arXiv:2602.02955v1 [cs.SD] 3 Feb 2026

# 医学音频分类的合成数据增强：初步评估

David McShannon[1],Anthony Mella[1],Nicholas Dietrich[2]1独立研究者2多伦多大学

## 摘要

医学音频分类仍然面临诸多挑战，包括信噪比较低、判别特征细微以及类内变异性显著，且类别不平衡和训练数据有限往往会使情况更加复杂。合成数据增强被提出作为缓解这些限制的一种潜在策略；然而，既往研究报告的方法不一致，实证结果也不尽相同。在这项初步研究中，我们使用一个在中度不平衡数据集（73%:27%）上训练的深度卷积神经网络基线模型，探究合成增强对呼吸音分类的影响。我们在受控实验条件下评估了三种生成式增强策略（变分自编码器、生成对抗网络和扩散模型）。未采用数据增强的基线模型 F1 分数为 0.645。单独采用各种增强策略均未带来性能提升，若干配置的分类性能持平或有所下降。只有增强模型的集成取得了小幅提升，F1 分数达到 0.664。这些发现表明，对于医学音频分类，将合成增强应用于标准卷积神经网络分类器时，未必能稳定提升性能。未来工作应着重厘清任务特定的数据特征、模型与增强方法的兼容性，以及使合成增强在医学音频应用中发挥效用所需的评估框架。

## 1 引言

医学音频分类具有挑战性，原因在于信噪比较低、判别性声学特征较为细微，以及显著的类内变异性 [1, 2, 3]。类别不平衡和标注数据的有限可得性进一步加剧了这些困难 [4]。在呼吸音分析中，疾病阳性录音尤为稀缺，这反映了疾病患病率、临床数据收集的实际障碍以及隐私约束 [5, 6, 7]。因此，在此类数据集上训练的模型对少数类的性能可能下降，而这些少数类往往对应于临床上重要的疾病状态 [6, 7]。

利用深度生成模型进行合成数据增强已被提出作为一种潜在方法，通过生成额外的疾病阳性样本来应对数据稀缺和类别不平衡 [8, 9, 10, 11]。生成建模的最新进展，包括变分自编码器（变分自编码器）、生成对抗网络（生成对抗网络）以及基于扩散的模型，在自然图像合成和数据增强任务中取得了优异表现 [12, 13]。这些进展促使人们越来越有兴趣将合成增强技术扩展到医学音频数据 [1, 2, 6, 14, 15]。

尽管受到广泛关注，生成高质量合成医学音频仍然困难。合成数据的保真度受可用源录音的质量和多样性限制，而这些录音在临床环境中往往数量有限且带有噪声 [10, 13, 15, 16]。因此，在现实的低数据条件下，合成增强对医学音频分类的实际价值仍不确定。

在这项初步研究中，我们探讨合成数据增强对医学音频分类性能的影响。我们评估了常用的生成式增强方法，并考察其对下游分类的影响；下游分类使用从随机初始化开始训练的深度卷积神经网络（卷积神经网络，CNN）。

## 2 方法

### 2.1 数据集

选取公开可用的 COVID-19数据集（Project Coswara[17]）来代表中度不平衡场景。该数据集包含通过网络应用程序收集的自录呼吸音，包括咳嗽、呼吸和语音样本。在本研究中，我们使用了标记为健康或感染者（确诊为COVID-19阳性）的参与者的咳嗽录音，排除了标记为暴露者或康复者的参与者。最终获得4,963条咳嗽样本：其中3,847条来自健康参与者，1,116条来自COVID-19感染者，类别比例约为3.4:1。所有录音均重采样为16 kHz单声道，并通过对较长样本进行中心裁剪、对较短样本进行零填充，将其分割为固定时长3秒的片段。该数据集被划分为

训练集（70%）、验证集（15%）和测试集（15%），采用分层随机抽样，并在各个划分中保持类别比例一致。仅对训练集进行了合成增强；验证集和测试集仅包含真实录音。

## 2.2 数据预处理

音频预处理遵循了声学分类任务的标准做法。将原始波形通过短时傅里叶变换转换为梅尔频谱图，采用 2048 点快速傅里叶变换、512 个采样点的步长以及汉宁窗。梅尔频率表示使用 128 个梅尔频带，覆盖 0–8000 Hz。频谱图采用基于训练集计算的逐通道 Z 分数归一化进行归一化。

## 2.3 基线卷积神经网络模型

该卷积神经网络架构专为频谱图分类设计，由四个卷积块组成（分别使用 32、64、128、256 个滤波器，每个卷积核为 $3 \times 3$），并包含批归一化、ReLU 激活和 $2 \times 2$ 最大池化。模型使用交叉熵损失和 Adam 优化器训练 100 个周期（lr$=0.001$，采用余弦退火），并根据验证集宏平均 F1 分数进行早停（耐心值为 $=15$ 个周期）。分类器从头开始训练，以隔离合成增强的效果，避免外部知识可能带来的影响。

## 2.4 合成数据生成

**变分自编码器：** 我们训练变分自编码器，以学习梅尔频谱图的连续潜在表示。编码器由四个卷积块组成（每个卷积块包含两个二维卷积层，滤波器数量分别为 64、128、256 和 512，随后进行批归一化和 LeakyReLU 激活），接着是全局平均池化和全连接层，将数据映射到 128 维潜在空间。解码器使用转置卷积，采用与编码器镜像对称的架构。模型使用复合损失进行训练，该损失结合了均方误差重建损失和 KL 散度正则化（$\beta = 0.1$），使用 Adam 优化器（lr$=0.0001$）训练 200 个周期。合成样本通过从潜在先验$\mathcal{N}(0, I)$中采样并解码生成。

**生成对抗网络：** 我们采用带梯度惩罚的 Wasserstein 生成对抗网络（WGAN-GP）来解决训练不稳定性问题。生成器架构由扩展 128 维噪声向量的全连接层组成，后接四个逐步上采样的转置卷积块。判别器使用了四个带谱归一化的卷积块。两个网络均使用批归一化和 LeakyReLU 激活。训练共进行 300 个周期，判别器与生成器的更新比例为 5:1，使用 RMSprop 优化器（lr$=0.00005$）和梯度惩罚系数 $\lambda = 10$。样本由高斯噪声 $z \sim \mathcal{N}(0,\ I)$生成。

**扩散模型：** 我们实现了基于 U-Net[18] 的去噪扩散概率模型。U-Net 架构包含五个编码器块和解码器块，并采用跳跃连接，在 $64 \times 64$分辨率瓶颈层引入多头自注意力。前向扩散过程按照线性方差调度，在 $T = 1000$ 个时间步内添加高斯噪声。模型使用简化的 L2 损失预测噪声分量，并使用 AdamW 优化器（lr$=0.0001$，权重衰减$=0.01$）训练 400 个周期。在推理过程中，我们采用去噪扩散隐式模型采样，以 50 步加快生成，同时保持样本质量。

所有生成模型均仅使用少数类（COVID阳性）样本进行训练。每次训练时，生成合成样本，将少数类的样本数量增加 50%（即在 1,116 个真实 COVID阳性样本的基础上增加 558 个合成样本）。

## 2.5 集成方法

我们采用集成方法，结合在不同增强条件下训练的多个模型所做出的预测。在推理时，四个模型——基线模型以及分别采用 VAE、生成对抗网络和基于扩散模型的数据增强进行训练的模型——独立处理输入的梅尔频谱图，并生成类别概率估计。

对于每个测试样本，模型 $i$ 对 COVID阳性类别的预测概率记为 $p_i$(COVID+)，并在各模型间取平均值，以得到集成预测：

$$P_{\text{ensemble}}(\text{COVID+}) = \frac{1}{4}\sum_{i=1}^{4} p_i(\text{COVID+}) \qquad (1)$$

最终类别标签根据平均概率最高的类别确定。这种集成策略无需在各个模型之外进行额外训练，并为每个组成模型赋予相同权重。

## 2.6 评估协议

分类性能通过宏平均F1分数和受试者工作特征曲线下面积（AUROC）进行评估。宏平均F1分数被用作主要指标，因为它对各类别的性能赋予相同权重，适用于类别不平衡的分类场景。AUROC通过改变应用于预测概率的分类阈值计算得出，用于概括模型在所有可能阈值下的区分性能。两项指标均在留出测试集上计算，该测试集仅包含真实（非合成）样本。

# 3 结果

基线卷积神经网络的宏平均F1分数为0.645。基于变分自编码器的数据增强后，F1分数为0.646（+0.001）。基于生成对抗网络的数据增强后，F1分数为0.609（−0.036）。基于扩散的增强后，F1分数为0.644（−0.001）。集成方法的F1分数为0.664（+0.019）。

AUROC值呈现相似的分布。基线卷积神经网络的AUROC为0.745。基于变分自编码器的数据增强得到的AUROC为0.748，基于扩散的增强达到0.746，而基于生成对抗网络的数据增强得到的AUROC较低，为0.726。集成模型取得了最高的AUROC（0.761）。图1展示了不同数据增强方法下的宏平均F1分数和AUROC。

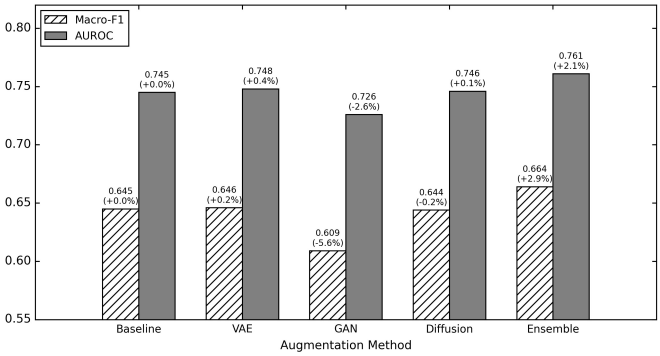


图1：不同数据增强方法下卷积神经网络分类器的宏平均F1分数和AUROC。

# 4 讨论

本初步研究评估了在中度不平衡、低数据量条件下，合成数据增强是否能够改善呼吸音分类。在三种常用的生成式方法中，当合成样本被直接纳入训练集时，我们未观察到分类性能的一致提升。这些结果表明，在该条件下，使用标准卷积神经网络进行医学音频分类时，合成数据增强可能无法可靠地提升性能。

我们的研究结果总体上与医学音频分类领域的既有研究一致 [3, 5, 6, 7, 19, 20], ，即合成增强带来的收益并不稳定，且取决于具体任务。若干因素可能解释了为何单独采用各种数据增强策略所观察到的收益有限。首先，医疗音频信号具有细微且相互重叠的声学特征，呼吸音（如咳嗽声）尤其如此，因为病理样本与健康样本具有相似的时序结构和频谱结构[6, 21]。使用数量有限且含噪的少数类数据训练的生成模型，可能难以学习具有判别性的变化，反而会复现源录音中占主导地位的模式或噪声 [4, 22]。当少数类包含的独特样本相对较少时，这一局限可能会进一步加剧，从而限制生成式学习可利用的数据多样性。

有趣的是，集成方法带来了小幅性能提升。尽管大多数单独使用增强数据的模型表现不及基线模型，仍然出现了这一提升。这一发现表明，不同的数据增强策略可能会导致彼此不同且部分不相关的误差模式，而集成可以利用这一点 [1, 23, 24]。在这种情况下，合成增强的价值似乎不在于提升单个模型的训练效果，而更多在于促进模型间的多样性。不过，这种优势需要付出计算复杂度和训练时间增加的代价，因此可能会限制此类方法在资源受限的临床环境中的实用性 [11]。

未来的研究应探索合成数据在直接扩充训练集之外的其他用途[2, 25, 26, 27]。合成样本或许更适用于预训练、表示学习或自监督目标，在这些场景中，分布失配造成的危害可能较小。此外，将领域约束或生理学先验融入生成模型，或许能提高合成医学音频的真实性和实用性。提升源数据质量（包括降噪和标准化录音协议）也可能是有效生成式数据增强的关键前提。此外，未来研究还应考察合成数据生成与对抗鲁棒性之间的相互作用，包括合成样本是否会影响模型对医疗音频输入中噪声、伪影或有意扰动的敏感性 [28, 29]。

本研究存在若干局限。首先，我们仅在单一数据集上开展评估，且研究重点为一项呼吸音分类任务。尽管该数据集呈现了较为真实的中度类别不平衡，但研究结果未必能推广至其他医疗音频领域，例如肺部听诊、心音或言语病理学。其次，我们评估了一个采用随机初始化训练的单一基线卷积神经网络。虽然这一设计旨在单独考察合成数据的贡献，但替代架构或预训练策略与增强数据的交互方式可能有所不同。第三，我们采用宏平均 F1 分数和 AUROC 评估模型性能，且实验使用的是单次训练实例，而非多次重复运行；因此，未进行正式的推断性检验。鉴于此，应将这些发现视为初步探索性结果，而非针对所有可能的模型与增强组合开展的全面推断性评估。

## 5 结论

我们的结果表明，使用当前生成模型进行合成数据增强并直接应用于训练数据，并不能稳定提升呼吸音分类性能。集成方法虽可借助数据增强带来的多样性获得有限收益，但将合成样本简单混入训练集仍需谨慎。这些发现凸显了严格评估的重要性，也表明对于医疗音频任务而言，合成增强的前景仍取决于具体条件和应用情境。

## 参考文献


[1] L. Nanni、G. Maguolo、S. Brahnam 和 M. Paci。用于音频分类的卷积神经网络集成。应用科学，2021年。[2] L. Xiao、L. Fang、Y. Yang 和 W. Tu。LungAdapter：面向肺音分类的高效音频频谱图变换器适配方法。载于*Interspeech 2024*，第 4738–4742 页，2024年。[3] N. Dietrich、D. McShannon 和 M. F. Rzepka。使用多模态语言模型评估基于频谱图的肺音分类中的少样本提示。*PLOS Digital Health*，5:e0001179，2026年。[4] D. Dablain、K. N. Jacobson、C. Bellinger、M. Roberts 和 N. Chawla。理解卷积神经网络在不平衡数据学习中的脆弱性。*arXiv* 预印本，2022年。[5] S. Yu、J. Yu、L. Chen、B. Zhu、X. Liang、Y. Xie 等。呼吸音分析的进展与挑战：基于 ICBHI2017 数据库的技术综述。*Electronics*，14:2794，2025年。[6] B. M. Rocha、D. Pessoa、A. Marques、P. Carvalho 和 R. P. Paiva。异常呼吸音的自动分类：一个（尚）未解决的问题？*Sensors*，21，2021年。[7]B. M. Rocha、D. Filos、L. Mendes、G. Serbes、S. Ulukaya、Y. P. Kahya 等。用于评估呼吸音分类算法的开放获取数据库。*Physiological Measurement*，40，2019年。[8] V. C. Pezoulas、D. I. Zaridis、E. Mylona、C. Androutsos、K. Apostolidis、N. S. Tachos 等。医疗保健领域的合成数据生成方法：开源工具与方法综述。*Computationaland Structural Biotechnology Journal*，23:2892–2910，2024年。[9] G. Zhu、Y. Wen 和 Z. Duan。基于分数的生成建模实现音频生成：设计原则与实现。*arXiv* 预印本，2026年。[10]H. M. Waseem、S. U. Islam、N. Matragkas、G. Epiphaniou、T. N. Arvanitis 和 C. Maple。《生成式人工智能合成数据生成综述：医疗保健视角》。《人工智能评论》，59:55，2025年。[11] N. Dietrich 和 K. Hanneman。《绿色设计：权衡放射学人工智能开发的环境影响》。《加拿大放射科医师协会杂志》，2025年。[12] A. Gonzales、G. Guruswamy 和 S. R. Smith。《医疗保健中的合成数据：叙述性综述》。*PLOSDigital Health*，2:e0000082，2023年。[13] G. Pasculli、M. Virgolin、P. Myles、A. Vidovszky、C. Fisher、E. Biasin 等。《医疗保健与药物开发中的合成数据：定义、监管框架与问题》。《*CPT*：药物计量学与系统药理学》，14:840–852，2025年。[14] S. Bae、J.-W. Kim、W.-Y. Cho、H. Baek、S. Son、B. Lee 等。《基于音频频谱图变换器的 Patch-Mix 对比学习，用于呼吸音分类》。*arXiv* 预印本，2024年。[15] A. Kala 和 M. Elhilali。《用于肺爆裂音数据增强的约束合成采样》。载于*IEEE* 医学与生物学工程学会年度国际会议，第 1–5 页，2023年。[16] S. Ghosh、S. Kumar、Z. Kong、R. Valle、B. Catanzaro 和 D. Manocha。《Synthio：利用合成数据增强小规模音频分类数据集》。*arXiv* 预印本，2025年。[17] D. Bhattacharya、N. K. Sharma、D. Dutta、S. R. Chetupalli、P. Mote、S. Ganapathy 等。《Coswara：用于远程筛查严重急性呼吸综合征冠状病毒2型感染的呼吸音和症状数据集》。《科学数据》，10:397，2023年。[18] O. Ronneberger、P. Fischer 和 T. Brox。《U-Net：用于生物医学图像分割的卷积网络》。*arXiv* 预印本，2015年。[19] P. Prodhan、R. S. Dela Rosa、M. Shubina、K. E. Haver、B. D. Matthews、S. Buck 等。《儿科重症监护病房中的哮鸣音检测：医师、护士、呼吸治疗师与计算机化呼吸音监测仪之间的比较》。《呼吸护理》，53，2008年。[20]Y. Kim、Y. Hyon、S. S. Jung、S. Lee、G. Yoo、C. Chung 等。《在临床领域使用深度学习对肺爆裂音、哮鸣音和鼾音进行呼吸音分类》。《科学报告》，11:17186，2021年。

[21] M. Sarkar、I. Madabhavi、N. Niranjan 和 M. Dogra。《呼吸系统听诊》。《胸部医学年鉴》，10:158，2015年。[22] H.-C. Chu、Y.-L. Zhang 和 H.-C. Chiang。《一种使用数据增强的卷积神经网络声音分类机制》。*Sensors*，23，2023年。[23] A. Karim、S. Ryu 和 I. C. Jeong。《生物医学信号分类的集成学习：基于叩诊和触诊频谱图的高准确率框架》。《科学报告》，15:21592，2025年。[24]R. Nadkarni、E. Nikolakakis 和 R. Marinescu。《AFEN：使用集成学习进行呼吸系统疾病分类》。*arXiv* 预印本，2024年。[25] N. Dietrich。《放射学中的智能体人工智能：新兴潜力与尚未解决的挑战》。《英国放射学杂志》，2025年。[26] A. M. Qureshi、A. Kaushik、R. Loughran、K. McDaid 和 F. McCaffery。《通过合成数据生成提升医疗数据质量：综述》。《健康信息学与生物信息学中的网络建模分析》，15:46，2026年。[27] S. Chakraborty、P. Kochhar、S. Patil、K. Kotecha、S. Gite、G. Selvachandran 等。《利用生成对抗网络增强数据改进心音异常检测》。《生物学与医学中的计算机》，195:110623，2025年。[28] N. Dietrich、B. Gong 和 M. N. Patlas。《放射学中的对抗性人工智能：攻击、防御与未来考量》。《诊断与介入影像学》，2025年。[29] S. G. Finlayson、J. D. Bowers、J. Ito、J. L. Zittrain、A. L. Beam 和 I. S. Kohane。《针对医疗机器学习的对抗性攻击》。《科学》，363:1287–1289，2019年。