加权平均召回率更容易受到多数类ND-CI分类性能的影响。

进一步的探索性分析包括使用CI和性别平衡数据集训练用于预测认知障碍状态的模型，并在抑郁标签上对其进行测试，反之亦然。这些跨条件评估结果并不理想，性能接近随机水平（见附录表V），表明这两种状况之间的表征重叠极小。

B. 偏差分析

为评估认知障碍与无认知障碍分类中亚组层面的性能差异，采用了各数据集上表现最佳的配置

表 IV 展示了使用 W2V2第9层和支持向量机分类器进行认知障碍与无认知障碍分类时，各人口统计学和临床亚组的性能及差异指标。为进一步确定这些性能差异的来源，还分析了模型的得分分布，如图4所示。

组内性能不平衡 δ （亚组特异性特异度 - 亚组特异性敏感度）刻画了某一亚组内灵敏度与特异度之间的权衡，而组间性能差异 Δ衡量的是在给定指标（灵敏度或特异度）上各亚组之间的模型性能差异。较高的 Δ 以及较大的AUC差距是算法偏差的主要指标，表明底层声学特征对特定亚组的区分能力较弱。而较高的 Δ 但相近的AUC则表明存在阈值设定或校准问题，可通过事后缩放加以缓解。

在所有数据集配置中都观察到系统性的性别相关性能差异。组内分析显示存在显著的权衡失衡（δ）：男性参与者始终被分类为具有更高的特异度（有利于

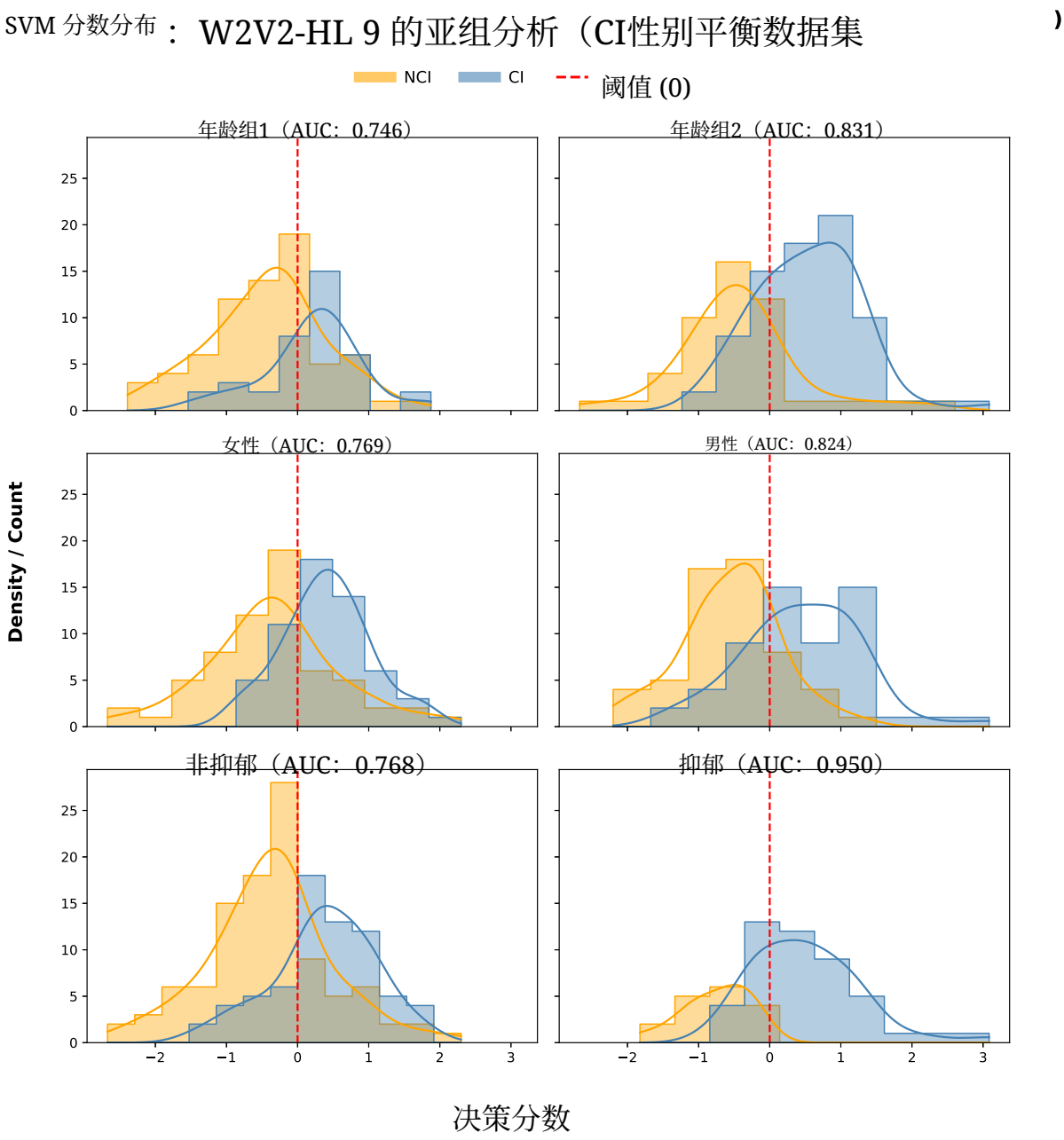


图4. 使用 CI性别平衡数据集时，Wav2Vec 2.0 隐藏层 9 在不同人口统计与临床亚组中的 SVM 决策分数分布。直方图和核密度估计（KDE）曲线展示了无认知障碍组与认知障碍组之间的区分情况。位于 0 的虚线表示 SVM 决策阈值。每个亚组均给出了 AUC 值，以表示分类性能。

NCI检测），其 δ 值为 +10%至 +16%（$p<0.05$）。相反，女性参与者明显偏向灵敏度（有利于认知障碍检测），尤其是在不平衡数据集和CI平衡数据集中（分别为δ = -15%和-8%；两者均$p<0.05$）。

这种内部失衡也导致了显著的组间性能差异（Δ）。模型在特异度方面系统性地更有利于男性（Δ_{spec} 范围为 +13%至 +18%，$p<0.05$），这表明健康女性被误分类的可能性显著高于男性。亚组特异性AUC值也支持这一点：男性的判别能力更高（AUC_{Male} = 0.824），而女性为（AUC_{Female} = 0.769）。女性亚组中更高的直方图重叠表明存在固有的算法偏差：无论决策阈值如何，模型都难以区分女性语音中的类别。

年龄人口统计特征分析显示，模型性能存在明显差距，且更有利于年龄较大的参与者（年龄组2）。虽然模型在年龄组1中保持了相对均衡的组内性能（δ 范围为 +4%至 +6%），但年龄组2则表现出强烈的偏向，过度倾向于检测认知障碍（δ 低至 -12%）。

组间调查发现特异度差距较大

（高达Δ_{spec} +15%，$p<0.05$），有利于年龄组1的NCI检测。这种差异主要源于模型判别能力的差异（算法偏差）；年龄组 1 AUC 的判别能力仅为0.746，而年龄组2 AUC 为0.831。密度图显示，较年轻亚组（年龄组1）的类别重叠程度更高，这表明底层声学特征对年轻参与者认知障碍相关变化的代表性较弱。

抑郁参与者中的性能差异最为显著，也最令人担忧。组内分析显示，非抑郁参与者的表现较为均衡（δ 介于 -1% 至 -8%）。而抑郁参与者则表现出倾向于检测为无认知障碍的保守偏差，δ 最高达 +30%，$p<0.05$），即使在进行数据平衡后仍是如此。

该模型对抑郁个体表现出较强的判别性能（$AUC_{Depressed}$ = 0.950），相比之下，$AUC_{Non-depressed}$ = 0低得多，仅为 .768。这种算法优势转化为较大的特异度差距（Δ_{spec} 低至 -26%，$p<0.05$）。抑郁参与者的得分分布重叠较少，而非抑郁亚组则存在高度重叠的密度分布。

此外，还对采用多层感知机分类器的 W2V2 第10隐藏层进行了比较性偏差分析（参见附录表VI）。尤其是在女性和抑郁参与者中，发现了相似的子组性能差异和偏差模式。重要的是，未对 D-CI 与 ND-CI 分类任务进行偏差分析，因为分类性能不尽如人意（非加权平均召回率约为 50% ），无法进行有意义的亚组比较。

总之，在认知障碍检测方面，高层W2V2嵌入的表现远优于传统声学特征，但也表现出明显的人口学与临床偏差。抑郁症分类仍是一项更具挑战性的任务，认知障碍与抑郁症之间的泛化能力有限，这凸显了开发病症特异性表征的必要性。

五、讨论

本研究采用多个分类器，评估了三种声学表征——梅尔频率倒谱系数、eGeMAPS和W2V2，用于检测认知障碍参与者中的认知障碍和抑郁症。

在认知障碍分类方面，梅尔频率倒谱系数的表现中等，而eGeMAPS的特异度明显较差，低至5.2%，凸显了其在认知障碍检测方面的有限效用。相比之下，来自第9层和第10层的W2V2嵌入始终表现更优，在所有数据集配置和分类器中的非加权平均召回率均介于76.5%至80.6%之间。这些改进可能源于更深层Transformer层对语境和语义语音特征的编码能力 [34], ，而这些特征对于检测认知障碍的典型语言与认知紊乱至关重要 [38]。在认知障碍和性别均衡的数据集中表现稳定，进一步凸显了其鲁棒性。多项研究报告的准确率高达82%，与我们的结果吻合良好；我们的结果显示准确率最高可达81.56% [13], [14],[10]。

抑郁症分类任务更具挑战性。梅尔频率倒谱系数和 eGeMAPS 的表现均不理想，非加权平均召回率约为 50%。尽管 W2V2嵌入略有改善，但整体表现仍不够理想，最高非加权平均召回率仅为 61.3%（第 2 层搭配随机森林分类器）。在W2V2嵌入中，较低层和中间层（尤其是第 2 层和第 6 层）的表现优于较高层，这与相关研究发现相符：抑郁症的声学标记（如单调音高、韵律变化减少，以及基频微扰和微扰增加）编码在较早的层中 [39], [34]。鉴于这些表现始终偏低，我们不对抑郁症分类作出明确结论，也未将偏差评估扩展到这一任务，因为如果没有足够可靠的基线模型，这类评估就没有意义。

尝试对认知障碍与抑郁症之间的重叠进行建模，但泛化效果不佳。训练用于检测认知障碍的分类器无法迁移至抑郁症检测，反之亦然，性能接近随机水平（参见附录表V）。这表明，尽管认知障碍与抑郁症可能存在共同的临床症状，但二者的声学表现各不相同，因此需要采用不同的建模策略。

使用支持向量机分类器和W2V2第9层进行的偏差分析显示，所有数据集中都存在一致的差异。分析采用组内不平衡（δ）和组间差异（Δ），并结合亚组特异性曲线下面积，以区分算法偏差与单纯的阈值设定效应。观察到系统性的性别偏差：男性始终被判定为具有较高的特异度（δ 最高达 +16%，偏向NCI检测），而女性则倾向于具有较高的灵敏度（δ 最低为-15%，偏向认知障碍检测）。这导致了显著的组间特异性差异（Δ_{spec} 最高达 +18%，$p<0.05$）。值得注意的是，女性的判别能力较低（AUC_{Female} = 0.769 对比 AUC_{Male} = 0.824），且直方图重叠程度更高（图4），这表明存在固有的算法偏差。这说明W2V2嵌入可能难以区分女性语音中与认知障碍相关的特征，而这种特性很可能源自LibriSpeech等不平衡的预训练数据[31]。

年龄相关分析显示，老年人（年龄组2）的表现明显更优。虽然年龄组2表现出偏向认知障碍检测的激进偏差（δ = -12%），但组1受限于较低的 AUC （0.746 vs. 0.831）。年龄组1的密度图中类别重叠更明显，这证实了底层声学特征对年轻参与者认知障碍相关变化的代表性较弱，导致特异度差距更大（Δ_{spec} 最高达 +15%） [38]。这些结果与先前报告的语音自监督学习模型存在性别和年龄相关差异的研究一致 [15], [16], [17]。

抑郁状态方面观察到的差异最大。非抑郁参与者的表现较为均衡，而抑郁个体则明显偏向保守地进行NCI检测（δ 最高达 +30%）。然而，得分分布近乎完美的分离，以及抑郁个体更强的判别能力（$AUC_{Depressed}$= 0.950 vs. $AUC_{Non-depressed}$ = 0.768），揭示了显著的临床状态偏差，这可能是由于数据集中抑郁的NCI个体数量较少所致（参见表I）。W2V2第10层使用多层感知机分类器时呈现出一致的趋势，这表明这些表征偏差并非仅限于单一层或分类器。然而，在CI性别平衡数据集中，与第9层相比，抑郁NCI受试者的结果不一致，

这可能反映了样本量较小的问题：一次误分类就可能显著影响准确率（表VI）。

这些发现凸显了临床语音应用中的一个关键权衡：尽管数据平衡可以缓解部分差异，但可能无法完全消除模型表征或数据中固有的偏差。这支持了开展公平感知的模型开发与评估，而非仅依赖数据平衡来确保公平的临床决策支持。重要的是，我们的工作通过明确评估基于声学的机器学习模型在不同人口统计与临床亚组中的认知障碍检测偏差，拓展了以往研究，而这一空白此前已在文献中得到强调 [17]。

然而，本研究仅关注DB的Pitt语料库，该语料库可能无法充分代表更广泛临床人群的人口统计与语言多样性。这一局限可能会影响我们研究结果的泛化性。此外，本研究的一个关键局限在于Pitt语料库中可用的抑郁标注样本数量有限 [40]。由于该语料库并非主要为抑郁症研究而设计，对抑郁病例的有限覆盖可能导致结果出现变异，并限制与抑郁症特异性分类性能相关结论的鲁棒性。

未来的工作应通过纳入规模更大、更多样且更加均衡的数据集，并采用经临床验证的抑郁标注，来解决这些差异。还应开展受试者层面分析，以更全面地刻画临床语音的变异性。将此分析扩展至其他现代自监督学习（SSL）变体（如 HuBERT、WavLM、W2V-BERT-2.0），将有助于进一步阐明不同预训练范式如何影响偏差。此外，自适应加权并融合多个嵌入层，是未来一项有前景的研究方向，尤其是在更深入了解逐层偏差模式之后。最后，还应探索减偏方法，以缓解亚组差异，支持公平、可靠的模型评估，以及人工智能在医疗领域的负责任部署。

VI. 结论

本研究考察了声学表征在认知障碍（CI）和共病抑郁症分类中的有效性，并特别关注识别不同人口统计学及临床亚组之间的性能差异。研究结果表明，W2V2嵌入，尤其是来自较高层Transformer层的嵌入，在认知障碍检测方面始终优于基线特征（梅尔频率倒谱系数和e GeMAPS），非加权平均召回率最高达到80.6%。相比之下，对认知障碍参与者进行抑郁症分类更具挑战性，使用较低层和中间层W2V2的效果相对更好。跨条件分析显示，认知障碍与抑郁症之间的泛化能力较差，这表明尽管二者存在一些重叠症状，其声学模式仍有显著差异，因此需要针对具体任务采用不同的建模策略

偏差分析揭示了不同人口统计学和临床亚组之间的性能差异。具体而言，在女性和年轻参与者中，模型表现出明显的类别重叠和较低的判别能力（AUC_{Female}= 0.769；$AUC_{Agegroup1}$ = 0.746），导致组间特异度差异显著（Δ_{spec} 最高达 +18%）。相比之下，男性、年长参与者以及患有抑郁症的参与者，获得了显著更高的判别准确率，且类别重叠较少。这些差异凸显了当前基于语音的认知障碍检测模型的关键局限，表明即使是最先进的方法，也可能无意中反映并延续训练数据中固有的偏差。

尽管我们的分析仅聚焦于单个数据集，但深入的亚组级评估为理解偏差如何产生于自监督模型中提供了见解。未来的工作应将该分析扩展到规模更大、更加多样化且经临床验证的抑郁症样本分布更加均衡的数据集，从而开展有意义的受试者级研究，更好地捕捉临床语音中的变异性。此外，还应系统评估模型去偏策略，以提升公平性和鲁棒性，并探索其他自监督学习架构（如 HuBERT、WavLM 和 W2V-BERT-2.0）的偏差特征。总体而言，本研究表明，尽管 W2V2 等自监督模型能够推动临床语音分析的发展，但它们也可能编码反映训练数据中系统性不平等的偏差。因此，基于语音的诊断工具不仅必须根据准确率进行评估，还必须评估其公平性，尤其是在多样化的临床人群中部署时。

附录

本附录提供了额外的评估结果和亚组分析，以补充正文中报告的研究结果。

表V总结了跨条件评估的结果：在该评估中，训练用于对认知障碍进行分类的模型接受抑郁标签测试，反之亦然。结果显示，模型在不同条件之间的泛化能力较差，准确率和非加权平均召回率均徘徊在50%左右，这证实了尽管认知障碍与抑郁症存在一定的症状重叠，但二者表现出不同的声学模式，无法被共享模型有效捕捉。

表VI报告了使用 W2V2第10层和多层感知机分类器，在不同数据集配置下区分认知障碍与无认知障碍的亚组特异性敏感度和特异度。与第9层的分析类似，结果显示出显著偏差，尤其是对抑郁参与者持续存在无认知障碍偏差（Δ_{spec}低至 -26%），且不同性别和年龄组之间的性能差异不一，进一步凸显了基于语音的认知障碍检测中表征和分布不平衡的问题。这些发现进一步表明，在基于临床语音的人工智能模型中进行系统的偏差评估至关重要，以确保诊断支持公平可靠。

参考文献

[1] B. Roark、M. Mitchell、J.-P. Hosom、K. Hollingshead 和 J. Kaye，《用于检测轻度认知障碍的口语衍生测量指标》，*IEEE*音频、语音与语言处理汇刊，第 19 卷，第 2081–2090 页，2011 年 9 月。[2] M. Grundman、R. Petersen、S. Ferris、R. Thomas、P. Aisen、D. Bennett、N. Foster、C. Jack、D. Galasko、R. Doody、J. Kaye、M. Sano、R. Mohs、S. Gauthier、H. Kim、S. Jin、A. Schultz、K. Schafer、R. Mulnard 和 L. Thal，《轻度认知障碍可与阿尔茨海默病及正常衰老区分开来，适用于临床试验》，《神经病学档案》，第 61 卷，第 59–66 页，2004 年 2 月。

表V CI和性别平衡数据集上的跨条件评估分类性能。使用认知障碍状态训练的模型在抑郁状态上进行测试，反之亦然。结果以准确率、非加权平均召回率、灵敏度和特异度表示，取5次固定随机数据划分的均值，并以均值（标准差）%的形式呈现。分类器：分类器；训练集-CI 测试集-D：使用认知障碍状态训练并针对抑郁状态进行测试；训练集-D 测试集-CI：使用抑郁状态训练并针对认知障碍状态进行测试；Wav2Vec 2.0隐藏层：Wav2Vec 2.0隐藏层。

Task	特征	Clf	准确率	UAR	灵敏度	特异度
训练集-CI 测试-D	W2V HL 7	SVM	48.70 (7.48)	**52.37 (6.13)**	69.56 (6.81)	35.18 (9.46)
	W2V HL 9	RF	45.22 (6.51)	49.31 (6.03)	67.69 (14.10)	30.93 (16.01)
	W2V HL 10	RF	46.09 (6.51)	49.64 (4.75)	67.46 (9.79)	31.81 (13.93)
训练集-D 测试集-CI	W2V HL 2	MLP	52.34 (5.65)	51.76 (5.61)	38.00 (9.54)	65.53 (10.84)
	W2V HL 6	MLP	57.45 (2.33)	**55.86 (3.09)**	29.41 (5.32)	82.30 (3.75)

表VI

使用 Wav2Vec 2.0 第10隐藏层及多层感知机分类器进行 CI 与 NCI 分类的亚组性能和差异指标。 δ 表示组内性能不平衡（SP - SE）；正值反映对 NCI 的偏差（特异度较高），负值则表示对 CI 的偏差（灵敏度较高）。 Δ 量化组间性能差异。星号（*）表示配对t检验结果$p<0.05$的比较。所有数值均以百分比表示，并突出显示存在偏差的结果。SE：亚组特异性敏感度，SP：亚组特异性特异度，IMB：不平衡，CI-B：认知障碍平衡，CIG-B：认知障碍与性别平衡，N-DEPR.：非抑郁，DEPR.：抑郁。

数据集	年龄组								性别								临床状态							
	组1			组2			年龄 Δ		Male			女性			性别 Δ		非抑郁			抑郁			临床 Δ	
	Sp	Se	δ	Sp	Se	δ	Δ_{Sp}	Δ_{Se}	Sp	Se	δ	Sp	Se	δ	Δ_{Sp}	Δ_{Se}	Sp	Se	δ	Sp	Se	δ	Δ_{Sp}	Δ_{Se}
IMB	79	78	1	73	79	-6	6	-1	75	76	-1	78	80	**-2**	-3	-4	73	77	-4	95	81	**14***	**-22***	-4
CIB	80	78	2	75	70	**5**	5	8	78	69	**9***	78	75	3	0	-6	74	75	-1	100	67	**33***	**-26***	8
CIG-B	72	74	-2	82	78	**4**	-10	-4	82	74	**8***	70	79	**-9***	**12***	-5	74	77	-3	88	76	**12***	**-14***	1

[3] H. B. Lee 和 C. G. Lyketsos，《阿尔茨海默病中的抑郁症：异质性及相关问题》，《生物精神病学》，第 54 卷，第 3 期，第 353–362 页，2003 年。[4] 阿尔茨海默病协会，《2023年阿尔茨海默病事实与数据》，《阿尔茨海默病与痴呆》：阿尔茨海默病协会期刊，第 19 卷，第 4 期，第 1598–1695 页，2023 年。[5] A. D. Korczyn 和 I. Halperin，《抑郁症与痴呆》，《神经科学杂志》，第 283 卷，第 139–142 页，2009 年。[6] P. Modrego 和 J. Ferr´andez，《轻度认知障碍患者的抑郁症会增加其发展为阿尔茨海默型痴呆的风险》，《神经病学档案》，第 61 卷，第 1290–3 页，2004 年 9 月。[7] S. Gregory、N. Linz、A. K¨onig、K. Langel、H. Pullen、S. Luz、J. Harrison 和 C. W. Ritchie，《远程数据收集、语音分析及预测高危人群中阿尔茨海默病生物标志物的识别：电话语音评估（SPeAk）前瞻性观察研究方案》，《*BMJ Open*》，第 12 卷，第 3 期，文章编号 e052250，2022 年。[8] J. Chen、J. Ye、F. Tang 和 J. Zhou，《仅使用自发语音自动检测阿尔茨海默病》，载于*Interspeech*，第 2021 卷。NIH 公共访问，2021 年，第 3830–3834 页。[9] R. Pappagari、J. Cho、L. Moro-Velazquez 和 N. Dehak，《运用先进的说话人识别和自然语言处理技术检测阿尔茨海默病并评估其严重程度》，载于*Interspeech*，2020 年，第 2177–2181 页。[10] A. Balagopalan 和 J. Novikova，《比较用于检测阿尔茨海默病的基于声学的方法》，载于*Interspeech 2021*，2021 年，第 3800–3804 页。[11]B. Abdallah-Qasaimeh 和 S. Ratt´e，《通过语音分析检测阿尔茨海默病和轻度认知障碍中的抑郁症》，《理论与应用信息技术杂志》，第 99 卷，1171，2021 年 5 月，第 1162 页。[12] B. Mohammad Abdallah Qasaimeh、A. Abdallah 和 S. Ratte，《使用人工神经网络（ANN）检测阿尔茨海默病和轻度认知障碍中的抑郁症》，载于*2021*年数据科学、电子学习与信息系统国际会议，DATA'21 系列。美国纽约州纽约市：计算机协会，2021 年，第 250–253 页。

[13] P. P´erez-Toro, D. Rodr´ıguez-Salas, T. Arias-Vergara, S. Bayerl, P. Klumpp, K. Riedhammer, M. Schuster, E. N¨oth, A. Maier, and J. Orozco-Arroyave，"利用 forestnets 迁移量化情绪知识以检测阿尔茨海默病中的抑郁症"，载于*ICASSP 2023-2023 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2023年，第1–5页。[14] F. Braun, S. P. Bayerl, P. A. P´erez-Toro, F. H¨onig, H. Lehfeld, T. Hillemacher, E. N¨oth, T. Bocklet, and K. Riedhammer，"抑郁症共病情况下的痴呆分类：一项跨语料库研究"，载于*Interspeech 2023*论文集，2023年，第2308–2312页。[15] M. Z. Boito, L. Besacier, N. Tomashenko, and Y. Est`eve，"自监督模型对语音转文本系统性别影响的研究"，*arXiv*预印本 *arXiv:2204.01397*，2022年。[16] M. Fuckner, S. Horsman, P. Wiggers, and I. Janssen，"揭示自动语音识别系统中的偏差：评估面向荷兰语使用者的 wav2vec2 和 whisper"，载于*2023*年语音技术与人机对话国际会议（*SpeD*）。IEEE，2023年，第146–151页。[17] Y.-C. Lin, T.-Q. Lin, H.-C. Lin, A. T. Liu, and H.-y. Lee，"论语音自监督模型的社会偏差"，载于*Interspeech 2024*，2024年，第4638–4642页。[18] S. Hu, X. Xie, M. Geng, Z. Jin, J. Deng, G. Li, Y. Wang, M. Cui, T. Wang, H. Meng, and X. Liu，"用于构音障碍者和老年人语音识别的自监督自动语音识别模型与特征"，*IEEE/ACM*音频、语音与语言处理汇刊，第32卷，2024年，第3561–3575页。[19] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle，"阿尔茨海默病的自然病程：研究队列描述与诊断准确性"，《神经病学档案》，第51卷，第6期，第585–594页，1994年。[20] H. Goodglass and E. Kaplan，《失语症及相关障碍评估》。Lea & Febiger，1983年。[21] P. Seath, L. E. Macedo-Orrego, and L. Velayudhan，"早发性与晚发性阿尔茨海默病的临床特征：系统综述与荟萃分析"，《国际老年精神病学》，第36卷，第12期，第1093–1109页，2024年。[22] M. F. Folstein, S. E. Folstein, and P. R. McHugh，"简易精神状态检查:

一种供临床医生评估患者认知状态的实用方法”，《精神病学研究杂志》，第12卷，第3期，第189–198页，1975年。[23] M. Hamilton，“汉密尔顿抑郁量表——抗抑郁药物研发的助力还是阻碍”，《精神病学》，第23卷，第1期，第56–62页，1960年。[24] M. Zimmerman, J. H. Martinez, D. Young, I. Chelminski, and K. Dal- rymple，“汉密尔顿抑郁量表的严重程度分类”，《情感障碍杂志》，第150卷，第2期，第384–388页，2013年。[25] V. T. Tiwari，“MFCC 及其在说话人识别中的应用”，Int. J. Emerg. Technol.，第1卷，2010年1月。[26] N. Dave，“语音识别中的特征提取方法：线性预测倒谱、感知线性预测和梅尔频率倒谱系数”，International Journal For Advance Research in Engineering And Technology (ISSN 2320-6802)，第1卷，2013年7月。[27] A. Meghanani, A. C. S., and A. G. Ramakrishnan，“利用对数梅尔频谱图和梅尔频率倒谱系数特征从自发语音中识别阿尔茨海默病痴呆”，载于2021年IEEE口语技术研讨会 (SLT)。IEEE，2021年，第670–677页。[28] M. L. B. Pulido, J. B. A. Hern´andez, M. ´A. F. Ballester, C. M. T. Gonz´alez, J. Mekyska, and Z. Sm´ekal，“阿尔茨海默病与自动语音分析：综述”，《专家系统与应用》，第150卷，文章113213，2020年。[29] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andr´e, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong，“用于语音研究与情感计算的日内瓦极简声学参数集 (GeMAPS)”，《IEEE情感计算汇刊》，第7卷，第2期，第190–202页，2016年。[30] F. Eyben，“标准基线特征集”，载于《通过大规模音频特征空间提取进行实时语音与音乐分类》，Springer博士论文丛书。Springer，Cham，2016年，第123–137页。[31] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli，“wav2vec 2.0：一种用于语音表征自监督学习的框架”，载于《神经信息处理系统进展》，第33卷，2020年，第12 449–12 460页。[32] X. Xu, Y. Kang, S. Cao, B. Lin, and L. Ma，“探索 wav2vec 2.0 在发音错误检测中的应用”，载于 Interspeech 2021论文集，2021年，第4428–4432页。[33] F. AI，“Wav2vec 2.0 基础版”，2024年，访问日期：2024年7月8日。 [在线]。网址：https://huggingface.co/facebook/wav2vec2-base-960h [34] A. Pasad, J.-C. Chou, and K. Livescu，“自监督语音表征模型的逐层分析”，载于2021年IEEE自动语音识别与理解研讨会 (ASRU)。IEEE，2021年，第914–921页。[35] S. Rathore, M. Habes, M. A. Iftikhar, A. Shacklett, and C. Davatzikos，“阿尔茨海默病及其前驱期的基于神经影像分类研究与相关特征提取方法综述”，《神经影像》，第155卷，第530–548页，2017年。[36] K. L´opez-de-Ipi˜na, J. B. Alonso-Hern´andez, J. Sol´e-Casals, C. M. Travieso-Gonz´alez, A. Ezeiza, M. Fa´undez-Zanuy, P. M. Calvo, and B. Beitia，“基于适用于阿尔茨海默病诊断的非线性语音建模，对情绪反应进行自动分析的特征选择”，《神经计算》，第150卷，第392–401页，2015年，信息处理与机器学习应用专刊。[37] A. Bonvino, E. Cornacchia, G. F. Scaramuzzi, D. Gasparre, V. Manippa, D. Rivolta, and P. Taurisano，“轻度认知障碍数字工具：诊断准确性与方法学挑战的系统综述和荟萃分析”，《神经心理学评论》，第1–13页，2025年。[38] K. C. Fraser, J. A. Meltzer, and F. Rudzicz，“语言特征可从叙述性语音中识别阿尔茨海默病”，《阿尔茨海默病杂志》，第49卷，第2期，第407–422页，2015年。[39] D. M. Low, K. H. Bentley, and S. S. Ghosh，“利用语音自动评估精神障碍：系统综述”，《喉镜调查耳鼻喉学》，第5卷，第1期，第96–116页，2020年。[40] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle，“阿尔茨海默病的自然病程：研究队列描述与诊断准确性”，《神经病学档案》，第51卷，第6期，第585–594页，1994年。