arXiv:2305.03711v2 [cs.LG] 8 Jan 2024

# 医疗记录凝缩：迈向医疗保健数据民主化的路线图

Yujiang Wang[1,4],Anshul Thakur[1*]、Mingzhi Dong[2], Pingchuan Ma[3], StavrosPetridis[3], Li Shang[2*]、Tingting Zhu[1] 和David A. Clifton[1,4]

[1*]牛津大学工程科学系，英国。[2*]复旦大学计算机科学学院，中国。[3]伦敦帝国学院计算机系，英国。[4]牛津苏州先进研究中心，中国。

*通讯作者。电子邮箱：anshul.thakur@eng.ox.ac.uk；lishang@fudan.edu.cn；贡献作者：yujiang.wang@oxford-oscar.cn；mingzhidong@gmail.com；pingchuan.ma16@imperial.ac.uk；stavros.petridis04@imperial.ac.uk；tingting.zhu@eng.ox.ac.uk；david.clifton@eng.ox.ac.uk；

**摘要**

人工智能（AI）的普及，预示着医疗保健民主化时代的到来，为每一位利益相关方带来全新且更美好的生活方式。然而，医疗保健领域数据民主化的缺失，严重阻碍了临床人工智能研究的发展。要真正实现人工智能研究的数据民主化，需要应对两方面的挑战：1. 应对临床数据中的敏感信息进行适当匿名化处理；2. 以人工智能为导向的临床知识应能在各组织之间自由流通。本文将近期深度学习领域的一项进展——数据集凝缩（DC）——视为一举两得、推动医疗保健数据民主化的利器。DC 后得到的凝缩数据可视为统计元数据，它概括了原始临床记录，并从个体层面不可逆地隐藏敏感信息；与此同时，仍保留了足够的知识，用于训练深度神经网络（DNNs）。更重要的是，凝缩数据占用空间更小、模型学习速度更快，这意味着临床知识共享将更加高效。

以及数据民主化所要求的流畅系统。我们通过对三个数据类型各异的医疗保健数据集进行实验并分析结果，阐明了数据凝缩（DC）在推动临床数据民主化方面的前景，尤其是通过人工智能研究实现电子医疗记录（EHR）的民主化。

**关键词：** 医疗保健数据民主化、数据集凝缩、死亡率预测、新冠肺炎诊断

# 1 引言

人工智能（AI）在医疗保健领域的普及，使这一领域以前所未有的方式发展。凭借自动化医师工作流程的能力，智能计算有望减轻临床从业者的工作负担，辅助疾病诊断与解读，改善医疗服务，促进患者就诊并实现个性化服务等。[1, 2]为了实现更具预测性、可及性和个性化的医疗服务，我们积极推动人工智能系统在临床领域的部署。例如，全球医疗保健人工智能市场预计将从2022年的194.5亿美元增长至2032年的2807.7亿美元，年增长率为30.6%。[3]智能临床技术的蓬勃发展预示着医疗保健民主化时代即将到来，[4], 这将为医疗生态系统中的每一位利益相关者开启一种全新且更美好的生活方式。借助人工智能，医师可以将耗时的日常工作自动化，并专注于能够最大限度发挥自身价值的领域；患者可以以更低的成本获得更优质、个性化的医疗服务；分析师则可以以前所未有的规模从临床数据中提取并提供洞见。这些只是其中一些潜在益处。

尽管医疗保健民主化展现了理想愿景，临床人工智能研究依然充满挑战且要求严苛，其“阿喀琉斯之踵”可归结为一个简单却长期存在的障碍：缺乏医疗保健数据民主化 [5, 6]。数据为智能算法提供动力。然而，与其他行业相比，医疗保健领域的数据民主化进程滞后，尤其是在组织之间实现真正开放且具备互操作性的数据共享方面。医疗记录的数字化，即电子健康记录（EHR） [7],，为患者提供了有效的方式来共享和管理自己的信息；然而，大多数电子健康记录仍与那些热衷于开发智能系统、却未获授权的人工智能研究人员相互隔绝。因此，医疗保健领域的数据民主化仍是遥不可及的乌托邦，严重阻碍了临床人工智能的发展，也妨碍医疗保健真正实现民主化。

在这项工作中，我们致力于解决推动医疗保健数据民主化、以供人工智能研究使用这一问题的一个关键环节。“数据民主化”一词特指临床人工智能研究人员之间开放、自由地共享医疗记录数据，尤其是电子健康记录（EHR），以开发和部署智能算法。然而，这面临两方面的挑战：

• **临床数据具有敏感性**。临床记录的敏感性是数据民主化的一大障碍，因为它们受数据监管法律保护，未经适当的匿名化或去标识化，不能公开。匿名化意味着无法将任何患者与特定的 EHR 条目关联起来。经过适当匿名化的医疗保健数据集不被视为敏感数据，因此可以自由共享，而不会违反法律或法规。然而在实践中，常用的医疗保健数据匿名化技术 [8]未必总能生成真正去标识化的数据集，在各种情形下，重新识别患者的风险不可避免地存在 [9——11]因此，亟需建立一种信息共享系统，既能消除医疗保健数据的敏感性，又能保留推动人工智能创新所必需的知识和资源。

• **临床数据的维护、传输和学习成本可能非常高昂**。为了实现人工智能研究的民主化，信息应当在各组织之间自由、开放地流通，因此需要一种快速便捷的数据共享策略。然而，自医疗保健数字化以来，电子健康记录（EHR）的数量呈指数级增长，维护或传输大规模临床数据集不仅会产生高昂的经济成本，还需要耗费大量时间。管理不善或遭受黑客攻击所引发的潜在数据泄露风险，可能会进一步加剧这一问题。另一个副作用是，利用此类大规模数据开发智能算法所需的计算成本可能极其高昂，因此需要配备强大但价格昂贵的图形处理器（GPU） [12], ，这会减缓人工智能研究中的试错迭代周期，并形成阻碍民主化进程的硬件壁垒。维护、传输和学习临床数据所产生的高昂成本，亟需建立一个更加高效、便捷的信息流通系统，以支持医学人工智能研究。

在这项工作中，我们关注人工智能领域近期的一项成果——数据集凝缩（DC）[13], ，它有望成为推动医疗保健数据民主化的有力候选方案。DC 是一种深度学习技术，旨在学习一个“凝缩”数据集，使其在深度模型学习的情境中，尽可能保留原始数据集中的知识。这个凝聚数据集能够引导深度神经网络（DNNs）在不观察任何具有相似模式的原始数据的情况下，仍能在真实的数据点上正常运行。尽管听起来颇具启发性，DC 在医疗保健领域仍是一项尚未探索的技术，尤其是其在推动数据民主化方面的潜力。我们发现，它具备两项鲜明优势，使其成为推动人工智能研究民主化的有力候选方案：一是 DC 能够移除敏感的个体层面信息，同时充分保留人工智能效用；二是能够有效压缩大规模数据。

•**压缩数据是不敏感的元数据，可以自由共享**。借助其独特的学习范式，数据凝缩不可逆地隐藏原始数据集中的个体层面信息，从而实现真正的去标识化。具体而言，数据凝缩旨在匹配嵌入分布

原始数据与凝缩数据之间的差异由深度神经网络计算。深度神经网络的参数并非通过显式学习获得，而是在每次训练迭代时随机初始化。随后，从凝缩数据和原始数据中分别选取两个随机批次，在这些深度神经网络的嵌入空间中学习前者，使其分布与后者相匹配。这一范式可谓一举两得：知识通过参数随机化、充当先天加密器的深度神经网络，不可逆地从原始数据流向凝缩数据。此外，将随机抽取的原始数据批次与凝缩数据批次进行匹配，会形成多对一的对应关系；也就是说，每个凝聚样本都能保留从众多原始实例中压缩而来的知识。由于个体层面的信息已被彻底移除，从理论和计算角度而言，都无法将患者与凝缩数据中的特定电子健康记录关联起来。我们可以将数据凝缩后的数据视为统计数据，其中包含足以驱动深度模型学习的群体层面知识。由于凝缩数据属于不敏感的元数据，我们可以自由共享这些数据，推动医疗保健人工智能研究的民主化，而无需担心法律问题。

- **压缩数据能够为人工智能研究提供更高效的信息流**。顾名思义，数据凝缩的另一项显著优势是能够大幅压缩原始电子健康记录的大小，从而加快组织之间的数据传输，促进互操作性。占用空间的大幅减少也意味着维护凝缩数据的经济成本更低。更重要的是，由于凝缩后的医疗保健数据不包含个人层面的敏感信息，因此可免受因重新识别患者而导致的数据泄露风险。数据量压缩还能带来额外好处：深度神经网络的学习效率可显著提升，其收敛速度远快于使用原始数据时。数据凝缩的指标描绘了人工智能研究人员之间自由开放的医疗知识交流图景。

本研究首次探讨了将 DC 应用于医疗保健人工智能研究民主化的前景。我们将 DC 应用于三个临床变量和任务各不相同的医疗保健数据集，以生成压缩后的训练集版本。我们采用了一个由 11 种不同架构的深度神经网络组成的队列，以评估原始数据和压缩数据在深度学习中的效用。我们还详细阐述了压缩数据在实现高效数据流动和加速模型收敛方面的优势。随后，我们讨论了压缩数据固有的元数据属性，并展示了其相对于原始数据的不同分布情况。最后，尽管数据集凝缩仍处于起步阶段，我们展示了其有望使我们摆脱去标识化与重新识别之间反复博弈的困境，并将重点放在医疗保健人工智能研究的知识民主化上。

# 2 结果

<escape>**数据**。 我们将数据凝缩（DC）应用于三个医疗保健数据集，以考察数据凝缩在共享这些数据集方面的潜力。第一个数据集记为 PhysioNet-2012 [14] 旨在</escape>

研究重症监护室（ICU）患者的死亡率预测，并在本研究中分析了 8,000 条公开可用的重症监护病房住院记录。每条 ICU 住院记录包含至多 42 个变量，这些变量在重症监护病房入院后的 48 小时内至少采集一次。研究目标是根据这些 ICU 住院记录预测院内死亡率。第二个数据集称为 MIMIC-III [15, 16],包含 21,156 条 ICU 住院记录，以及在进入 ICU 后最初 48 小时内采集的 17 个临床变量，任务是预测每次住院期间是否发生院内死亡。最后一个数据集名为 Coswara [17], 旨在利用各种呼吸音研究冠状病毒（新冠肺炎）的诊断。我们使用全部 1,368 个呼吸样本并提取声学特征，目标是诊断每个样本是否感染新冠肺炎。

在所有数据集中，我们将全部原始样本随机划分为训练集、验证集和测试集。在 PhysioNet-2012 中，我们将全部 8,000 条 ICU住院期间记录划分为上述三个集合，样本数分别为 5,120/1,280/1,600；而 MIMIC-III 中的 21,156 条住院记录则被划分为 14,698/3,222/3,236 个数据点。Coswara 的数据划分分别得到包含 987/175/206 个实例的三个集合。我们在补充说明1中详细介绍了这三个数据集的预处理和数据划分。

**压缩数据**。 对于每个数据集，我们采用第 4 节“方法”中所述的数据集凝缩方法，学习训练集的凝缩版本，同时保留验证集和测试集不变，以供后续评估。我们为每个数据集生成三个规模各异的凝缩数据集，以展示凝缩样本数量的影响。我们将 PhysioNet-2012 训练集中的 5,120 条 ICU住院期间记录压缩为三个集合，分别包含 20、40 和 80 个数据点。我们将 MIMIC-III 训练集（14,698 条 ICU住院期间记录）压缩为分别包含 400、800 和 1,200 个实例的集合，同时将 Coswara 的 987 个训练声学特征整合为分别包含 20、40 和 80 个样本的组。我们观察深度神经网络在原始数据集或凝缩数据集上训练时，在测试集上的表现。

**DNN 的队列**。 为考察凝缩数据对不同神经网络架构的普适性，我们收集了 11 种用于分析时间序列数据的常见 DNN，其中大多数已在计算机视觉领域展现出良好性能。该队列包括三种具有不同卷积核大小的多尺度时间卷积网络（MS-TCN） [18] ，分别记为 TCN-α、TCN-β 和 TCN-γ；两种具有不同注意力头数量的视觉变换器（ViT） [19], 简称为 ViT-α 和 ViT-β；两种具有不同注意力头数量和维度的变换器编码器 [20] ，分别称为 TRSF-α 和 TRSF-β；两种具有不同隐藏状态单元的长短期记忆网络（LSTM） [21] ，分别命名为 LSTM-α 和 LSTM-β；以及两种具有不同隐藏维度的循环神经网络（RNN） [22] ，分别称为 RNN-α 和 RNN-β。我们选用三种网络学习凝缩数据：TCN-α、ViT-α 和 LSTM-α；同时部署全部 11 种 DNN，以评估凝缩数据集所保留的深度学习效用。我们在补充说明2中概述了这些 DNN 的网络架构和训练设置。

**表 1**：在 PhysioNet-2012、MIMIC-III 和 Coswara 数据集上，基于原始训练集和三组压缩数据训练的深度神经网络的平均性能。我们对每个深度神经网络重复训练五次，并计算全部 11 个深度神经网络的均值和标准差（SD）。对于每个数据集，我们还报告其存储空间占用，单位为兆字节（MB）。

| **数据集** | **训练数据**（样本） | **大小**（MB） | **测试AUC**（标准差） |
|---|---|---|---|
| PhysioNet-2012 | 原始数据 (5120) | 88.13 | 0.858（0.011） |
| | 压缩数据（80） | 0.69 | 0.804（0.014） |
| | 压缩数据 (40) | 0.34 | 0.804 (0.016) |
| | 压缩数据 (20) | 0.17 | 0.803 (0.012) |
| MIMIC-III | 原始数据 (14698) | 322.95 | 0.840 (0.007) |
| | 压缩数据 (1200) | 13.18 | 0.756 (0.014) |
| | 压缩数据 (800) | 8.79 | 0.750 (0.014) |
| | 压缩数据 (400) | 4.39 | 0.741 (0.019) |
| Coswara | 原始数据 (987) | 46.27 | 0.737 (0.017) |
| | 压缩数据 (80) | 1.88 | 0.642 (0.016) |
| | 压缩数据（40） | 0.94 | 0.641（0.021） |
| | 压缩数据（20） | 0.47 | 0.632（0.030） |

**结局**。主要结局是考察 DC 凭借其独特的凝缩机制，在医疗保健数据民主化方面的适用性。我们主要从这一愿景的三个方面进行考察：1）DNN 在凝缩数据上训练后对真实数据的可用性；2）DC 在更紧凑的数据存储和更快的模型收敛率方面带来的益处；以及 3）凝缩数据的元数据属性。第一和第三个方面表明，凝缩的医疗保健数据可以不受限制地共享用于人工智能研究；第二个方面则有助于推动信息高效、便捷地流向数据民主化领域。

**凝缩数据在显著压缩数据量的同时，能够很好地保留深度学习效用**。我们期望凝缩数据能够充分保留原始数据的深度学习效用，使深度神经网络能够在真实样本上运行。为此，我们采用广泛使用的曲线下面积（AUC）作为主要评估指标。我们将基于原始训练集训练的全部11个深度神经网络的平均测试AUC作为基线。随后，我们在同一数据集的凝缩版本上训练相同队列的深度神经网络，并研究其测试AUC相对于基线的变化。为减小偏差，我们对每个深度神经网络重复训练五次，并报告平均AUC和标准差（SD）。

我们在表 1中总结了三个数据集上原始数据和凝缩数据的表现，并在图 1、图 2和图 3中分别展示了各数据集的结果。每个深度神经网络的数值结果见补充表 8。如表1 和图 1c 所示，PhysioNet-2012 的原始训练集包含 5,120 条 ICU 住院记录，可使全部 11 个深度神经网络达到以下

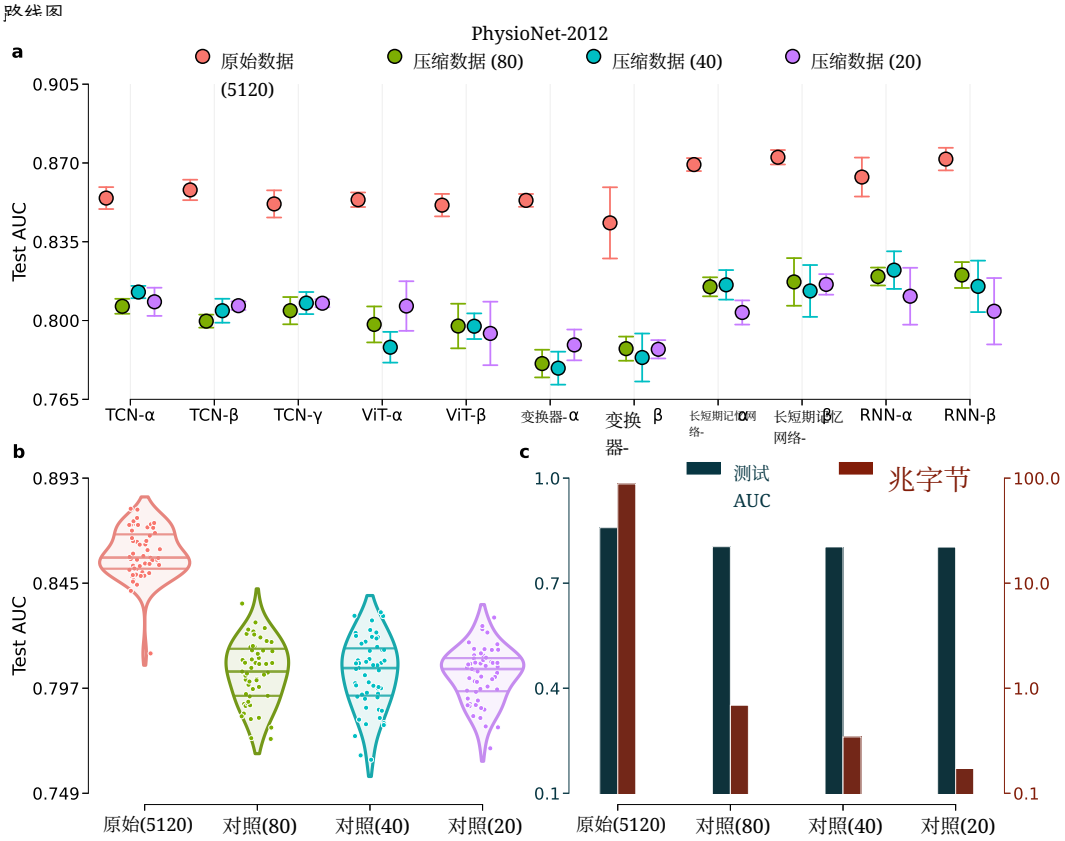



**图1：PhysioNet-2012 结果的可视化。**a，原始训练集与三个不同规模的凝缩数据在11个深度神经网络上的平均AUC值。我们在每个数据组后附加样本数量。b，四个数据组中所有深度神经网络测试AUC值的小提琴图。“Ori.”和“Con.”分别是“原始数据”和“压缩数据”的缩写。c，不同数据组的平均AUC与磁盘空间占用量的比较。请注意，我们采用对数刻度来显示空间使用情况。

预测院内死亡率的平均测试AUC $= 0.858$（标准差 0.011）。不过，该训练集占用 88.13 兆字节（MB）的磁盘空间。另一方面，使用包含 80、40 和 20 个样本的凝缩数据训练的深度神经网络，AUC值略有下降 $= 0。858 \rightarrow 0$分别为 0.804、0.804 和 0.803，降幅约为 5.5%。考虑到数据规模大幅压缩，这种准确率损失可以说是可以接受的，甚至令人印象深刻 $= 88。13 \rightarrow 0$分别为 0.69、0.34 和 0.17 兆字节，即使在最差的 80 个样本情形下，也减少了 87.44 兆字节（比原始数据少 99.2%）。此外，三个凝缩数据组的 AUC值相近且稳定，不受样本数量变化的影响；不过，包含 80 个和 40 个样本的数据集略优于包含 20 个样本的数据集。我们还可以从图 1a 中各个深度神经网络的表现对比看出一个有趣的现象：某个深度神经网络在原始数据上取得较高AUC时，同一架构在凝缩数据上的 AUC通常也较高，反之亦然。这一趋势可能与数据凝缩的机制相符，即匹配原始数据与凝缩数据之间的嵌入分布

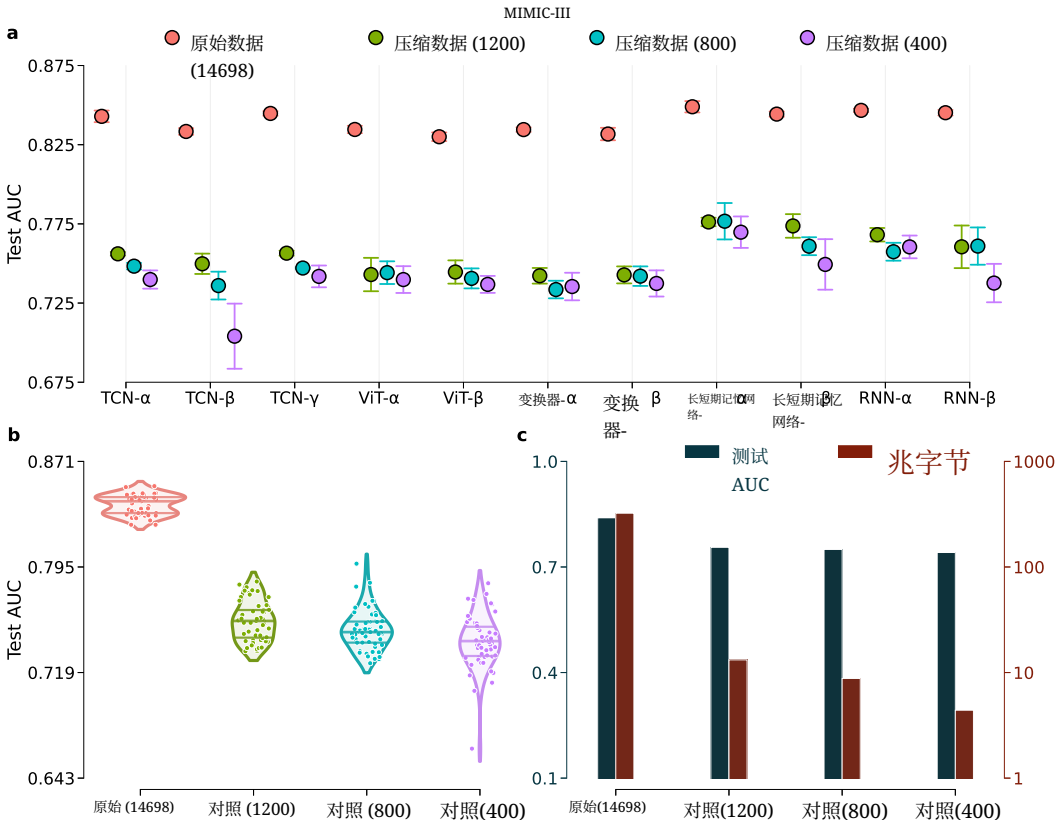



**图 2：MIMIC-III 结果的可视化。a，**原始数据与三个凝缩数据集在 11 个深度神经网络上的平均曲线下面积。**b，**四组数据中所有深度神经网络测试 AUC 的小提琴图。**c，**各数据组曲线下面积和数据大小的汇总。我们启用对数刻度来显示存储空间用量。

以及凝聚数据批次，这很可能会使相同深度神经网络结构产生相似的嵌入模式。此外，尽管除 TCN-α、ViT-α 和长短期记忆网络-α 之外的深度神经网络在数据凝缩过程中并未使用，但它们在凝缩数据上仍能取得与上述三种网络相当甚至更好的曲线下面积，这表明数据凝缩能够泛化到各种神经网络架构。

我们还可以从 MIMIC-III 结果中发现，压缩数据保留了深度学习效用。如表 1 和图2c 所示，在 MIMIC-III 原始训练集的 14,698 条重症监护病房住院记录上训练的 11 个深度神经网络，可实现 $= 0.840$ 的死亡率预测曲线下面积（标准差 0.007），而存储这些数据需要 322.95 MB 空间。使用包含 1200、800 和 400 个元素的凝缩数据集训练同一组深度神经网络，所得测试 AUC $= 0.840 \rightarrow 0.756$、0.750 和 0.741，分别低于原值。预测准确率下降 8.9%，但大幅降低的存储需求可有效弥补这一损失 $= 322.95 \rightarrow 13.18$、8.79 和 4.39 MB，分别减少了 95.9 %、97.2% 和 98.6% 的空间。各深度神经网络的曲线下面积结果

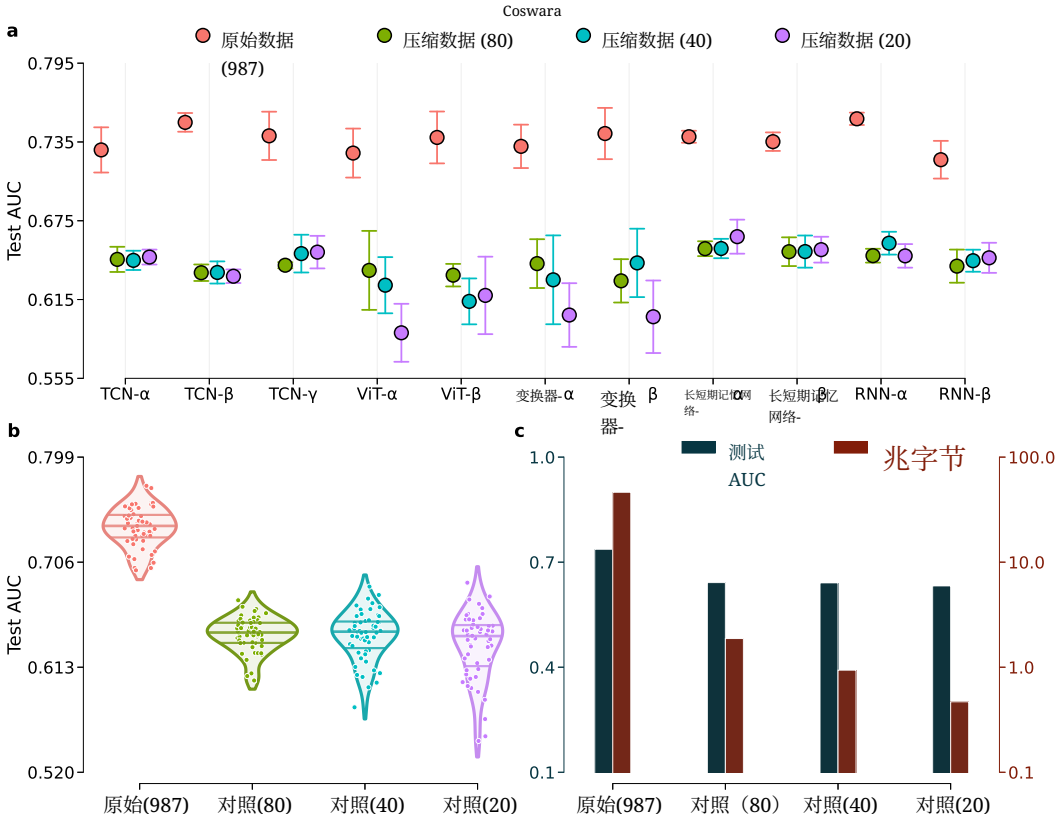



**图 3：Coswara 结果的可视化。a，**原始数据与三个凝缩数据集之间 11 个深度神经网络的平均 AUC值。**b，**四组数据中所有深度神经网络测试 AUC 的小提琴图。**c，**各数据组 AUC值和数据大小的汇总。我们采用对数刻度来显示空间占用情况。

图 2a 显示，在相同的深度神经网络架构上，原始数据与压缩数据的曲线下面积之间存在稳定的相关性，这一现象与 PhysioNet-2012 的结果一致。

在 Coswara 数据集上，我们将数据凝缩应用于一个截然不同的临床场景，即根据呼吸声的声学特征诊断新冠肺炎。考虑到所达到的深度神经网络准确率，这项任务比在 PhysioNet-2012 或 MIMIC-III 上进行死亡率预测更具挑战性。如表 1 和图 3c 所示，使用 46.27 MB 原始训练集训练的全部 11 个深度神经网络，平均测试曲线下面积仅为 $= 0.737$（标准差 0.017），效果不如 PhysioNet-2012 或 MIMIC-III。受原始数据曲线下面积的限制，包含 80、40 和 20 个实例的凝缩数据组使深度神经网络达到的曲线下面积分别为 $=0.737 \rightarrow 0.642$、0.641 和 0.632，曲线下面积下降约 9.6%。数据凝缩将磁盘占用量从原始数据的 46.21 MB 压缩至 1.88、0.94 和 0.47 MB（分别对应 80、40 和 20 个凝缩实例），节省了 96.0% 的存储空间。图 3a 展示了各神经网络架构的曲线下面积对比，整体趋势与 PhysioNet-2012 或 MIMIC-III 上的结果一致。

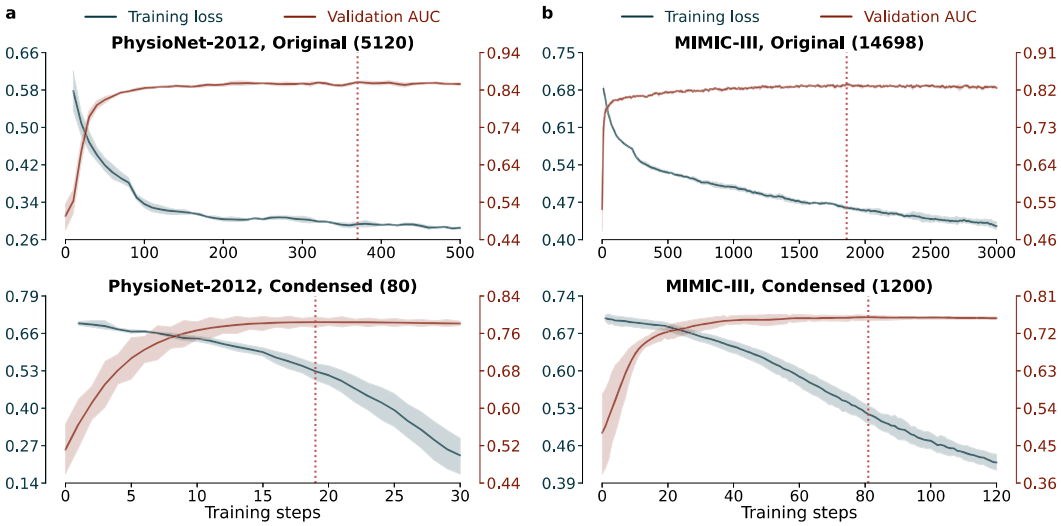



**图4**：a，在 PhysioNet-2012 原始数据和凝缩数据上训练的时间卷积网络-β 的学习曲线。b，在 MIMIC-III 原始数据和凝缩数据上训练的变换器-α 的学习曲线。红色虚线表示收敛步骤。

在三个数据集上，我们可以观察到，原始数据的深度学习效用（以 11 个深度神经网络的平均曲线下面积表示）在不同临床场景下都能在凝缩数据中得到充分保留，同时数据量大幅缩减，也便于数据存储与传输。值得注意的是，数据集凝缩对未见过的神经网络架构也具有普适性，这进一步凸显了数据凝缩在推动数据民主化方面的价值。

**表2**：在原始数据和压缩数据上训练的11个深度神经网络的收敛训练步数。“Ori.”和“Con.”分别表示“原始数据”和“压缩数据”，“平均值”（AVG）表示平均步数。我们在每个数据组后附加了样本数量。

| 数据集 | 列车 data | 收敛训练步数 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AVG | TCN-α | TCN-β | TCN-γ | ViT-α | ViT-β | 变换器-α | 变换器-β | 长短期记忆网络-α | 长短期记忆网络-β | RNN-α | 循环神经网络2-β |
| PhysioNet-2012 | 原始(5120) | 330 | 240 | 370 | 190 | 500 | 340 | 200 | 410 | 170 | 300 | 480 | 440 |
| | 对照（80） | 15 | 7 | 19 | 5 | 2 | 3 | 7 | 6 | 16 | 27 | 33 | 40 |
| MIMIC-III | 原始(14698) | 3083 | 4310 | 4400 | 3960 | 2030 | 1340 | 1860 | 2240 | 2500 | 3020 | 3870 | 4390 |
| | 对照（1200） | 273 | 356 | 311 | 361 | 32 | 358 | 81 | 361 | 186 | 323 | 333 | 304 |
| Coswara | 原始(987) | 217 | 148 | 384 | 144 | 132 | 108 | 164 | 104 | 204 | 324 | 324 | 360 |
| | 对照(80) | 26 | 3 | 40 | 25 | 25 | 25 | 40 | 32 | 33 | 24 | 5 | 35 |

**凝缩数据可以加速深度模型的收敛**。共享凝缩数据可以为人工智能研究带来另一项吸引人的优势：它能够加速深度神经网络的收敛，从而减少时间成本。为描述这一优势，我们绘制了在 PhysioNet-2012 的原始数据集和凝缩集上训练的 TCN-β 的学习曲线，如图4a 所示；学习

TRSF-α 在 MIMIC-III 上的曲线分别如图 4b 所示。我们可以从图 4a 观察到，在 PhysioNet-2012 上，TCN-β 在原始训练集上达到收敛（由红色虚线表示）大约需要 370 个训练步数。在 80 个凝缩样本上训练的同一网络在第 $= 370 \rightarrow 19$步收敛，节省了 93.1% 的训练时间。同样，如图4b 所示，TRSF-α 在原始 MIMIC-III 训练集上收敛平均需要 1,860 个训练步数，而在包含 1200 个实例的凝缩集上，所需的收敛步数显著减少至 $= 1{,}860 \rightarrow 81$，使训练会话提速 95.6%。我们在补充说明3中附上了其他网络的学习曲线。

我们在表格2中详细比较了各个深度神经网络原始数据与压缩数据的收敛步骤。可以发现，压缩数据加速学习是所有数据集和网络的共同特征。平均而言，在 PhysioNet-2012、MIMIC-III 和 Coswara 上，压缩数据相比原始数据分别可将训练速度提升 95.4%、91.1% 和 88.0%。因此，分享压缩数据能够以显著缩减的数据量和更快的模型学习速度，助力医疗保健人工智能研究。在研究数据量庞大的数据集时，这些特性尤为可取，因为完成模型训练可能十分耗时。

**凝缩数据有效隐藏了个体层面的知识**。除了良好保留的机器学习效用和压缩后的数据量之外，消除个体层面信息可能是数据凝缩最具吸引力的特性。正如第 4 节所述，数据凝缩通过随机化深度神经网络对原始数据和凝聚数据的两个随机批次进行嵌入，并在二者之间执行分布匹配。由于凝缩样本的数量远少于原始样本，这一机制形成了多对一对应。每个凝聚样本都从多个原始实例中获取知识，而凝缩集本质上是对原始数据集进行的宏观层面综合。从这一角度来看，共享凝聚数据集类似于共享元数据，例如临床变量的均值和标准差；这会隐藏个体层面信息，因此符合相关法规的要求。

为了说明这一元数据属性，我们在图中进行了可视化。5a 将从 MIMIC-III 及其凝缩集中随机选取的六个数据示例表示为二维热力图。原始实例的热力图呈现出高度规则且可能可识别的模式，构成患者身份被重新识别的潜在威胁。相比之下，凝聚样本的热力图呈现出完全混乱的模式，表明它们本质上是没有明确含义的随机信号。这些无序信号源自数据凝缩的学习过程；每个凝聚实例都会接收并融合多个原始样本的知识。因此，具有规则模式的原始数据被交织并随机化为混乱信号样本，可以说这能够掩盖医疗保健数据中的个体层面信息。

我们通过考察各种临床变量，进一步研究这一问题。图5b 描绘了根据

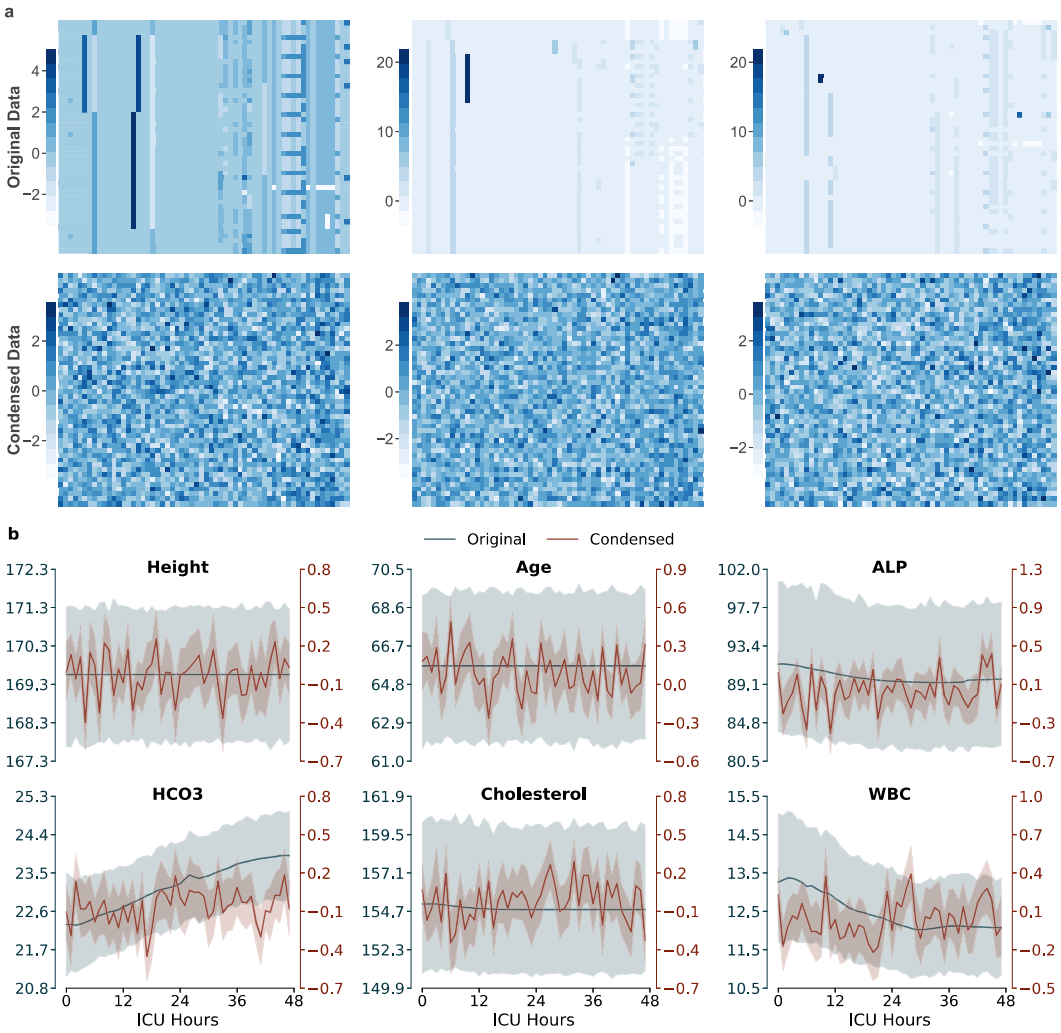



**图5**：原始数据和凝聚样本的可视化。a，从 MIMIC-III 中随机抽取的三个原始数据样本和三个凝聚样本的热力图。b，分别根据 80 个原始数据实例和 80 个凝聚实例计算的 PhysioNet-2012 中六个临床变量的 48小时趋势。

PhysioNet-2012 的 80 个原始样本和 80 个凝缩样本计算得出的 48小时趋势。这些变量包括：“身高”“年龄”“碱性磷酸酶”（“ALP”）、“血清碳酸氢盐”（“HCO3”）、“胆固醇”和“白细胞计数”（“WCB”）。作为一般性描述变量，原始样本中的“身高”和“年龄”在整个重症监护病房期间均呈现恒定值。然而，凝缩数据中的相同变量却显示出随时间显著波动的趋势，这在真实病例中不太可能出现，并且可能掩盖变量层面的信息。在其余原始变量和凝缩变量之间，我们也能观察到类似的偏差。例如，随着小时数增加，原始数据中的“ALP”和“WCP”总体值呈下降趋势，而其凝缩曲线则显示出带有随机噪声的不一致趋势。在原始数据呈上升趋势的“HCO3”上，也可以发现凝缩变量的波动。我们注意到，变量的不规则波动是凝缩数据的常见特征

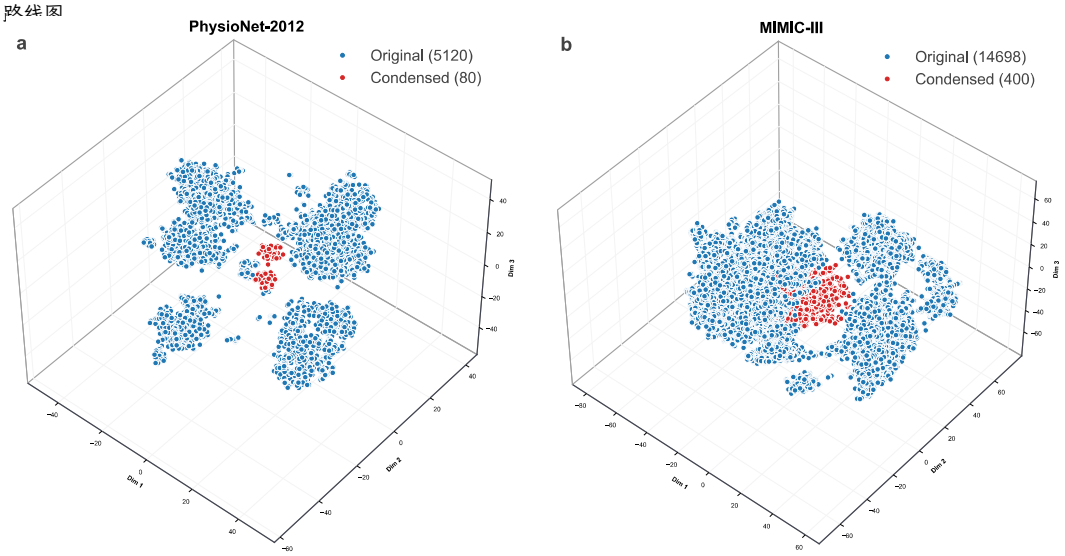



**图 6**：通过 t-SNE 将原始数据和压缩数据样本嵌入三维空间后的可视化结果。a，PhysioNet-2012 中 5,120 个原始数据样本和 80 个凝聚样本的三维分布。b，MIMIC-III 中 14,698 个原始数据样本和 400 个凝聚样本的三维分布。

样本，尽管原始变量中可能存在规律模式。这些观察结果与图 5a 的热图可视化结果一致，反映出临床变量层面无法恢复的个人知识隐匿。此外，每个压缩数据变量的范围都会自动偏移并重新缩放至约介于 $-1$ 和 1 之间的任意区间，从而进一步加密变量层面的知识。我们在补充说明3中补充了更多原始数据样本与凝缩样本的比较。

**压缩数据的分布与原始数据的分布截然不同**。个体层面和变量层面的隐匿意味着压缩数据可能存在不同的分布，从而有效隐藏真实的数据分布。因此，分享压缩数据，实际上就是分享一个独立于原始数据的不可见数据集。为解释这种分布偏移，我们采用 t分布随机邻域嵌入（t-SNE） [23] 来可视化图 6a 中 PhysioNet-2012 原始数据和压缩数据的三维分布，以及图6b 中 MIMIC-III 的三维分布。图 6a 中 PhysioNet-2012 的原始数据点大致构成四个相距较远的簇，中间留有一片空白区域。另一方面，压缩数据点呈现出偏离原始数据的分布。大多数压缩数据点位于原始簇之间的中央空白区域，我们只观察到两个明显的群组。这种分布可能是数据凝缩中的多对一对应所致。由于每个压缩数据点都从多个原始样本接收信息，其分布自然会移向中央区域，以尽可能平衡不同病例的知识。从这个角度来看，压缩数据实质上捕获了宏观

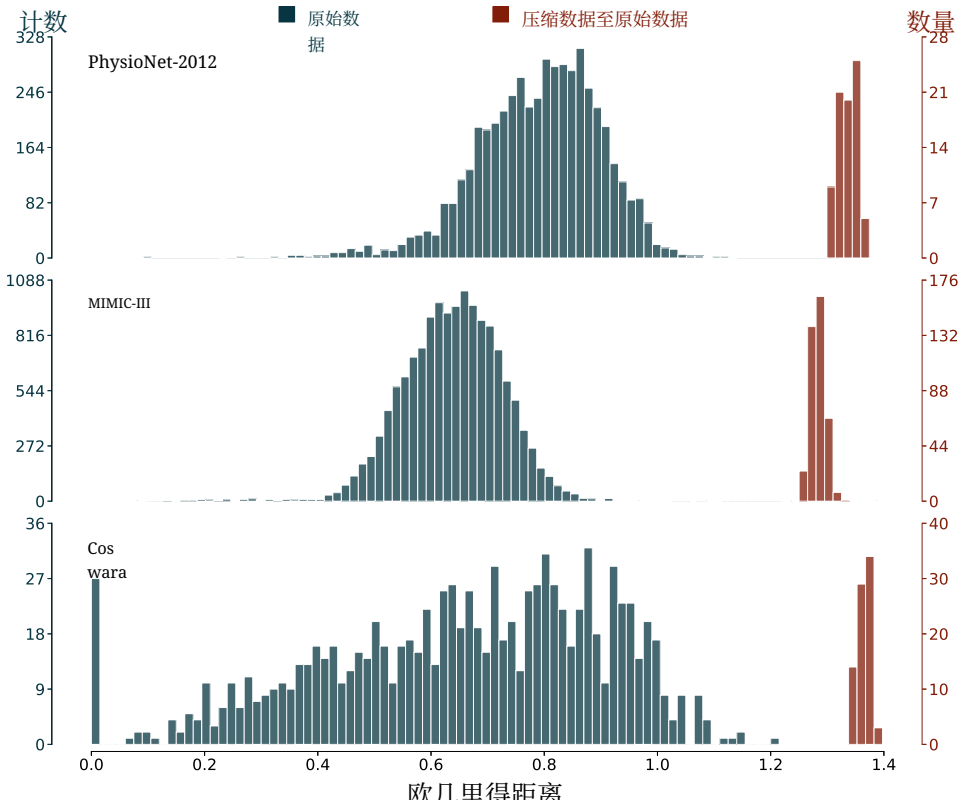



**图 7**：PhysioNet-2012、MIMIC-III 和 Coswara 的距离直方图。我们计算每个凝聚样本与每个原始实例之间的最小欧几里得距离，并将所有凝聚样本的最小距离绘制为直方图（红色的“压缩数据到原始数据”）。对于每个原始样本，我们计算其与其他所有实例之间的最小欧几里得距离，并将所有原始样本的最小距离绘制为直方图（蓝色的“原始数据”）。

原始数据集的统计特征，而共享这些特征就如同共享均值或标准差（STD）等统计数据一样。我们可以从图 6b 中看到，MIMIC-III 上原始数据与凝缩数据的分布也与这一分析结果一致。尽管数据密度显著增加，MIMIC-III 的原始数据仍由多个簇构成，其中心区域较为稀疏，而所有凝缩数据点均位于该区域内，这表明凝缩数据具有元数据的本质特征。

我们通过计算样本之间的欧氏距离，进一步考察分布偏移。我们在图 7 中分别绘制了原始训练集内部的欧氏距离直方图，以及凝缩实例与原始数据之间的距离直方图。在全部三个数据集中，原始数据内部的距离均显著小于凝缩数据与原始数据之间的距离。例如，大多数 MIMIC-III 原始样本之间的距离介于 0.4 和 0.9 之间，而凝缩实例与原始样本之间的距离则显著增大，达到 1.3 和 1.4 的范围。换言之，凝缩数据与原始数据之间的任意距离约为原始数据内部距离的两倍，这表明凝缩点位于远离原始数据分布的空间中

。在图 7中，从 PhysioNet-2012 和 Coswara 的直方图也可以观察到类似的分布差异，其中凝缩样本与原始样本之间的距离显著大于原始样本彼此之间的距离。在分布发生偏移的情况下，凝缩数据的民主化共享可以避免在共享具有真实分布的个体级数据时可能持续存在的患者重新识别风险。

# 3 讨论

人工智能研究处于医疗保健民主化的核心；然而，缺乏令人满意的数据民主化机制长期以来一直是个棘手问题，阻碍了临床人工智能的发展。本文探讨了数据集凝缩这一近期人工智能领域的成果，在推动医疗数据民主化、服务人工智能研究方面的潜力。针对三个医疗保健数据集 PhysioNet-2012、MIMIC-III 和 Coswara 的实验结果表明，数据凝缩（DC）能够保留原始临床记录的深度学习效用，同时以压缩后的数据量促进信息流通。凝缩数据还能够让深度神经网络（DNN）以快得多的速度收敛，显著加快深度网络的学习过程。本文还分析并讨论了凝缩数据的元数据属性：通过多对一的学习机制和不同的分布，它不可逆地隐藏了个体层面的信息。作为群体层面的知识，凝缩数据（如统计元数据）可以共享并实现民主化，从而推动医学人工智能研究。数据凝缩前景广阔。它或许能让我们摆脱去标识化与重新识别之间无休止的竞赛，转而专注于推动面向人工智能的知识民主化，以促进深度学习研究。临床数据集的持有者掌握着高度敏感但也极具价值的知识，可利用数据凝缩创建凝缩集并免费共享，同时避免个体被重新识别的风险。医疗保健研究人员也可以分享其私有数据集的凝缩版本，而这些数据集通常需要耗费大量时间才能获得访问许可，甚至根本无法获批。其他研究人员则可使用经过大幅压缩的凝聚数据集，训练精度较低但仍可用的深度神经网络，并加快其收敛速度。尽管数据凝缩仍处于起步阶段，但它展现出令人鼓舞的前景，有望促进科学研究的开放性并推动医疗保健领域的人工智能研究民主化。

另一方面，数据集凝缩存在各种局限性，我们从*没有免费午餐*的视角对这些局限性进行详细阐述。任何用于数据共享的去标识化方法都需要付出代价，通常会损害数据效用。传统的匿名化方法通过删除直接标识符和准标识符来实现去标识化，从而降低数据效用。删除更多数据字段通常会降低重新识别的可能性，但代价是可用性进一步受损。去标识化与数据效用之间的这种权衡，同样适用于共享由生成对抗网络（GAN）生成的合成数据 [24, 25]。合成数据对敏感信息提供更好的保护，代价是可用性下降；反之，若要保持良好的数据效用，则患者易受攻击的风险会增加。

重新识别。此外，由于合成数据源自真实的数据分布，因此包含原始样本的个体层面信息，可能会泄露个人身份信息。

没有免费午餐的观点同样适用于数据凝缩。凝缩数据本质上是原始数据集的宏观统计描述符，能够有效隐藏个体层面的敏感信息。然而，这类隐藏也会消除在充分挖掘临床记录全部价值时至关重要的细粒度尺度上的数据可用性。对于要求访问微观数据的分析任务而言，凝缩数据的作用可能十分有限，而这正是自由共享所需付出的主要代价之一。此外，便于存储和模型学习的压缩数据量，是以不可避免的信息损失为代价的，这一点体现为所学习深度神经网络中预测准确率的下降。数据民主化并不存在免费午餐；我们必须同时审视其利弊。尽管个体层面的效用已被消除，数据凝缩最具吸引力的优点在于，它允许面向人工智能的知识自由共享与流动，从而有可能促成民主化医疗保健系统。通过对那些通常难以直接访问的临床数据集进行凝缩，医疗保健人工智能研究能够从数据民主化中获益匪浅。

我们将数据集凝缩视为一种新兴且前景广阔的医疗保健 AI 数据共享途径。在数据凝缩出现之前，去标识化机制大致分为三类：1). 经典的基于删改的数据匿名化 [8], 2). 生成对抗网络生成的合成数据共享 [26],以及 3). 用于提供群体层面见解的远程处理系统，例如安全多方计算或同态加密[27]。数据凝缩开辟了一条尚未探索的途径，具有独特优势。匿名化数据或合成数据中的个体层面知识意味着重新识别患者的风险不可避免。此外，还面临着难以在安全性与效用之间找到令人满意的权衡这一困境。另一方面，数据凝缩消除了个体层面知识，带来了更大的共享灵活性，而且应用数据凝缩也更加直接，无需进行安全性与效用的权衡。数据凝缩在多个方面也不同于这些远程处理系统。构建此类远程系统可能成本高昂且耗时，包括硬件配置、日常维护、框架构建等。建成后，这些系统只能支持预先定义的分析任务，缺乏适应不同架构学习模型的灵活性。相比之下，数据凝缩是一种成本低且易于实施的技术，我们可以部署它，直接、快速地输出凝缩数据（所需学习时间见补充表 10）。凝缩数据中得以充分保留的深度学习效用，也使数据凝缩与同态加密等远程系统显著不同。

总而言之，没有灵丹妙药能够同时实现完美的去标识化和毫无瑕疵的数据效用。不过，数据集凝缩凭借其独特属性，有望成为一种分享和传播面向人工智能的知识、惠及医疗保健人工智能研究的技术。未来的工作可以致力于进一步提升通过凝缩数据训练出的深度神经网络的性能

同时保留其他期望的特性；随着数据凝缩在计算机视觉领域的应用，真正实现医疗保健民主化或许已不像过去看起来那么遥不可及。

# 4 方法

**问题定义**。 我们将医疗保健数据集的原始训练集定义为 $\mathcal{O} = \{(\mathbf{x}_i, y_i) \mid 1 \leq i \leq |\mathcal{O}|\}$ 其中，$\mathbf{x}_i \in \mathbb{R}^{T \times F}$是第 $i$ 个时间序列样本， $T$ 和 $F$ 分别表示时间维度和特征维度， $y_i$ 是类别标签， $|\mathcal{O}|$ 表示样本总数。我们旨在学习一个凝聚数据集 $\mathcal{C} = \{(\mathbf{c}_i, y_i) \mid 1 \leq i \leq |\mathcal{C}|\}$ 其中，$\mathbf{c}_i \in \mathbb{R}^{T \times F}$是一个凝聚样本，$|\mathcal{C}|$ 是凝聚样本的数量，且 $|\mathcal{C}| \ll |\mathcal{O}|$。我们期望 $\mathcal{C}$ 能够保留 $\mathcal{O}$ 中的深度学习效用，使得用 $\mathcal{C}$替换训练集 $\mathcal{O}$ 时，深度神经网络仍能在未见过的测试集上取得相当的性能，可表示为

$$\mathbb{E}_{\mathbf{x} \sim P_{\mathcal{D}}}[\mathrm{L}(\varphi_{\boldsymbol{\theta}^{\mathcal{O}}}(\mathbf{x}), y)] \simeq \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{D}}}[\mathrm{L}(\varphi_{\boldsymbol{\theta}^{\mathcal{C}}}(\mathbf{x}), y)]. \tag{1}$$

在式 1中，$\mathbf{x} \in \mathbb{R}^{T \times F}$ 是从原始数据分布中抽取的数据样本$P_{\mathcal{D}}$， $y$ 是类别标签，L 是损失函数，例如二元交叉熵（BCE）， $\varphi$ 是一个具有参数的深度神经网络 $\boldsymbol{\theta}$，$\varphi_{\boldsymbol{\theta}^{\mathcal{O}}}$ 而 $\varphi_{\boldsymbol{\theta}^{\mathcal{C}}}$ 分别表示在 $\mathcal{O}$ 和 $\mathcal{C}$上训练的深度神经网络。式 1 是面向时间序列医疗保健数据的数据凝缩目标；然而，直接优化该方程颇具挑战。我们遵循[13]来求解式 1，旨在匹配嵌入空间中的分布，因为在嵌入空间中计算效率更高。

**通过分布匹配进行数据集凝缩**。 优化公式的主要挑战 1 在于估计数据分布 $P_{\mathcal{D}}$。考虑到高维医疗保健样本，获得准确且合理的$P_{\mathcal{D}}$ 可能在计算上代价高昂，实际上也难以实现。为解决计算机视觉中的同一问题， [13] 的作者假设每个图像样本都可以借助一组参数化函数嵌入到低维空间中，然后汇集结果以估计原始数据的图像数据分布。在本研究中，我们对医疗保健数据遵循相同的前提，将每个样本 $\mathbf{x} \in \mathbb{R}^{T \times F}$ 嵌入到一组参数化函数中$\phi_{\boldsymbol{\mu}}$：$\mathbb{R}^{T \times F} \mapsto \mathbb{R}^{T' \times F'}$其中 $T' \ll T$， $F' \ll F$以及 $\boldsymbol{\mu}$表示函数 $\phi$的参数 $\phi_{\boldsymbol{\mu}}$。每个 $P_{\mathcal{D}}$ 揭示原始样本分布的一部分，而将它们合并可以提供合理的估计。

在这一背景下，我们可以使用最大均值差异（MMD）距离[28], 来匹配原始数据集与凝聚数据集之间的嵌入，其定义如下：

$$\sup_{\|\phi_{\boldsymbol{\mu}}\|_{\mathcal{H}} \leq 1} \left(\mathbb{E}[\phi_{\boldsymbol{\mu}}(\mathcal{O})] - \mathbb{E}[\phi_{\boldsymbol{\mu}}(\mathcal{C})]\right) \tag{2}$$

其中 $\mathcal{H}$ 表示再生核希尔伯特空间（RKHS）。我们采用经验估计来计算公式2 中的 MMD 距离，遵循 [13], 其可以

应写为

$$\mathbb{E}_{\boldsymbol{\mu}\sim P_{\boldsymbol{\mu}}}\|\frac{1}{|\mathcal{O}|}\sum_{i=1}^{|\mathcal{O}|}\phi_{\boldsymbol{\mu}}(\mathbf{x}_i)-\frac{1}{|\mathcal{C}|}\sum_{j=1}^{|\mathcal{C}|}\phi_{\boldsymbol{\mu}}(\mathbf{c}_j)\|^2 \tag{3}$$

其中， $P_{\boldsymbol{\mu}}$ 指的是 $\phi_{\boldsymbol{\mu}}$的参数分布。

嵌入网络有多种方法 $\phi_{\boldsymbol{\mu}}$。作者[13] 采用固定的网络架构，并在每个训练步骤中采样随机化参数。我们将这一方法扩展到一组网络架构，以实现更好的结构泛化和去标识化。对于每个训练步骤，我们首先从一个集合中随机选取一个网络 $\phi^i_{\boldsymbol{\mu}_i}$，该集合为 $\Phi=\{\phi^i_{\boldsymbol{\mu}_i}\,|\mathbb{1}\le i\le|\Phi|\}$ 其中 $\phi^i_{\boldsymbol{\mu}_i}$ 表示第 $i$个具有参数 $\boldsymbol{\mu}_i$ 的 DNN，而 $|\Phi|$表示集合大小。我们从参数分布 $P_{\boldsymbol{\mu}_i}$ 中为选定的网络生成随机化权重。这种双重采样策略将目标转化为一个优化问题，可描述为

$$\min_{\mathcal{C}}\mathbb{E}_{\phi^k_{\boldsymbol{\mu}_k}\sim\Phi,\boldsymbol{\mu}_k\sim P_{\boldsymbol{\mu}_k}}\|\frac{1}{|\mathcal{O}|}\sum_{i=1}^{|\mathcal{O}|}\phi^k_{\boldsymbol{\mu}_k}(\mathbf{x}_i)-\frac{1}{|\mathcal{C}|}\sum_{j=1}^{|\mathcal{C}|}\phi^k_{\boldsymbol{\mu}_k}(\mathbf{c}_j)\|^2. \tag{4}$$

公式 4 指出，我们希望学习一个凝聚数据集 $\mathcal{C}$，使其能够最大限度地缩小任意网络投影到嵌入空间中的原始数据与凝缩数据的两个分布之间的差异 $\phi^k_{\boldsymbol{\mu}_k}$，其中网络参数是随机采样的 $\boldsymbol{\mu}_k$。

随机化网络架构和参数可能是 DC 最有趣的机制之一。乍看之下，由随机化参数构成的随机化网络竟能实现对原始数据的适当嵌入，这似乎有些反直觉；我们也可能会疑惑，深度学习知识如何能够从原始数据集顺畅地流向凝聚数据集。然而，相关研究 [29]，表明，随机初始化的网络能够为各种计算机视觉任务提取高质量的特征描述符。这些网络还能够在嵌入空间中保持原始样本之间的距离相似性，例如，如 [30]所示，同一类别的两个样本的嵌入距离小于不同类别样本之间的距离。借助高质量且保持距离的表示，我们可以预期，随机初始化的深度神经网络能够在嵌入空间中构建原始数据的有意义分布，从而充分指导凝聚数据集的学习。应用这种随机化网络策略还可在医疗保健领域带来一个有利特性：从原始数据到凝缩数据的不可逆知识流。在不知道网络 $\phi^k_{\boldsymbol{\mu}_k}$以及经过 DC 训练后被遗忘的采样参数 $\boldsymbol{\mu}_k$ 的情况下，从凝聚数据集中追溯任何原始数据在计算上都是不可能的。

**学习凝聚数据集**。为了学习凝聚数据集 $\mathcal{C}$ 于式 4中，我们采用了如 [13]所示的迷你批次训练算法。我们首先随机初始化一个凝聚数据集 $\mathcal{C}$，并对其进行总计 $M$ 次迭代学习。每次迭代中，我们按如下方式进行随机采样：1）。采样

**算法 1** 学习凝缩医疗保健数据集

**输入：** 原始医疗保健数据集 $\mathcal{O}$

**要求：** 随机初始化的凝缩数据集 $\mathcal{C}$，包含 $|\mathcal{C}|$ 个 $S$类别的样本，深度网络集合 $\Phi$，参数分布 $P_{\boldsymbol{\mu}_k}$，用于任意网络，$\Phi$ 中的训练迭代次数 $M$、学习率 $\xi$

**对于** $m = 1 : M$ **执行**
  采样网络 $\phi^k_{\boldsymbol{\mu}_k} \sim \Phi$
  采样网络参数 $\boldsymbol{\mu}_k \sim P_{\boldsymbol{\mu}_k}$
  **对于** $s = 1 : S$ **执行**
    采样一个原始批次 $B_s^{\mathcal{O}} \sim \mathcal{O}$
    采样一个压缩数据批次 $B_s^{\mathcal{C}} \sim \mathcal{C}$
  **循环结束**
  计算损失 $\mathcal{L}$ 如公式所示5
  更新 $\mathcal{C} \leftarrow \mathcal{C} - \xi\nabla_{\mathcal{C}}\mathcal{L}$
**循环结束**

**输出：** 凝缩医疗保健数据集 $\mathcal{C}$

一个网络 $\phi^k_{\boldsymbol{\mu}_k} \sim \Phi$，2）。采样该网络的参数 $\boldsymbol{\mu}_k \sim P_{\boldsymbol{\mu}_k}$，3）。对于每个类别 $s$，采样一个原始数据批次 $B_s^{\mathcal{O}} \sim \mathcal{O}$ 和一个凝聚数据批次 $B_s^{\mathcal{C}} \sim \mathcal{C}$。计算所有类别中原始数据批次与凝聚数据批次之间的平均差异，作为本次迭代的损失 $\mathcal{L}$，可写作

$$\mathcal{L} = \sum_{s=1}^{S} \| \frac{1}{|B_s^{\mathcal{O}}|} \sum_{(\mathbf{x},y)\in B_s^{\mathcal{O}}} \phi^k_{\boldsymbol{\mu}_k}(\mathbf{x}) - \frac{1}{|B_s^{\mathcal{C}}|} \sum_{(\mathbf{c},y)\in B_s^{\mathcal{C}}} \phi^k_{\boldsymbol{\mu}_k}(\mathbf{c}) \|^2 \tag{5}$$

其中 $S$ 表示 $\mathcal{O}$中的类别总数。最后，我们使用随机梯度下降（SGD）算法，根据损失 $\mathcal{C}$ 更新凝聚数据集 $\mathcal{L}$，可表示为 $\mathcal{C} \leftarrow \mathcal{C} - \xi\nabla_{\mathcal{C}}\mathcal{L}$ 其中 $\xi$ 为学习率。我们在算法 1 中展示了训练过程。

公式 5展示了原始样本与凝聚样本之间的多对一对应关系。由于 $|\mathcal{C}| \ll |\mathcal{O}|$，训练迭代期间，凝缩批次 $B_s^{\mathcal{C}}$中的每个样本都会接触到同一类别的多个原始实例，这意味着它融合了多个案例中的知识。这种宏观层面理解有效地隐藏了原始数据中的个体层面信息，从而实现了“真正”的去标识化。

**超参数**。在全部三个数据集上，我们使用 Adam [31]优化器，并将初始学习率设为 0。001，以更新 $\mathcal{C}$ 作为算法的第9步操作1，并对每个凝聚数据集训练 24,000 次迭代。将 $B_s^{\mathcal{O}}$ 和$B_s^{\mathcal{C}}$ 的批大小统一设为 256。在 PhysioNet-2012 和 Coswara 上，我们生成包含不同数量样本的三个凝聚数据集，即 $|\mathcal{C}| = 20$、40 和 80；而 MIMIC-III 的数据集规模 $|\mathcal{C}|$ 分别设为 400、800 和 1,200。我们从头开始随机初始化每个凝聚

数据集，并将正标签和负标签的比例设为 50% 和 50%。在本研究中，我们将集合 $\Phi$设定为包含三种网络架构，即 TCN-α、ViT-α 和 LSTM-α。我们还探究了凝聚样本中时间维度变化的影响（见补充表 9）。

# 参考文献

[1] Topol, E.J.：高性能医疗：人类智能与人工智能的收敛。《自然医学》**25**（1），44–56（2019）[2] Zhang, A., Xing, L., Zou, J., Wu, J.C.：推动医疗保健领域的机器学习从开发走向部署，并从模型转向数据。《自然·生物医学工程》，1–16（2022）[3]全球医疗保健人工智能市场规模将于2032年达到2807.7亿美元，复合年增长率为30.6%。SPHERICAL INSIGHTS LLP。访问日期：2023年8月15日（2023）。
https://www.globenewswire.com/en/news-release/2023/07/04/2698952/0/en/Global-Artificial-Intelligence-In-Healthcare-Market-Size-To-Worth-USD-280-77-

比利

html[4]医疗保健的民主化。斯坦福医学。访问日期：2023年6月15日（2018）。
https://med.stanford.edu/content/dam/sm/school/documents/Health-Trends-Report/Stanford-Medicine-Health-Trends-Report-2018.pdf[5] Wang, Y., Blobel, B., Yang, B.：通过数据民主化加强健康数据共享。《个性化医学杂志》**12**（9），1380（2022）[6] Lewis, K., Pham, C., Batarseh, F.A.：医疗保健领域的数据开放与民主化：医院排名方法评估。载于：《数据民主》，第109–126页。爱思唯尔，???（2020）[7] Rajkomar, A., Oren, E., Chen,K., Dai, A.M.,Hajaj,N.,Hardt,M.,Liu,P.J., Liu, X., Marcus, J.,Sun, M., 等：基于电子健康记录实现可扩展且准确的深度学习。《NPJ数字医学》**1**（1），1–10（2018）[8] El Emam, K., Arbuckle, L.：《健康数据匿名化：案例研究与入门方法》，第1版。O'Reilly Media, Inc.，塞瓦斯托波尔（2013）[9] Narayanan, A., Shmatikov, V.：“个人身份信息”的种种迷思与谬误。《ACM通讯》**53**（6），24–26（2010）

[10] Sweeney, L., Yoo, J.S.：根据处方数据对共享的韩国居民登记号码进行去匿名化。《Technology Science》（2015）[11] De Montjoye, Y.-A., Radaelli, L., Singh, V.K., Pentland, A.S.：购物中心中的独特性：信用卡元数据的可重新识别性。《Science》**347**（6221），536–539（2015）[12] Pandey, M., Fernandez, M., Gentile, F., Isayev, O., Tropsha, A., Stern, A.C., Cherkasov, A.：GPU计算与深度学习在药物发现中的变革性作用。《Nature Machine Intelligence》**4**（3），211–221（2022）[13] Zhao, B., Bilen, H.：通过分布匹配实现数据集凝缩。载于：IEEE/CVF冬季计算机视觉应用会议论文集，第6514–6523页（2023）[14] Silva, I., Moody, G., Scott, D.J., Celi, L.A., Mark, R.G.：预测重症监护病房患者的院内死亡率：PhysioNet/Computing in Cardiology 2012挑战赛。载于：2012年Computing in Cardiology，第245–248页（2012）。IEEE[15]Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.-w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., Mark, R.G.：MIMIC-III：一个可免费访问的重症监护数据库。《Scientific Data》**3**（1），1–9（2016）[16] Johnson, A., Pollard, T., Mark, R.：Mimic-iii 临床数据库（版本1.4）。PhysioNet**10**，2–26（2016）[17] Sharma, N., Krishnan, P., Kumar, R., Ramoji, S., Chetupalli, S.,Nirmala, R., Kumar Ghosh, P., Ganapathy, S.：Coswara——用于新冠肺炎诊断的呼吸、咳嗽和语音数据集。载于：国际语音通信协会年会论文集，INTERSPEECH，第2020卷，第4811–4815页（2020）。国际语音通信协会[18] Martinez, B., Ma, P., Petridis, S., Pantic, M.：使用时间卷积网络进行唇读。载于：ICASSP 2020-2020 IEEE 国际声学、语音与信号处理会议（ICASSP），第6319–6323页（2020）。IEEE[19] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.：一张图像胜过16×16个词：用于大规模图像识别的变换器。载于：国际学习表征会议（2021）。
https://openreview.net/forum?id=YicbFdNTTy

[20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. 神经信息处理系统进展 **30** (2017)[21] Hochreiter, S., Schmidhuber, J.: 长短期记忆. 神经计算 **9**(8), 1735–1780 (1997)[22] Hopfield, J.J.: 具有涌现集体计算能力的神经网络和物理系统. 美国国家科学院院刊 **79**(8), 2554–2558 (1982)[23] Van der Maaten, L., Hinton, G.: 使用 t-SNE 可视化数据. 机器学习研究杂志 **9**(11) (2008)[24] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: 生成对抗网络. ACM通讯 **63**(11), 139–144 (2020)[25] Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: 使用条件生成对抗网络对表格数据建模. 神经信息处理系统进展 **32** (2019)[26] Stadler, T., Oprisanu, B., Troncoso, C.: 合成数据——匿名化的土拨鼠日. 载于：第 31 届 USENIX 安全研讨会（USENIX Security 22），第 1451–1468 页（2022）[27] Froelicher, D., Troncoso-Pastoriza, J.R., Raisaro, J.L., Cuendet, M.A., Sousa, J.S., Cho, H., Berger, B., Fellay, J., Hubaux, J.-P.: 利用多方同态加密实现精准医学中真正保护隐私的联邦分析. 自然通讯 **12**(1), 1–10 (2021)[28] Gretton, A., Borgwardt, K.M., Rasch, M.J., Sch¨olkopf, B., Smola, A.: 核双样本检验. 机器学习研究杂志 **13**(1), 723–773 (2012)[29] Saxe, A.M., Koh, P.W., Chen, Z., Bhand, M., Suresh, B., Ng, A.Y.: 关于随机权重与无监督特征学习. 载于：ICML（2011）[30]Giryes, R., Sapiro, G., Bronstein, A.M.: 具有随机高斯权重的深度神经网络：一种通用分类策略？IEEE信号处理汇刊 **64**(13), 3444–3457 (2016)[31] Kingma, D.P., Ba, J.: Adam：一种随机优化方法. arXiv 预印本 arXiv:1412.6980 (2014)

arXiv:2305.03711v2 [cs.LG] 8 Jan 2024

# 补充信息-医疗记录凝缩：迈向医疗保健数据民主化的路线图

Yujiang Wang[1,4],Anshul Thakur[1*]、Mingzhi Dong[2],Pingchuan Ma[3], Stavros Petridis[3], Li Shang[2*]、Tingting Zhu[1] 和David A. Clifton[1,4]

[1*]工程科学系，牛津大学，英国。[2*]计算机科学学院，复旦大学，中国。[3]计算机系，伦敦帝国学院，英国。[4]牛津苏州先进研究中心，中国。

*通讯作者。电子邮箱：anshul.thakur@eng.ox.ac.uk；lishang@fudan.edu.cn；贡献作者：yujiang.wang@oxford-oscar.cn；mingzhidong@gmail.com；pingchuan.ma16@imperial.ac.uk；stavros.petridis04@imperial.ac.uk；tingting.zhu@eng.ox.ac.uk；david.clifton@eng.ox.ac.uk；

## 补充说明1 - 数据集

### 预处理

**PhysioNet-2012**。 PhysioNet-2012 数据集 [2]包含 8,000 条重症监护病房住院记录，可公开用于研究院内死亡率相关问题。每次住院记录包含 42 个变量，每个变量在进入重症监护病房后的前 48 小时内至少记录一次。其中 37 个变量是时间序列，在这 48 小时内有多次观测；另有 6 个变量是重症监护病房入院时采集的一般描述信息。“体重”既是一般描述信息（入院时测量），也是时间序列（每小时记录）。我们将每个时间序列变量转换为一个 48 维向量，每个值表示对应小时段的状态。我们将每个通用...

## 2 补充信息

**补充表1**：PhysioNet-2012 中的临床变量。我们为每个变量创建一个或多个48维向量。我们使用独热编码表示“性别”和“ICU类型”，并用“→”指定类别值。“计数（向量）”指预处理后该变量对应的向量数量。每次 ICU住院期间都会生成一个大小为 $48 \times 47$ 的矩阵，并以“院内死亡”作为结局标签。

| 临床变量 | | Type | 计数（向量） |
|---|---|---|---|
| · 白蛋白<br>· 血尿素氮<br>· 格拉斯哥昏迷评分<br>· 血清钾<br>· 血清钠<br>· 血清肌酐<br>· 胆红素<br>· 机械通气呼吸<br>· 丙氨酸转氨酶<br>· 有创收缩压血压<br>· 无创舒张动脉血压<br>· 血清碳酸氢盐<br>· 无创收缩压血压<br>· 动脉血pH<br>· 血小板<br>· 肌钙蛋白I<br>· 心率<br>· 血红蛋白氧饱和度<br>· 体重 | · 碱性磷酸酶<br>· 胆固醇<br>· 血清葡萄糖<br>· 乳酸<br>· 尿量<br>· 体温<br>· 吸入氧分数<br>· 动脉分压CO2<br>· 天冬氨酸转氨酶<br>· 有创舒张压血压<br>· 无创平均动脉压血压<br>· 血细胞比容<br>· 有创平均动脉压<br>· 血清镁<br>· 呼吸频率<br>· 肌钙蛋白T<br>· 白细胞计数<br>· 动脉血氧分压 | Time系列 | 37(37) |
| · 年龄<br>· 身高<br>· ICU类型→外科重症监护病房<br>· ICU类型→心脏外科手术复苏单元 | · 性别→女性<br>· 性别→男性<br>· ICU类型→内科重症监护病房<br>· ICU类型→冠心病监护病房Unit | 通用描述项 | 4(8) |
| · SAPS-I评分 | · SOFA评分 | 评分 | 2(2) |
| · 院内死亡 | | 结局 | 1(1) |

将描述符转换为具有恒定时间值的48维向量（不包括“记录ID”）。值得注意的是，对于两个分类变量“性别”和“ICU类型”，我们采用独热编码策略，分别使用两个48维二进制向量描述前者，使用四个48维二进制向量描述后者。此外，还使用两个额外的评分变量，即SAPS-I评分 [3] 和SOFA评分 [4], ，并分别将其转换为两个48通道向量。我们将一次ICU住院期间的所有向量堆叠成一个形状为 $48 \times 47$的代表性矩阵，其中48对应小时数，47代表特征数。结局是一个二元变量“院内死亡”，用于指示患者是存活（阴性）还是在医院内死亡（阳性），其中阴性样本总数为6,878个，阳性样本总数为1,122个。补充表 1 总结了PhysioNet-2012中的临床变量。

**补充表2**：MIMIC-III 中的临床变量。对于每个变量，我们按照[1]创建一个或多个包含插补值的 48维向量。我们使用独热编码表示每个分类变量，并使用“→”详细说明分类值。表示插补情况的掩码向量被视为输入。“GCS”指“格拉斯哥昏迷评分”，“言语反应”指语言反应。每次 ICU住院期间都会生成一个以“院内死亡”为结局的 $48 \times 60$ 矩阵。

| 临床变量 | | Type | 计数（向量。） |
|---|---|---|---|
| · 舒张压<br>吸入氧分数<br>· 葡萄糖<br>· 心率<br>· 身高<br>· 平均血压 | · 血氧饱和度<br>· 呼吸频率<br>· 收缩压<br>· 体温<br>· 体重<br>· pH | 继续-续 | 12(12) |
| · 毛细血管再充盈率 → 0<br>· GCS睁眼反应 → 对言语有反应<br>· GCS睁眼反应 → 对疼痛有反应<br>· GCS睁眼反应 → 自动睁眼<br>· GCS言语反应 → 无法理解的声音<br>· GCS言语反应 →否反应<br>· GCS运动反应 → 异常伸展<br>· GCS运动反应 → 无反应<br>· GCS运动反应 → 疼痛定位<br>· GCS总分 → 3<br>· GCS总分 → 4<br>· GCS总分 → 5<br>· GCS总分 → 6<br>· GCS总分 → 7<br>· GCS总分 → 8<br>· GCS总分 → 9 | · 毛细血管再充盈率 → 1<br>· GCS睁眼反应 → 无<br>· GCS言语反应 → 困惑<br>· GCS睁眼反应 → 无反应<br>· GCS言语反应 →不恰当的言语<br>· GCS言语反应 → 定向正常<br>· GCS运动反应 →异常屈曲<br>· GCS运动反应 → 屈曲撤退<br>· GCS运动反应 →遵命令<br>· GCS总分 → 10<br>· GCS总分 → 11<br>· GCS总分 → 12<br>· GCS总分 → 13<br>· GCS总分 → 14<br>· GCS总分 → 15 | 分类-别 | 5(31) |
| · 掩码 → 毛细血管再充盈率<br>· 掩码 → 吸入氧分数<br>· 掩码 → GCS运动反应<br>· 掩码 → GCS总分<br>· 掩码 → 心率<br>· 掩码 → 平均血压<br>· 掩码 → 呼吸频率<br>· 掩码 → 体温<br>· 掩码 → pH | · 掩码 → GCS睁眼反应<br>· 掩码 → 舒张压<br>· 掩码 → GCS言语反应<br>· 掩码 → 葡萄糖<br>· 掩码 → 身高<br>· 掩码 → 血氧饱和度<br>· 掩码 → 收缩压<br>· 掩码 → 体重 | Mask | 17(17) |
| · 院内死亡 | | 结局 | 1(1) |

**MIMIC-III**。 我们通常按照 [1] 对 MIMIC-III 数据集进行预处理[5, 6]。具体而言，本研究使用的 MIMIC-III 数据集包含 21,156 条重症监护病房住院记录，涉及从“CHARTEVENTS”和“LABEVENTS”表中选取的 17 个临床变量（有关这些表的详情，请参阅 [5] ）。这些变量记录于重症监护病房入院后的前 48 小时内，可分为 5 个分类变量和 12 个连续变量。每个连续变量都会转换为一个 48维数值向量，而每个分类变量则编码为

独热向量。所有缺失值均按照 [1], 进行插补，并且我们将 17 个二进制掩码向量作为额外输入信息，用以标示变量值是真实值还是插补值。预处理后，我们可以生成 60 个 48维向量来表示每条 ICU住院期间的数据，即一个 $48 \times 60$ 矩阵。我们将“院内死亡率”作为结局标签，其中 18,357 例为阴性（“存活患者”），2,799 例为阳性（“在院内死亡”）。补充表2展示了 MIMIC-III 中的所有临床变量。

**Coswara**。 Coswara数据集 [7]包含从接受聚合酶链式反应（PCR）检测并获得检测结果的参与者中采集的1,368段呼吸声样本，其结局为诊断冠状病毒-19（新冠肺炎）。原始声音的时长约为3至8秒，采样频率为48 kHz。为统一时间长度，我们从每个样本中截取声强最高的3秒片段。随后，我们采用短时傅里叶变换（STFT），使用重叠率为50%的100 ms FFT窗口，提取64个对数梅尔频带的梅尔频谱图。提取后，每段呼吸声均得到一个 $94 \times 64$ 表示矩阵，其中94和64分别代表时间维度和声学维度。该数据集包含948个PCR检测结果为阳性的样本和420个PCR检测结果为阴性的样本。

## 数据集划分

**补充表3**：三个数据集中训练集、验证集和测试集的样本数量（阳性/阴性样本数量）。

| 数据集 | 训练集 | 验证集 | Test |
|---|---|---|---|
| PhysioNet-2012 | 5,120 (728/4,392) | 1,280 (167/1,113) | 1,600 (227/1,373) |
| MIMIC-III | 14,698（1,989/12,709） | 3,222 (436/2,786) | 3,236 (374/2,862) |
| Coswara | 987 (683/304) | 175 (121/54) | 206（144/62） |

在三个数据集中，我们将数据划分为训练集、验证集和测试集。训练集用于学习凝缩数据和训练深度神经网络，而验证集和测试集则用于评估深度神经网络的性能。补充表格3概述了本文所采用的数据划分方式。

# Su补充说明 2——深度神经网络 ks

## 网络架构

我们选取了 11 种常见的时间序列数据深度神经网络，以评估原始数据和压缩数据训练集的深度学习效用。这 11 种深度神经网络来自五大神经网络族：多尺度时间卷积网络（MS-TCN） [8], 变换器（TRSF） [9], 视觉 Transformer（ViT） [10], 长短期记忆网络（LSTM） [11],以及循环神经网络（RNN） [12]。

**多尺度时间卷积网络**。多尺度时间卷积网络（MS-TCN） [8]是时间卷积网络（TCN） [14] 的一种变体，并且在唇读等多种计算机视觉任务中取得了最先进的性能。TCN 利用一维因果卷积分析时间序列数据中的时间信息，而 MS-TCN 通过引入具有不同卷积核尺寸的多尺度卷积来改进其网络架构，从而更好地反映不同尺度的时间信息。本研究采用了三个 MS-TCN，分别记为 TCN-α、TCN-β和 TCN-γ。所有网络均由两个时间卷积（TC）层、一个时间池化层和一个全连接（FC）层组成。TCN-α 中的每个 TC 层均包含三个分支，其卷积核尺寸分别为 3、5 和 7，且每个分支的隐藏神经元数量均设为 64。对于 TCN-β，每个 TC 层仅包含一个卷积核尺寸为 3、隐藏维度为 64 的卷积分支。TCN-γ 中的每个 TC 层均采用双分支结构，隐藏神经元数量为 64，两个分支的卷积核尺寸分别为 3 和 5。补充表4 展示了 TCN-α、TCN-β和 TCN-γ的网络架构。

**Transformer和视觉Transformer**。Transformer（TRSF） [9]最初是为解决自然语言处理（NLP）问题而提出的。自注意力机制使 Transformer能够捕捉输入信号的全局依赖关系，因此其在众多领域得到广泛应用，并取得了令人瞩目的结果。原始Transformer [9] 由一个编码器和一个解码器组成。在本研究中，我们采用Transformer编码器提取时间序列数据的嵌入，并舍弃解码器。我们主要考察了两种架构不同的Transformer编码器：TRSF-α 和TRSF-β。这两种网络均由两个多头自注意力（MSA）层组成，其后接一个时间池化层和一个由两个全连接层构成的多层感知机（MLP）头。为简化起见，我们忽略位置嵌入。在TRSF-α中，我们将每个MSA层设为包含16个自注意力头，每个头包含64个注意力通道。MLP中的隐藏神经元数量设为64。TRSF-β 中的每个MSA层包含4个注意力头，每个头有256个通道，其MLP的隐藏维度为128。

视觉Transformer（视觉变换器，ViTs） [10] 是专为图像识别设计的一种Transformer变体。通过将每幅图像划分为图像块，并进行线性投影。

**补充表4**：TCN-α、TCN-β和 TCN-γ的架构。核大小标注在每个一维卷积操作之后，每个随机失活后也标明了随机失活率。我们采用参数化修正线性单元 [13] 作为主要激活函数。“RC$i\rightarrow$”表示第 $i$个残差连接的起始处，“RC$i\leftarrow$”表示其结束处。“拼接”表示按通道拼接，“Chmop1D”是因果卷积的实现方式[14]。

| 层 | 时间卷积网络-α 分支1 | 时间卷积网络-α 分支2 | 时间卷积网络-α 分支3 | 时间卷积网络-β | 时间卷积网络-γ 分支1 | 时间卷积网络-γ 分支2 |
|---|---|---|---|---|---|---|
| | | RC1→ | | RC1→ | RC1→ | |
| 1 | 一维卷积(3)<br>批归一化<br>一维裁剪<br>参数化修正线性单元 | 一维卷积(5)<br>批归一化<br>一维裁剪<br>参数化修正线性单元 | 一维卷积(7)<br>批归一化<br>一维裁剪<br>参数化修正线性单元 | 一维卷积(3)<br>批归一化<br>一维裁剪<br>参数化修正线性单元 | 一维卷积(3)<br>批归一化<br>一维裁剪<br>参数化修正线性单元 | 一维卷积(5)<br>批归一化<br>一维裁剪<br>参数化修正线性单元 |
| | 拼接 & 随机失活（0.5） | | | 随机失活（0.5） | 拼接 & 随机失活（0.5） | |
| | 一维卷积(3)<br>批归一化<br>一维裁剪<br>参数化修正线性单元 | 一维卷积(5)<br>批归一化<br>一维裁剪<br>参数化修正线性单元 | 一维卷积(7)<br>批归一化<br>一维裁剪<br>参数化修正线性单元 | 一维卷积(3)<br>批归一化<br>一维裁剪<br>参数化修正线性单元 | 一维卷积(3)<br>批归一化<br>一维裁剪<br>参数化修正线性单元 | 一维卷积(5)<br>批归一化<br>一维裁剪<br>参数化修正线性单元 |
| | 拼接 & 随机失活（0.5） | | | 随机失活（0.5） | 拼接 & 随机失活（0.5） | |
| | 一维卷积(1)<br>RC1←<br>参数化修正线性单元 | | | 一维卷积(1)<br>RC1←<br>参数化修正线性单元 | 一维卷积(1)<br>RC1←<br>参数化修正线性单元 | |
| | | RC2→ | | RC2→ | RC2→ | |
| 2 | 一维卷积(3)<br>批归一化<br>一维裁剪<br>参数化修正线性单元 | 一维卷积(5)<br>批归一化<br>一维裁剪<br>参数化修正线性单元 | 一维卷积(7)<br>批归一化<br>一维裁剪<br>参数化修正线性单元 | 一维卷积(3)<br>批归一化<br>一维裁剪<br>参数化修正线性单元 | 一维卷积(3)<br>批归一化<br>一维裁剪<br>参数化修正线性单元 | 一维卷积(5)<br>批归一化<br>一维裁剪<br>参数化修正线性单元 |
| | 拼接 & 随机失活（0.5） | | | 随机失活（0.5） | 拼接 & 随机失活（0.5） | |
| | 一维卷积(3)<br>批归一化<br>一维裁剪<br>参数化修正线性单元 | 一维卷积(5)<br>批归一化<br>一维裁剪<br>参数化修正线性单元 | 一维卷积(7)<br>批归一化<br>一维裁剪<br>参数化修正线性单元 | 一维卷积(3)<br>批归一化<br>一维裁剪<br>参数化修正线性单元 | 一维卷积(3)<br>批归一化<br>一维裁剪<br>参数化修正线性单元 | 一维卷积(5)<br>批归一化<br>一维裁剪<br>参数化修正线性单元 |
| | 拼接 & 随机失活（0.5） | | | 随机失活（0.5） | 拼接 & 随机失活（0.5） | |
| | 一维卷积(1)<br>RC2←<br>参数化修正线性单元 | | | 一维卷积(1)<br>RC2←<br>参数化修正线性单元 | 一维卷积(1)<br>RC2←<br>参数化修正线性单元 | |
| 3 | 池化<br>线性层<br>Sigmoid | | | 汇聚<br>线性层<br>Sigmoid | 池化<br>线性层<br>Sigmoid | |

将图像块转换为序列嵌入后，视觉变换器能够以最先进的性能分析图像。视觉变换器在获取序列表示方面也不同于 Transformer。输入前会添加一个可学习令牌，其在输出端对应的部分会馈入多层感知机。本研究中，我们舍弃线性投影组件，直接将时间序列数据馈入视觉变换器编码器。我们考察了两种架构不同的视觉变换器：ViT-α 和ViT-β。两种视觉变换器均包含四个多头自注意力层，并且

补充表 5：变换器-α、变换器-β、视觉变换器-α和视觉变换器-β的架构。“多头自注意力”指多头自注意力层，头数和注意力通道紧随其后。“[令牌]←”指向输入中追加一个学习令牌，而“[令牌]→”表示在输出端获取该令牌。“前馈层”由两个全连接层组成。我们将 GELU [15] 作为主要激活函数。

| 层 | 变换器-α | 变换器-β | 视觉变换器-α | 视觉变换器-β |
|---|---|---|---|---|
| 1 | RC1a→<br>层归一化<br>多头自注意力(16,64)<br>RC1a←<br>RC1b→<br>层归一化<br>F.F.<br>RC1b← | RC1a→<br>层归一化<br>多头自注意力(4,256)<br>RC1a←<br>RC1b→<br>层归一化<br>F.F.<br>RC1b← | [token]←<br>RC1a→<br>层归一化<br>多头自注意力(16,64)<br>RC1a←<br>RC1b→<br>层归一化<br>F.F.<br>RC1b← | [token]←<br>RC1a→<br>层归一化<br>多头自注意力(4,256)<br>RC1a←<br>RC1b→<br>层归一化<br>F.F.<br>RC1b← |
| 2 | RC2a→<br>层归一化<br>多头自注意力(16,64)<br>RC2a←<br>RC2b→<br>层归一化<br>F.F.<br>RC2b← | RC2a→<br>层归一化<br>多头自注意力(4,256)<br>RC2a←<br>RC2b→<br>层归一化<br>F.F.<br>RC2b← | RC2a→<br>层归一化<br>多头自注意力(16,64)<br>RC2a←<br>RC2b→<br>层归一化<br>F.F.<br>RC2b← | RC2a→<br>层归一化<br>多头自注意力(4,256)<br>RC2a←<br>RC2b→<br>层归一化<br>F.F.<br>RC2b← |
| 3 | 池化<br>层归一化<br>线性层<br>GELU<br>丢弃(0.5)<br>线性层<br>Sigmoid | 池化<br>层归一化<br>线性层<br>GELU<br>丢弃(0.5)<br>线性层<br>Sigmoid | RC3a→<br>层归一化<br>多头自注意力(16,64)<br>RC3a←<br>RC3b→<br>层归一化<br>F.F.<br>RC3b← | RC3a→<br>层归一化<br>多头自注意力(4,256)<br>RC3a←<br>RC3b→<br>层归一化<br>F.F.<br>RC3b← |
| 4 | N/A | N/A | RC4a→<br>层归一化<br>多头自注意力(16,64)<br>RC4a←<br>RC4b→<br>层归一化<br>F.F.<br>RC4b← | RC4a→<br>层归一化<br>多头自注意力(4,256)<br>RC4a←<br>RC4b→<br>层归一化<br>F.F.<br>RC4b← |
| 5 | | | [token]→<br>层归一化<br>线性层<br>GELU<br>丢弃(0.5)<br>线性层<br>Sigmoid | [token]→<br>层归一化<br>线性层<br>GELU<br>丢弃(0.5)<br>线性层<br>Sigmoid |

可学习令牌对应的输出会馈入一个包含两个全连接层的多层感知机；为简化起见，也省略了位置嵌入。ViT-α 的每个多头自注意力层包含16个自注意力头，每个头有64个通道；而ViT-β 的每个多头自注意力层由4个注意力通道数为128的头组成。补充表5展示了TRSF-α、TRSF-β、ViT-α以及

**补充表6**：LSTM-α、LSTM-β、RNN-α 和 RNN-β 的架构。我们在每个“RNN单元”和“LSTM单元”后追加隐藏状态单元。

| 层 | LSTM-α | LSTM-β | RNN-α | 循环神经网络-β |
|---|---|---|---|---|
| 1 | LSTM单元 (256) | LSTM单元 (128) | RNN单元 (256) | RNN单元 (128) |
| 2 | 线性层<br>Sigmoid | 线性层<br>Sigmoid | 线性层<br>Sigmoid | 线性层<br>Sigmoid |

视觉变换器-β 详解。

**LSTM 和 RNN**。 长短期记忆网络（LSTM） [11]和循环神经网络（RNN） [12] 均属于循环网络，但实现方式各不相同。RNN 因架构较为简单，无法捕捉数据中的长期时间依赖关系；LSTM 则通过引入更复杂的门控机制改善了这一局限。在本研究中，我们分别采用了两种 LSTM 和两种 RNN，记为 LSTM-α 和 LSTM-β 、RNN-α和 RNN-β。这四种架构均将时间序列数据输入一个循环层（基于 LSTM 或 RNN），再接一个全连接层进行最终预测。LSTM-α与 LSTM-β 的区别在于，前者的循环层包含 256 个单元的隐藏状态，而后者为 128 个单元。我们将 RNN-α的循环层隐藏状态维度设为 256，将 RNN-β设为 128。补充表6 说明了 LSTM-α 和 LSTM-β、RNN-α和 RNN-β这几种架构。

## 训练设置

**补充表7**：三个数据集上11个深度神经网络的训练设置。所有实验均使用二元交叉熵损失和Adam优化器。“初始学习率”指Adam优化器的初始学习率，“min(64, $|\mathcal{C}|$)”表示取64与 $|\mathcal{C}|$ （凝缩集的大小）中的较小值作为批大小。

| 数据集 | 训练数据 | 超参数 | | |
|---|---|---|---|---|
| | | 批大小 | 初始学习率 | 训练轮数 |
| PhysioNet-2012 | 原始数据 | 64 | $5 \times 10^{-4}$ | 30 |
| | 压缩数据 | min(64, $\lvert\mathcal{C}\rvert$) | | |
| MIMIC-III | 原始数据 | 64 | $5 \times 10^{-5}$ | 20 |
| | 压缩数据 | | | |
| Coswara | 原始数据 | 64 | $1 \times 10^{-4}$ | 30 |
| | 压缩数据 | min(64, $\lvert\mathcal{C}\rvert$) | | |

在全部三个数据集上，我们使用原始训练集及其压缩数据版本训练全部 11 个网络，以评估数据凝缩（DC）所保留的深度学习效用。我们采用二元交叉熵（BCE）损失训练所有深度神经网络（DNN），并在所有实验中使用 Adam [16] 优化器，同时针对不同数据集采用不同的初始学习率。我们还在将数值型输入数据馈入 DNN 之前进行数据标准化（减去均值并除以标准差）。详细的训练超参数汇总于补充表7。请注意，这些设置应与主稿“方法”部分所述的凝缩数据学习超参数区分开来。

# 补充说明3 - 结果

**11个深度神经网络的AUC值**

**补充表8**：在全部3个数据集和11个网络上，深度神经网络在原始数据集和压缩数据集上训练所得的曲线下面积。我们将每个训练会话重复五次，并报告测试集上的平均AUC及标准差。“平均值”指所有会话的平均AUC。

| 数据集 | 训练数据（样本） | 测试AUC（标准差） | | | |
|---|---|---|---|---|---|
| | | 时间卷积网络-α | 时间卷积网络-β | 时间卷积网络-γ | 视觉变换器-α |
| PhysioNet-2012 | 原始(5120) | 0.854 (0.005) | 0.858 (0.005) | 0.852 (0.006) | 0.854 (0.003) |
| | 对照（80） | 0.806（0.003） | 0.800 (0.003) | 0.804（0.006） | 0.798（0.008） |
| | 对照(40) | 0.813 (0.003) | 0.804 (0.005) | 0.808 (0.005) | 0.788 (0.007) |
| | 对照(20) | 0.808 (0.006) | 0.807 (0.002) | 0.808 (0.002) | 0.806 (0.011) |
| MIMIC-III | 原始(14698) | 0.843 (0.004) | 0.833（0.002） | 0.845（0.001） | 0.835（0.002） |
| | 对照（1200） | 0.756（0.001） | 0.750（0.006） | 0.756 (0.002) | 0.743 (0.011) |
| | 对照(800) | 0.748 (0.002) | 0.736 (0.009) | 0.747 (0.000) | 0.744 (0.007) |
| | 对照(400) | 0.740 (0.006) | 0.704 (0.021) | 0.742 (0.007) | 0.740 (0.008) |
| Coswara | 原始(987) | 0.729 (0.017) | 0.750 (0.007) | 0.740 (0.018) | 0.727 (0.019) |
| | 对照(80) | 0.646 (0.010) | 0.635 (0.006) | 0.641 (0.002) | 0.637 (0.030) |
| | 对照(40) | 0.645（0.007） | 0.636（0.008） | 0.650 (0.014) | 0.626 (0.021) |
| | 对照（20） | 0.647 (0.006) | 0.633 (0.005) | 0.651 (0.012) | 0.590 (0.022) |
| | | 视觉变换器-β | 变换器-α | 变换器-β | 长短期记忆网络-α |
| PhysioNet-2012 | 原始(5120) | 0.851 (0.005) | 0.853 (0.003) | 0.843 (0.016) | 0.869 (0.003) |
| | 对照(80) | 0.798 (0.010) | 0.781 (0.006) | 0.788 (0.005) | 0.815 (0.004) |
| | 对照(40) | 0.797 (0.006) | 0.779 (0.007) | 0.784 (0.011) | 0.816 (0.007) |
| | 对照(20) | 0.794 (0.014) | 0.789 (0.007) | 0.787 (0.004) | 0.804 (0.005) |
| MIMIC-III | 原始(14698) | 0.830 (0.003) | 0.835 (0.001) | 0.832 (0.004) | 0.849 (0.004) |
| | 对照(1200) | 0.745 (0.007) | 0.742 (0.005) | 0.743 (0.005) | 0.776 (0.003) |
| | 对照(800) | 0.741 (0.006) | 0.734 (0.006) | 0.742 (0.006) | 0.777 (0.011) |
| | 对照(400) | 0.737 (0.005) | 0.735 (0.009) | 0.737 (0.008) | 0.770 (0.010) |
| Coswara | 原始(987) | 0.738 (0.020) | 0.732 (0.017) | 0.741 (0.020) | 0.739 (0.005) |
| | 对照(80) | 0.634 (0.009) | 0.642 (0.019) | 0.629 (0.017) | 0.654 (0.006) |
| | 对照(40) | 0.614（0.017） | 0.630 (0.034) | 0.643（0.026） | 0.654 (0.007) |
| | 对照（20） | 0.618 (0.029) | 0.603 (0.024) | 0.602 (0.028) | 0.663 (0.013) |
| | | AVG | 长短期记忆网络-β | 循环神经网络-α | 循环神经网络-2-β |
| PhysioNet-2012 | 原始(5120) | 0.858 (0.011) | 0.873（0.003） | 0.864（0.009） | 0.872 (0.005) |
| | 对照(80) | 0.804 (0.014) | 0.817 (0.011) | 0.820 (0.004) | 0.820 (0.006) |
| | 对照(40) | 0.804 (0.016) | 0.813 (0.012) | 0.822 (0.008) | 0.815 (0.011) |
| | 对照(20) | 0.803 (0.012) | 0.816（0.005） | 0.811 (0.013) | 0.804（0.015） |
| MIMIC-III | 原始(14698) | 0.840 (0.007) | 0.844 (0.002) | 0.847 (0.001) | 0.845 (0.002) |
| | 对照(1200) | 0.756 (0.014) | 0.774 (0.007) | 0.768 (0.004) | 0.761 (0.013) |
| | 对照(800) | 0.750 (0.014) | 0.761 (0.006) | 0.757（0.006） | 0.761 (0.012) |
| | 对照（400） | 0.741 (0.019) | 0.749（0.016） | 0.760（0.007） | 0.738（0.012） |
| Coswara | 原始(987) | 0.737 (0.017) | 0.735（0.007） | 0.753（0.005） | 0.721 (0.014) |
| | 对照(80) | 0.642 (0.016) | 0.651 (0.011) | 0.648 (0.005) | 0.640 (0.013) |
| | 对照(40) | 0.641 (0.021) | 0.652 (0.012) | 0.658 (0.009) | 0.645 (0.008) |
| | 对照(20) | 0.632 (0.030) | 0.653 (0.010) | 0.648 (0.009) | 0.647 (0.011) |

## 数据凝缩中改变时间维度的影响

**补充表 9**：压缩数据在 3 个数据集上不同时间维度的影响。 $|\mathcal{C}|$ 表示凝缩集中的样本数，而“$T$ (原始)”和“$T$ (对照)”分别是原始数据和凝缩样本中时间维度的缩写。减小$T^*$ 可以进一步压缩数据量，同时不会显著影响测试 AUC，体现了数据凝缩在大小控制方面的高度灵活性。

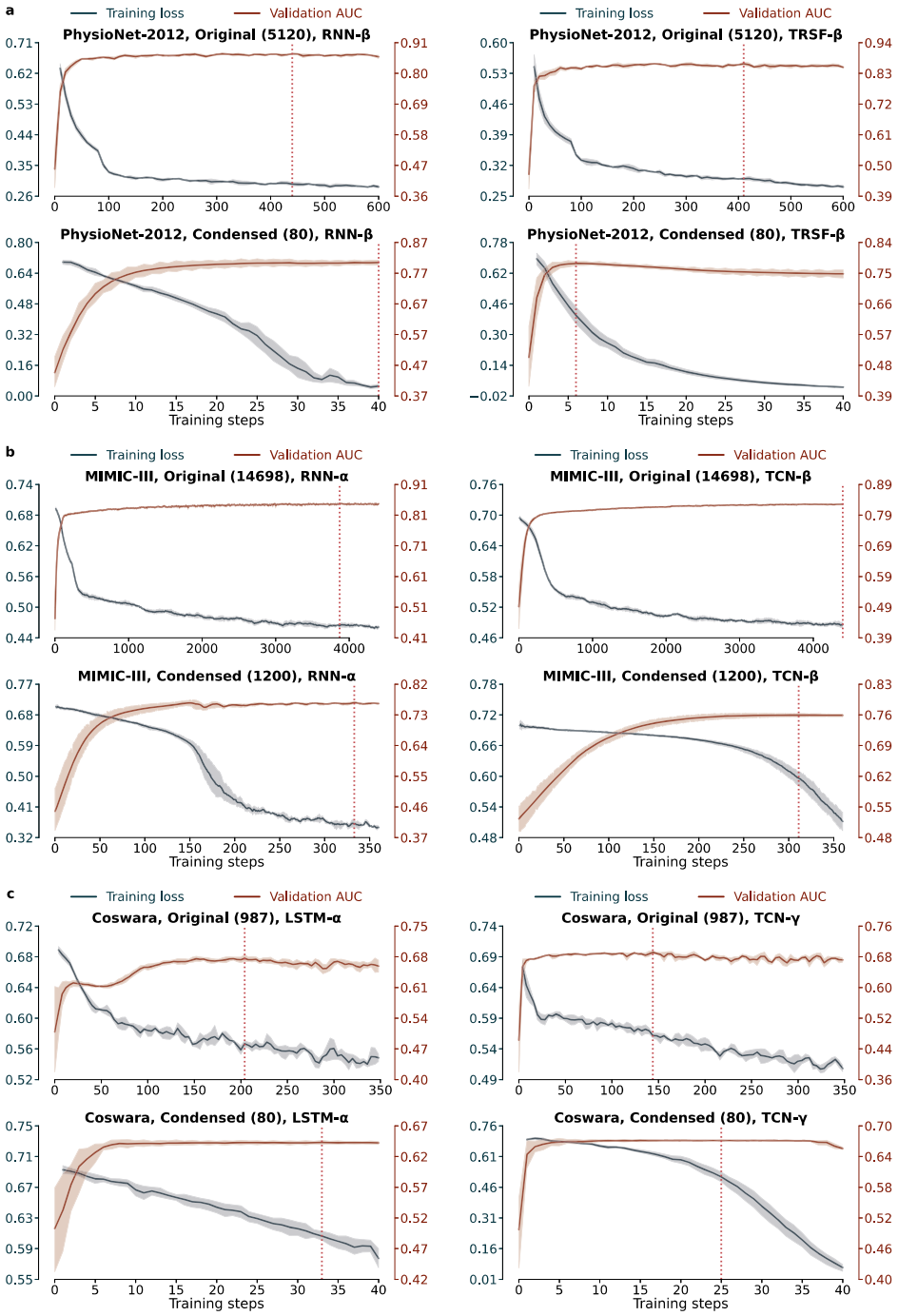
| 数据集 | $\|\mathcal{C}\|$ | 时间维度 $T$ （原始） | $T^*$ （对照） | 大小 (MB) | 测试AUC（标准差） |
|---|---|---|---|---|---|
| PhysioNet-2012 | 80 | 48 | 48 | 0.69 | 0.804 (0.014) |
| | | | 36 | 0.52 | 0.799 (0.015) |
| | | | 24 | 0.34 | 0.801 (0.016) |
| | | | 12 | 0.17 | 0.801 (0.016) |
| MIMIC-III | 1200 | 48 | 48 | 13.18 | 0.756 (0.014) |
| | | | 36 | 9.89 | 0.745 (0.017) |
| | | | 24 | 6.59 | 0.747 (0.017) |
| | | | 12 | 3.30 | 0.746 (0.018) |
| Coswara | 80 | 96 | 96 | 1.88 | 0.642 (0.016) |
| | | | 64 | 1.25 | 0.639（0.021） |
| | | | 32 | 0.63 | 0.631（0.034） |
| | | | 16 | 0.31 | 0.632（0.030） |

在整篇论文中，我们预先设定每个凝聚样本保留与原始样本相同的大小，即对于原始实例 $\mathbf{x} \in \mathbb{R}^{T\times F}$，其中 $T$ 和 $F$分别对应时间维度和特征维度，我们将学习形状相同的凝聚样本 $\mathbf{c} \in \mathbb{R}^{T\times F}$。然而，原则上，只要数据凝缩（DC）中两者的嵌入分布仍然能够匹配，我们也可以将凝聚样本设置为与原始样本具有不同的大小。在本研究中，我们采用11种深度神经网络（DNNs）作为嵌入函数，而所有架构均可接受具有任意时间维度的输入。也就是说，我们可以定义凝聚样本 $\mathbf{c} \in \mathbb{R}^{T^*\times F}$，其中 $T^* \neq T$或 $T^* = T$。

我们在补充表9中研究了凝缩样本中不同时间维度（TD）的影响。我们发现，使用较小的$T^*$ 可以在不显著影响测试 AUC的情况下进一步压缩数据量。这一观察反映出数据凝缩的另一项优势：大小控制的灵活性。在选择 $|\mathcal{C}| \ll |\mathcal{O}|$ 后，我们可以通过选取一个 $T^* < T$ 来进一步压缩凝缩数据的大小，但准确率会略有下降。

## 学习曲线

**补充图1**：在3个数据集上，使用原始数据和凝缩数据训练的不同网络的学习曲线。红色虚线表示收敛步骤。**a**，PysioNet-2012上的循环神经网络-β 和变换器-β 。**b**，MIMIC-III上的循环神经网络-α 和时间卷积网络-β 。**c**，Coswara上的长短期记忆网络-α 和时间卷积网络-γ 。使用凝缩数据训练的深度神经网络，其收敛速度显著优于使用原始数据训练的相同网络。



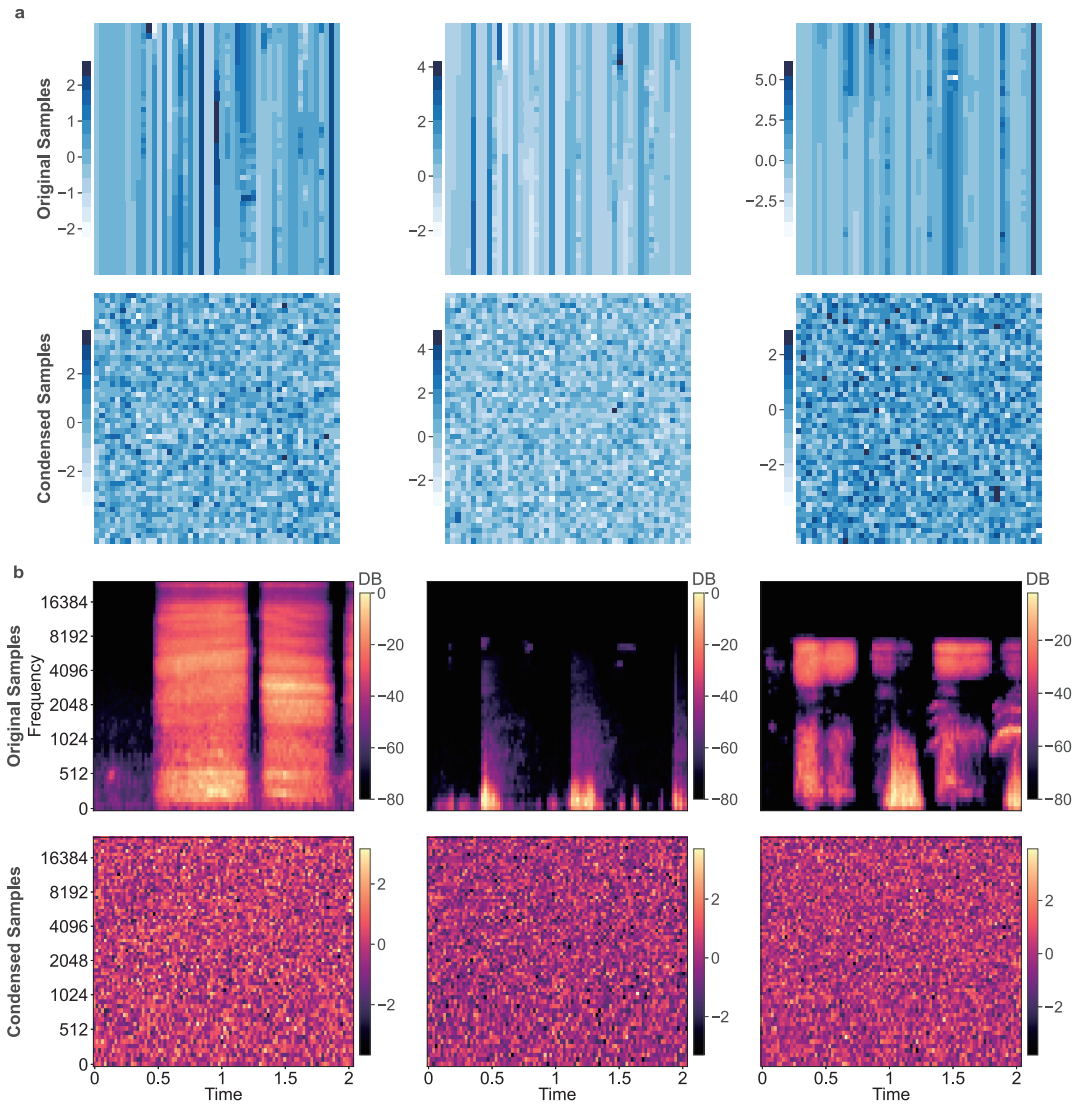
## 数据凝缩的学习时间

**补充表10**：在3个数据集上学习不同大小凝缩集所需的训练时间。我们测量了使用 NVIDIA GeForce RTX 3060 显卡进行的每次数据凝缩学习（24,000 次迭代）的时长（分钟）。所有凝聚数据集均可在约1小时内完成学习。

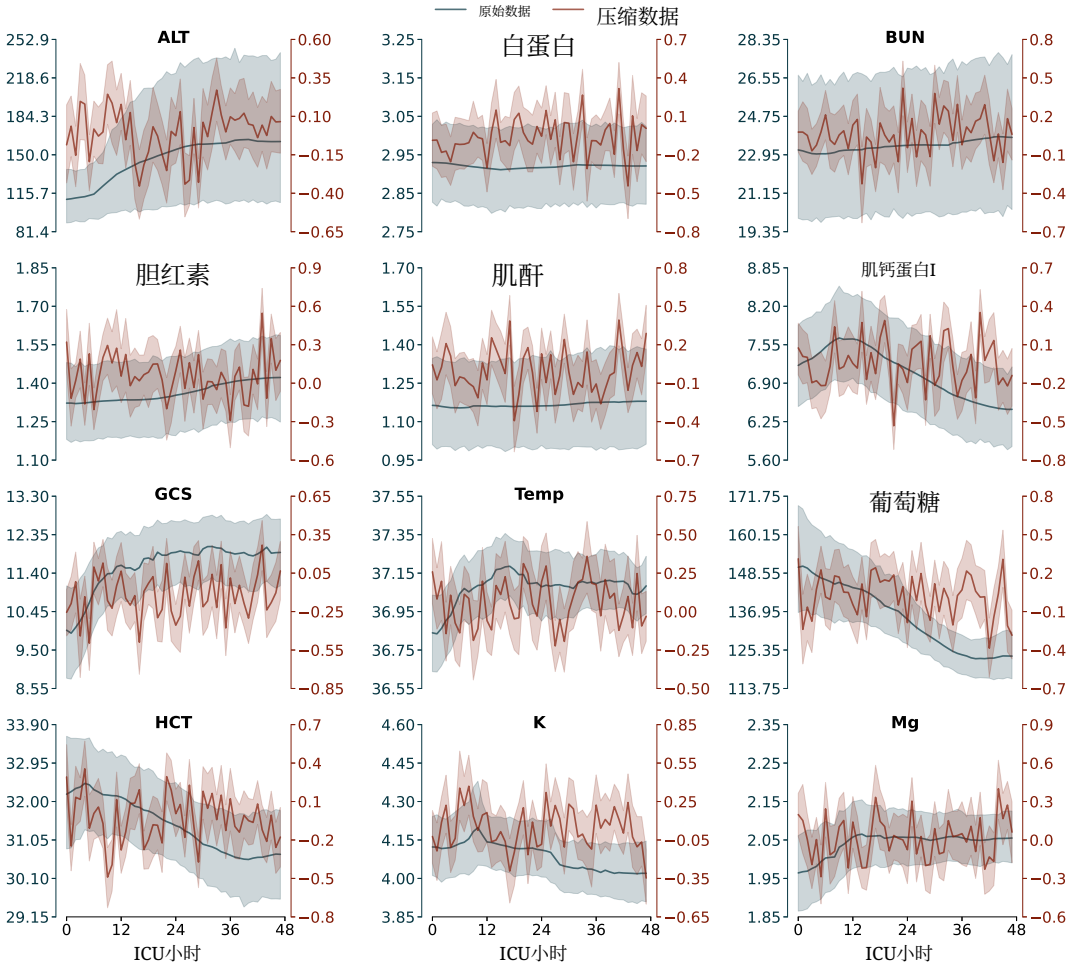
| **数据集**（样本） | **压缩数据大小** | **数据凝缩时间**（分钟） |
|---|---|---|
| PhysioNet-2012（5120） | 80 | 18.3 |
| | 40 | 16.9 |
| | 20 | 16.2 |
| MIMIC-III（14698） | 1200 | 61.8 |
| | 800 | 46.0 |
| | 400 | 30.6 |
| Coswara（987） | 80 | 31.1 |
| | 40 | 28.2 |
| | 20 | 27.4 |

## 数据可视化

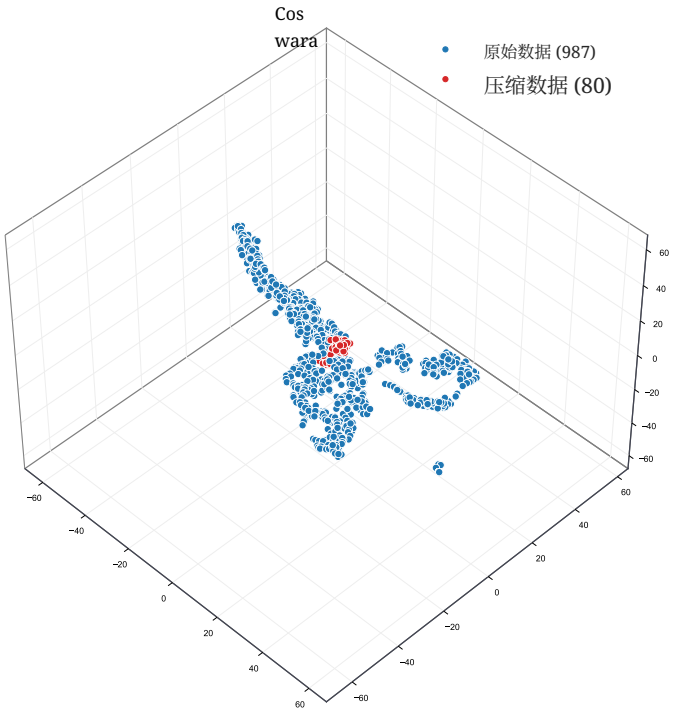
**补充图2**: a，从 PhysioNet 中随机选取的三个原始数据样本和三个凝缩样本的热力图。b，Coswara 上随机选取的三个音频片段和三个随机凝缩样本的梅尔频谱图。凝缩样本表现得像随机噪声，不像原始样本那样具有规律模式。



**补充图3**：分别基于80个原始数据实例和80个压缩数据实例，计算得出的PhysioNet-2012 中另外12项临床变量的48小时趋势。这些变量包括：丙氨酸转氨酶（“ALT”）、白蛋白、血尿素氮（“BUN”）、胆红素、血清肌酐（“Creatinine”）、肌钙蛋白I、格拉斯哥昏迷评分（“GCS”）、体温（“Temp”）、血清葡萄糖（“Glucose”）、血细胞比容（“HCT”）、血清钾（“K”）和血清镁（“Mg”）。



**补充图4**：使用 t-SNE 将 Coswara 中的 987 个原始数据样本和 80 个压缩数据样本嵌入三维空间后得到的三维分布。




# 参考文献

[1] Harutyunyan, H., Khachatrian, H., Kale, D.C., Ver Steeg, G., Galstyan, A.：基于临床时间序列数据的多任务学习与基准测试。Scientific data6(1), 1–18 (2019)[2] Silva, I., Moody, G., Scott, D.J., Celi, L.A., Mark, R.G.：预测重症监护病房患者的住院死亡率：PhysioNet/Computing in Cardiology 2012挑战赛。载于：2012 Computing in Cardiology，第245–248页 (2012)。IEEE[3] Le Gall, J.-R., Loirat, P., Alperovitch, A., Glaser, P., Granthil, C., Math- ieu, D., Mercier, P., Thomas, R., Villers, D.：重症监护病房患者简化急性生理评分。Critical care medicine12(11), 975–977 (1984)[4] Ferreira, F.L., Bota, D.P., Bross, A., M´elot, C., Vincent, J.-L.：连续评估SOFA评分以预测危重症患者的结局。Jama286(14), 1754–1758 (2001)[5]Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.-w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., Mark, R.G.：

Mimic-iii，一种免费开放访问的重症监护数据库。Scientific data**3**(1),1–9 (2016)[6] Johnson, A., Pollard, T., Mark, R.：Mimic-iii 临床数据库（版本1.4）。PhysioNet**10**，2–26 (2016)[7] Sharma, N., Krishnan, P., Kumar, R., Ramoji, S., Chetupalli, S., Nirmala, R., Kumar Ghosh, P., Ganapathy, S.：Coswara——用于新冠肺炎诊断的呼吸、咳嗽和语音数据库。载于：国际语音通信协会年会论文集，INTERSPEECH，第2020卷，第4811–4815页 (2020)。国际语音通信协会[8] Martinez, B., Ma, P., Petridis, S., Pantic, M.：使用时间卷积网络进行唇读。载于：ICASSP 2020-2020 IEEE 国际声学、语音与信号处理会议 (ICASSP)，第6319–6323页 (2020)。IEEE[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.：Attention is all you need。神经信息处理系统进展**30** (2017)[10] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.：一张图像胜过16×16个词：用于大规模图像识别的变换器。载于：国际学习表征会议 (2021)。https://openreview.net/forum?id=YicbFdNTTy[11] Hochreiter, S., Schmidhuber, J.：长短期记忆。Neural com-putation**9**(8),1735–1780 (1997)[12] Hopfield, J.J.：具有涌现集体计算能力的神经网络与物理系统。美国国家科学院院刊**79**(8),2554–2558 (1982)[13] He, K., Zhang, X., Ren, S., Sun, J.：深入研究整流器：在ImageNet分类上超越人类水平的表现。载于：IEEE 国际计算机视觉会议论文集，第1026–1034页 (2015)[14] Bai, S., Kolter, J.Z., Koltun, V.：用于序列建模的通用卷积网络和循环网络的实证评估。arXiv预印本 arXiv:1803.01271 (2018)[15] Hendrycks, D., Gimpel, K.：高斯误差线性单元 (GELU)。arXiv预印本 arXiv:1606.08415 (2016)

[16] Kingma, D.P., Ba, J.：Adam：一种随机优化方法。arXiv 预印本 arXiv:1412.6980（2014）