医学人工智能领域数据稀缺与小样本学习问题调研报告

引言

人工智能 (AI) 在医疗健康领域的应用展现出革命性的潜力，其应用范围涵盖了从疾病诊断、个性化治疗方案制定到药物研发等多个方面。然而，尽管前景广阔，医学 AI 的全面发展仍面临一项根本性挑战：高质量、大规模数据集的稀缺性。这一问题在医学领域尤为突出，其成因复杂，涉及伦理、法规、技术以及数据本身的特性。数据匮乏直接导致了 AI 模型，特别是深度学习模型，难以达到理想的性能和泛化能力。

在此背景下，小样本学习 (Small-Sample Learning, SSL) 或少样本学习 (Few-Shot Learning, FSL) 成为一项至关重要的研究方向。其核心目标是开发能够在数据量有限的情况下有效学习的 AI 模型，这对于许多现实世界中的医学应用而言是必不可少的。鉴于医学数据的敏感性和获取难度，依赖海量数据训练模型的传统范式在医学 AI 领域往往难以为继。

本报告旨在对医学 AI 领域的数据稀缺和小样本学习问题进行全面而深入的调研。报告将系统梳理该问题的概要、产生的背景、重要的应用价值，剖析当前的研究现状与技术瓶颈，回顾该领域过往的关键研究里程碑，并列举相关的产业化案例。通过本次调研，期望为相关领域的研究人员和从业者提供有价值的参考。

一、 问题概要

数据稀缺在医学 AI 中的定义与表现

医学 AI 领域的数据稀缺，指的是用于训练稳健且具有良好泛化能力的 AI 模型的数据在数量和多样性上均显不足 [1]。值得注意的是，数据稀缺并非总是指数据在技术层面上的完全不存在，而更多地表现为数据的可获取性、可用性和适用性问题 [1]。许多因素共同导致了这一困境，其中，患者隐私保护的严格规定、数据采集的高昂成本以及医学专家进行数据标注的巨大工作量是主要原因 [2]。这种稀缺性已成为一个显著障碍，对 AI 在医药行业的进一步发展和潜力释放构成了威胁 [1]。例如，在许多医疗 AI 应用中，由于使用的是相

对较小的数据集，其模型准确性往往低于临床实践中的标准，导致模型鲁棒性和泛化能力不足[1]。

这种状况不仅仅是数据量的不足，更深层次地看，是“可用、已标注、具有代表性”数据的缺乏。虽然原始医疗数据可能大量存在，但将其转化为 AI 模型可以直接使用的、高质量的标注数据集，需要耗费大量的人力、物力和财力，并且依赖于高度专业的医学知识[1]。因此，解决数据稀缺问题，不仅要关注模型学习能力的提升，还需要关注数据处理流程和标注效率的优化。

小样本学习 (SSL) 的挑战

小样本学习旨在解决仅用少量样本（而非成千上万的样本）训练 AI 模型的问题[1]。这在医学领域至关重要，因为对于许多疾病，特别是罕见病而言，获取大规模数据集几乎是不可能的[2]。其核心挑战在于如何使模型能够从极为有限的数据中高效地识别模式并具备良好的泛化能力[1]。传统的深度学习模型通常依赖大规模标注数据集进行训练，在数据量不足时，其性能会显著下降[2]。

对于罕见病而言，数据稀缺问题形成了一个自我延续的困境。罕见病本身的病例数量就非常有限[2]，这意味着可用于训练针对这些疾病的 AI 模型的数据自然就少。如果缺乏有效的小样本学习方法，就无法开发出针对这些罕见病的 AI 工具。而这些工具的缺失，反过来又可能阻碍罕见病的早期诊断和深入研究，进一步限制了相关数据的收集机会。打破这一恶性循环，是小样本学习在医学领域应用的一个关键价值所在。

数据稀缺与小样本学习的内在联系及其影响

数据稀缺的现实直接催生了对小样本学习技术的需求。在普遍存在数据匮乏的背景下，如果 AI 模型不具备从小样本中学习的能力，那么针对广泛医学任务（包括图像分析中的分类、分割、配准，以及疾病诊断和治疗规划等）的有效 AI 工具的开发将举步维艰[2]。

其后果是多方面的：模型可能因信息不足而导致预测准确性低下，在不同临床环境中的适用性受到限制[1]。更为严重的是，这还可能引发关于 AI 模型有效性和公平性的伦理问

题，特别是当模型是基于有限的、可能存在偏倚的数据进行训练时[1]。如果 AI 模型只能针对拥有海量数据的常见疾病进行有效训练，那么 AI 技术带来的益处将无法惠及罕见病患者，这无疑会加剧医疗领域的不平等现象。因此，数据稀缺和小样本学习不仅是技术层面的挑战，更是实现普惠医疗 AI 的关键瓶颈。

二、 问题背景及应用价值

2.1 问题背景

医学 AI 领域数据稀缺的成因复杂多样，既有客观条件的限制，也有技术和管理层面的因素。

数据稀缺的深层原因

- **严格的隐私法规：** 如美国的 HIPAA 法案和欧盟的 GDPR 条例，为保护患者隐私，对医疗数据的访问、使用和共享施加了严格限制[2]。这是导致数据难以获取和流通的首要障碍。
- **数据孤岛现象：** 医疗数据常常分散存储于不同医疗机构的独立系统中，缺乏统一标准和互操作性，导致难以汇聚形成大规模数据集。
- **高昂的采集与标注成本：** 获取高质量的医学数据（如影像扫描、基因测序）本身成本不菲。而更具挑战性的是数据的标注工作，例如由放射科医生、病理科医生进行的图像区域勾画或病灶性质判断，不仅耗时费力，成本高昂，还容易受到观察者间主观差异的影响[2]。这种“标注成本”并非单纯的资金投入，更包含了高级医疗专家时间的巨大机会成本，这些时间本可用于患者护理。同时，具备标注能力的专家资源本身也是稀缺的。因此，标注瓶颈是一个涉及资金、专家时间、专业知识的多维度难题，单纯依靠增加资金投入难以彻底解决，这也凸显了能够减轻标注负担的技术（如小样本学习、自监督学习）的重要性。
- **罕见病的数据困境：** 对于罕见病而言，由于患病人数少，可供研究的病例数据自然也极为有限[2]。
- **数据异构性与标准化难题：** 医疗数据来源广泛，模态多样（如 MRI、CT、X 射线、

电子健康记录 EHR 等），格式各异。同时，不同医疗设备、扫描参数、以及患者群体的差异也会导致数据特征的多样性和不一致性，这给数据的标准化和整合带来了巨大挑战[3]。

数据稀缺对 AI 模型开发的影响

- **泛化能力差：** 在有限数据上训练的模型可能在训练集上表现良好，但难以推广到新的、未见过的数据，尤其当这些新数据来自不同的患者群体、医疗机构或采集设备时[1]。
- **模型偏倚：** 如果稀缺的数据未能充分代表广泛的患者人群（例如，在年龄、性别、种族等方面缺乏多样性），AI 模型可能会习得这些偏倚，导致其在代表性不足的群体中表现不佳，甚至产生错误的诊断或治疗建议，加剧医疗不平等[3]。
- **准确性和鲁棒性降低：** 数据不足会导致模型学习不够充分，准确性难以保证，并且对数据中的噪声或微小变化更为敏感，即鲁棒性较差[1]。
- **伦理风险：** 部署基于不充分或有偏倚数据训练的 AI 模型，会引发对其可靠性、公平性以及潜在危害的严重伦理关切[1]。
- **过拟合问题：** 深度学习模型通常参数众多，当训练数据量较小时，极易发生过拟合现象，即模型仅仅记住了训练数据的特征，而未能学习到普适的规律[10]。

2.2 应用价值

有效解决医学 AI 领域的数据稀缺和小样本学习问题，具有极其重要的应用价值，不仅能推动 AI 技术自身的进步，更能为医疗健康事业带来深远影响。

推动医学 AI 前沿发展

- **赋能罕见病诊疗：** 小样本学习技术对于开发罕见病的诊断工具至关重要，因为这些疾病难以获得大规模数据集[2]。
- **实现个性化医疗：** 能够适应个体患者特征或从特定小规模患者队列中学习的 AI 模型，是个性化医疗发展的关键支撑[2]。
- **加速医学研究与药物发现：** 克服数据限制能够显著加快医学研究的步伐，促进新疗

法、新药物的发现与验证过程。

提升医疗服务的可及性与效率

- **普及优质医疗资源：** 正如比尔·盖茨等人士所展望的，即使是在有限数据上通过有效方法训练的 AI，也有潜力将专业的医疗知识和经验普及到更广泛的地区，缓解医疗资源分布不均的问题 [11]。
- **优化临床工作流程：** 稳健的 AI 工具能够辅助临床医生完成例如医学影像解读等任务，减轻其工作负担，提高工作效率 [2]。
- **辅助临床决策：** AI 可以为医生提供有价值的“第二意见”，或提示医生可能忽略的细微模式，尤其是在处理复杂病例时。

值得注意的是，应对医疗 AI 数据稀缺的努力，正在推动隐私保护技术的创新，形成一种积极的反馈循环。隐私法规是数据稀缺的主要原因之一 [2]，这促使研究者开发如联邦学习（在不共享原始数据的情况下本地训练模型 [12]）和合成数据生成（旨在创建逼真且匿名的模拟数据 [1]）等技术。这种在符合隐私要求的前提下克服数据稀缺的需求，推动了这些隐私保护 AI 技术的进步。反过来，这些技术的进步使得处理敏感医疗数据变得更加可行，从而可能释放新的数据集或促进合作机会。这种问题（受隐私限制的数据稀缺）与解决方案（隐私保护技术）之间的协同进化，正在重塑医疗 AI 的研发生态。

因此，解决数据稀缺问题不仅仅是构建更好的 AI 模型，更深远地看，它关乎重新设计医疗知识的产生、共享和应用方式。在这一领域的成功，有望催生更具协作性的研究环境，加速科研成果向临床实践的转化，并最终构建更公平、更高效的医疗体系。然而，这也意味着必须同步发展针对这些新型数据形态（如合成数据 [6]）的伦理框架和监管机制。

三、 研究现状及瓶颈

针对医学 AI 领域的数据稀缺和小样本学习问题，学术界和产业界已经探索了多种技术路径，并取得了一定的进展。然而，这些方法也各自面临着不同的挑战和局限。

3.1 主要研究方向

3.1.1 小样本学习 (Few-Shot Learning - FSL)

小样本学习的核心思想是使模型能够像人类一样，从极少数的样本中进行学习和泛化 [2]。这对于医学数据集往往规模有限的现状至关重要 [2]。目前，FSL 的主要研究方向包括基于度量学习的方法和基于优化的方法（通常属于元学习范畴）。

基于度量学习的方法 (Metric-based Learning)

此类方法致力于学习一个优质的嵌入空间 (embedding space)，在该空间中，来自同一类别的样本彼此靠近，而来自不同类别的样本则相互远离。分类任务通过比较查询样本与各类别原型的距离来完成。

- **原型网络 (Prototypical Networks):** 该网络学习一个度量空间，其中分类通过计算查询样本嵌入与各类别原型（该类别支持样本嵌入的均值）之间的距离来实现 [1]。原型网络采用了一种在有限数据情况下更为有效的简单归纳偏置 [14]。
  - *医学应用案例：* 在跨年龄组的头影测量标志点检测中，通过比较图像特征与标志点原型进行定位，并采用特定策略改进原型对齐 [16]。在去中心化学习场景下，原型网络也被提议用于医学图像分类，以最小化不同客户端之间的异质性差异 [17]。
  - *可解释性潜力：* 原型网络通过学习具有代表性的图像模式（即原型），为临床医生理解模型决策提供了一定的可能性 [18]。然而，在空间定位的精确性以及将原型与具体文本描述相关联方面仍存在挑战 [18]。

基于优化的方法 (Optimization-based Learning / Meta-Learning)

这类方法的目标是“学会学习” (learning to learn)。模型在一系列不同的学习任务上进行训练，使其能够利用少量样本快速适应新的、未见过的任务。

- **模型无关元学习 (Model-Agnostic Meta-Learning - MAML):** MAML 的核心思想是训练模型的初始参数，使其能够在少量新任务数据上通过少数几次梯度下降步骤就达到良好的性能 [1]。其“模型无关”的特性使其能够应用于各种基于梯度下降训练的模型和不同类型的学习问题 [19]。

- *医学应用案例：* MAML 已被用于医学图像分割任务，在仅有少量标注数据的情况下，能够有效提升如 U-Net 等成熟分割网络的适应性和性能 [10]。在关于医学影像元学习的文献中，MAML 被认为是主流方法之一 [21]。

其他元学习方法 (Other Meta-Learning Approaches)

元学习构建了一套计算机制，能够系统而高效地使模型适应新任务，从而应对深度神经网络在数据需求量大、训练成本高以及任务间迁移能力有限等方面的根本性挑战 [21]。除了 MAML，元学习的流行变体还包括记忆增强网络 (memory augmentation) 和学习优化 (learning to optimize) 等 [21]。一项针对医学影像 FSL 的系统综述表明，元学习是处理此类问题的常用选择，尤其是在心脏、肺部和腹部等领域的应用中，模型能够凭借少量标注样本快速适应新任务 [2]。在医学图像分类任务中，元学习通过在源自不同 MRI 数据集的多个小型分类任务上训练模型，使其能够提取可迁移的模式，从而快速适应新的分类任务 [22]。

下表总结了医学 AI 领域中几种关键的小样本学习技术：

表 1: 医学 AI 领域关键小样本学习技术

技术名称	核心原理	在医学背景下的优势	在医学背景下的局限性/挑战	医学应用案例 (部分)
原型网络	学习一个嵌入空间，通过计算查询样本与类别原型的距离进行分类；原型是支持集样本嵌入的均值。	概念简单，归纳偏置适合小样本；可提供一定的可解释性。	对嵌入空间的质量高度敏感；原型可能无法充分代表类别复杂性；空间定位精度和语义关联有待提升 [18]。	头影测量标志点检测 [16]，罕见病分类，皮肤病变分类 [22]，联邦学习环境下的医学图像分类 [17]。

模型无关元学习 (MAML)	训练模型的初始参数，使其能通过少量梯度更新快速适应新任务。	模型无关，适用性广；能够学习快速适应的能力。	训练过程计算成本高（双层优化）；对超参数敏感；可能存在梯度病态问题。	医学图像分割 (如结合 U-Net) [10]，皮肤病变分割 [10]，医学图像分类 [23]。
其他元学习方法	例如记忆增强网络、学习优化器等，旨在从一系列任务中学习通用的“学习策略”或“知识表示”。	能够学习更复杂的适应机制；可能提升泛化能力。	理论理解和实现相对复杂；训练数据（任务）的多样性和质量要求高。	各种医学图像分析任务，如分类、分割，特别是在需要模型快速适应新设备、新疾病类型或个体化医疗场景中 [2]。

3.1.2 迁移学习 (Transfer Learning - TL)

迁移学习的核心思想是将在源任务或源领域（通常拥有充足数据，如 ImageNet 上的自然图像）中学到的知识（如特征表示、模型权重）应用于数据量有限的目标任务或目标领域（如特定类型的医学图像）[1]。在医学影像领域，一个常见的策略是使用在自然图像上预训练的模型，然后针对具体的医学任务进行微调 (fine-tuning) [4]。迁移学习有助于减少对大规模标注数据的依赖，但并不能完全消除这一需求 [3]。然而，迁移学习在医学领域的有效性并非总是确定的，相关研究结果有时甚至相互矛盾 [4]。当源领域和目标领域之间差异较大时，迁移学习的效果会显著下降 [3]。例如，自然图像与某些医学影像模态（如超声图像）之间的巨大差异，会削弱迁移学习的有效性 [9]。

3.1.3 数据增强 (Data Augmentation)

数据增强通过对现有图像进行受控的、合理的变换来人为地扩充训练数据集，从而帮助模型更好地泛化，提升其鲁棒性，尤其是在训练数据有限的情况下 [1]。

- **传统方法：** 包括简单的几何变换（如翻转、旋转、缩放、裁剪）和像素强度变换（如调整亮度、对比度）[1]。值得注意的是，对于医学图像，垂直翻转有时是合理的

（例如，肿块的垂直翻转仍然是真实的肿块），这与自然图像通常只进行水平翻转不同 [25]。

- **高级方法：** 包括弹性变形（模拟组织的可变性）、添加高斯噪声（模拟低质量扫描图像）、基于主成分分析 (PCA) 的颜色抖动、高斯模糊等 [24]。此外，可微分增强 (differentiable augmentation) 技术也被证明有助于稳定 GAN 的训练过程 [13]。
- **应用考量：** 在医学图像分割等任务中，数据增强操作必须保持图像内容与其对应掩模之间的空间关系的一致性。同时，所有增强方法都应尊重生物学上的合理性，避免引入不切实际的伪影或误导模型 [24]。因此，在选择和应用数据增强策略时，通常需要与临床医生合作进行验证。并非所有增强策略都是有益的，例如，不恰当地添加噪声可能导致模型性能显著下降 [25]。

3.1.4 合成数据生成 (Synthetic Data Generation)

合成数据生成旨在创建人工的、但与真实数据特征相似的数据，用于训练 AI 模型，尤其是在真实数据稀缺或高度敏感（如涉及患者隐私）的情况下 [1]。

- **生成对抗网络 (Generative Adversarial Networks - GANs):** GAN 由一个生成器 (generator) 和一个判别器 (discriminator) 组成，两者在极小极大博弈中进行对抗训练 [13]。GAN 在自然图像生成方面取得了令人瞩目的成果，激发了其在医学数据生成领域的应用探索 [13]。
  - *医学应用：* GAN 已被用于数据增强、数据匿名化、图像重建、病灶检测和图像分割等任务，涵盖 MRI、CT、乳腺 X 线摄影等多种模态 [26]。常见的 GAN 架构包括 DCGAN, LSGAN, CycleGAN, StyleGAN2 等 [26]。
  - *性能表现：* 部分先进的 GAN 模型能够生成在视觉上与真实医学图像难以区分的图像（通过 FID 等指标衡量，并通过视觉图灵测试），但在完整再现医学数据集的丰富细节和统计特性方面仍有不足，其生成的合成数据在特定下游任务中的表现通常仍逊于真实数据 [13]。
- **扩散模型 (Diffusion Models):** 扩散模型是近年来兴起的一种强大的生成模型，在高保真图像合成方面展现出巨大潜力。
  - *医学应用：* 在医学领域，扩散模型已被用于文本到图像的合成，例如根据临床文

本描述生成结肠镜图像或放射影像。研究路径包括微调大型预训练模型（如 FLUX, Kandinsky）或训练小型的领域专用模型（如 MSDM）[23]。针对三维医学影像（如 CT、MR 扫描），基于小波的扩散模型 (Wavelet Diffusion Models - WDM) 能够生成高分辨率（例如 128×128×128 和 256×256×256）的图像，并在图像保真度方面达到了当前最佳水平 [27]。

- **其他方法 (如变分自编码器 - VAEs):** VAE 是另一种常用的生成建模技术，也被用于合成数据的生成 [28]。

下表对比了医学领域中主要的合成数据生成方法：

表 2: 医学合成数据生成方法对比

方法名称	工作原理	医学数据生成优势	医学数据生成劣势/主要挑战	医学应用案例	著名架构举例
生成对抗网络 (GANs)	生成器和判别器通过对抗学习，生成器学习生成逼真数据，判别器学习区分真实与生成数据。	可生成多样性和新颖性较高的样本；在特定条件下可生成高分辨率图像。	训练不稳定 (模式崩溃、梯度消失、收敛困难 [13])；难以完全捕捉真实数据的复杂分布；验证困难；可能生成不符合医学常识的图像。	MRI 合成、CT 重建、病理图像生成、数据增强 [26]。	DCGAN, StyleGAN, CycleGAN
变分自编码器 (VAEs)	通过编码器将数据映射到潜在空间，再通过解码器从潜	训练相对稳定；潜在空间具有良好结	生成图像通常较模糊，细节不如 GAN；难以生成非常	异常检测、数据增强、特征学习。	-

	在空间重构数据，学习数据的概率分布。	构，便于插值和生成。	高分辨率的清晰图像。		
扩散模型	通过逐步向数据中添加噪声（前向过程），然后学习从噪声中逐步恢复数据（反向过程）来生成样本。	生成图像质量高，保真度好，多样性强；训练过程相对稳定。	计算成本高，采样速度慢（尽管已有改进）；理论理解和实现相对复杂。	高分辨率 2D/3D 医学影像合成 (CT, MRI) [27]，文本引导的医学图像生成 [23]。	WDM, MSDM

3.1.5 联邦学习 (Federated Learning - FL)

联邦学习的核心思想是允许多个机构（如医院）在不直接交换其本地私有数据的情况下，协同训练一个共享的 AI 模型。每个参与方在本地数据上训练模型，然后仅将模型的更新（如权重参数）发送到中央服务器进行聚合 [12]。联邦学习对于解决医学 AI 领域的数据稀缺和隐私保护问题至关重要，它使得在保护数据隐私的前提下，能够利用来自多个医疗机构的更大规模、更多样化的数据集进行模型训练 [7]。例如，NVIDIA Clara Train SDK 就提供了基于服务器-客户端架构的联邦学习功能 [12]。为了进一步增强隐私保护，还可以采用例如对模型参数进行截断处理或添加噪声阈值等技术 [29]。

3.1.6 自监督学习 (Self-Supervised Learning - SSL)

自监督学习允许 AI 模型从未标注的原始数据中进行学习。其核心机制是通过设计“借口任务” (pretext task)——例如预测图像中缺失的部分、或为灰度图像上色——来让模型自我学习数据的内在表示。这种方法显著减少了对大规模标注数据集的依赖 [1]。在医学领域，自监督学习可用于在海量的未标注医学数据（如影像、病理报告）上预训练模型，学习到有用的特征表示，然后这些预训练模型可以在特定下游任务中，利用少量标注数据进行微调，从而提升模型性能。

3.2 主要技术瓶颈与挑战

尽管上述研究方向为解决医学 AI 数据稀缺问题提供了多种途径，但每种方法自身也存在不容忽视的瓶颈和挑战。

FSL/元学习的局限性

- **任务泛化能力：** 通过元学习训练的模型可能难以泛化到与元训练阶段任务差异较大的新任务上。
- **计算复杂度：** 元学习，特别是像 MAML 这样基于优化的方法，由于其嵌套的优化循环，训练过程可能计算成本高昂 [21]。
- **可解释性：** 理解元学习模型是如何适应新任务的，或者为何某个度量空间是有效的，往往具有挑战性 [7]。
- **元训练数据质量：** 用于元训练的任务的多样性和质量对最终模型的性能至关重要。
- **医学领域的特殊挑战：** 医学图像本身高度复杂且变异性大，这使得 FSL 难以完全捕捉这些细微特征 [4]。此外，由于计算资源限制而进行的图像下采样操作，可能导致对医学诊断至关重要的细微结构信息的丢失 [7]。这种医学数据的复杂性与 FSL 模型为实现良好泛化而追求的简洁归纳偏置之间，存在一种内在的张力。如何在不牺牲必要复杂性的前提下实现有效的小样本学习，或者如何将领域知识更有效地融入模型以指导学习过程，是未来研究需要突破的关键点。

迁移学习的挑战

迁移学习是应对数据有限情况的常用策略，但在医学影像领域的应用面临诸多限制，如下表所示：

表 3: 迁移学习在医学影像中的局限性概述

局限性	在医学背景下的详细描述	对医学 AI 模型性能和信任度的影响	潜在缓解策略 (部分)

领域不匹配 (Domain Mismatch / Shift)	源领域（如自然图像）与目标医学领域（如 X 射线、MRI）的特征分布差异巨大，导致从源领域学到的特征与医学任务不相关 [3]。	导致模型性能欠佳，特征迁移效果差，降低模型可靠性。	领域自适应技术；选择更相关的源领域数据；设计更鲁棒的特征提取器。
有限的标注数据	迁移学习减少但不能完全消除对标注数据的需求；医学标注数据获取成本高、耗时长、易出错 [3]。	即使使用迁移学习，在极度缺乏标注数据或数据不平衡时，模型性能仍受限。	结合半监督学习、主动学习；改进标注工具和流程。
类别不平衡	许多医学数据集中不同类别的样本数量分布不均（如罕见病样本远少于常见病样本），影响模型准确性 [3]。	模型可能偏向于多数类，对少数类的识别能力差，导致漏诊或误诊。	代价敏感学习；过采样少数类、欠采样多数类；使用 FSL 技术处理不平衡类别。
可解释性有限	深度学习模型（包括迁移学习模型）通常被视为“黑箱”，决策过程不透明，难以获得临床医生的完全信任 [3]。	阻碍临床采纳；难以进行错误分析和模型改进；责任界定困难。	发展可解释 AI (XAI) 技术；可视化模型关注区域；结合临床知识进行模型验证。
泛化能力差	模型可能难以很好地泛化到来自不同扫描仪、不同参数设置或	限制模型在真实临床环境中的广泛应用；需要针对特定场景重	使用更多样化的训练数据；采用领域泛化技术；进行严格的多

	不同患者群体的图像，尤其当训练数据多样性不足时 [3]。	新训练或微调。	中心验证。
隐私和未见领域问题	少数研究关注隐私保护问题，或模型在完全未见过的领域（如新的疾病类型、新的成像设备）上的应用挑战 [3]。	限制模型在真实世界中的部署和协作共享；对突发新疾病的快速响应能力不足。	联邦学习；差分隐私；持续学习和模型更新机制。

数据增强的有效性与医学真实性

- **引入不真实伪影的风险：** 如果选择或应用不当，某些数据增强方法可能会引入不切实际的图像特征，或改变对诊断至关重要的信息，从而误导模型 [24]。
- **多样性有限：** 数据增强是在现有数据基础上进行变换，并不能从根本上引入全新的信息或未见过的病理模式。它不能替代真实世界中多样化的数据 [24]。
- **最佳策略选择：** 如何确定哪些增强策略最能捕捉医学图像的统计特性并有效提升模型性能，仍是一个持续研究的课题 [25]。

合成数据的质量、多样性与验证难题

- **GAN 训练不稳定性：** GAN 的训练过程 notoriously 困难，常伴有收敛困难、梯度消失和模式崩溃（即生成器仅能产生有限种类的样本）等问题 [13]。
- **模式崩溃 (Mode Collapse)：** 这是 GAN 训练中的一个主要障碍，会导致生成的合成数据集缺乏多样性，从而严重影响基于这些数据训练的下游模型的性能 [13]。
- **保真度与真实性：** 尽管 GAN 和扩散模型能够生成视觉上逼真的图像，但要确保它们能完整捕捉真实医学数据的丰富细节、微妙的病理特征以及准确的统计特性，仍然极具挑战性 [13]。在专门的下游任务中，合成数据通常仍无法超越真实数据集的表现 [26]。
- **验证困境：** 对合成医学数据进行严格的验证至关重要，但也异常困难。如何确保合成数据对下游任务确实有益，并且不会引入新的偏倚，是一个悬而未决的问题 [26]。

- **伦理与隐私风险：** 即便是合成数据，如果模型在训练过程中无意间"记住"了源数据的某些特征，也可能存在再识别的风险。确保公平性，避免偏倚放大也是关键的伦理考量 [6]。

模型的可解释性、可信度与临床接受度

- **"黑箱"本质：** 许多深度学习模型，包括用于小样本学习的模型，其决策过程缺乏透明度，使得临床医生难以理解其内部机制或完全信任其输出结果 [3]。这是阻碍 AI 在临床广泛应用的一个重要因素 [3]。
- **可解释 AI (XAI) 的需求：** 亟需有效的 XAI 工具来增强模型的透明度。像原型网络这类具有一定自解释性的模型展现了潜力，但仍面临挑战 [18]。
- **不确定性量化：** 对模型预测的不确定性（包括偶然不确定性和认知不确定性）进行量化，对于建立临床信任至关重要，它能帮助使用者了解预测结果的可靠程度 [7]。

数据隐私与安全问题

- 即使采用了联邦学习和合成数据生成等技术，确保稳健的隐私和安全保护依然是一个持续的挑战。对于联邦学习，共享模型更新仍可能存在一定的隐私泄露风险 [12]。对于合成数据，确保其不会泄露原始数据集的敏感信息至关重要 [6]。严格遵守 HIPAA、GDPR 等数据保护法规是基本前提 [5]。

缺乏标准化评估方法和基准数据集

- 由于缺乏针对医学 AI 领域的标准化基准数据集和统一的评估协议，对不同的小样本学习或合成数据生成技术的性能进行公平比较变得十分困难。MICCAI 等学术会议发起的开放数据计划 (Open Data initiative) 旨在通过推动医学影像数据集（特别是来自代表性不足人群的数据）的共享来缓解这一问题 [30]。

从这些研究现状和瓶颈中可以看出，医学 AI 领域的数据高效学习并非易事，不存在一劳永逸的解决方案。每种技术都在某些方面表现出优势，但在另一些方面则存在固有的局限性。这表明"没有免费的午餐"定理在此领域同样适用。例如，FSL 方法虽然前景广阔，

但在泛化性和计算开销上仍有待改进[4]；迁移学习应用广泛，但领域漂移和负迁移的风险不容忽视[3]；数据增强简单易行，却可能产生不真实的样本或提供的多样性有限[24]；合成数据生成技术（如 GAN 和扩散模型）能够创造新样本，但面临训练不稳定、模式崩溃和验证困难等挑战[13]。这意味着，单一技术不太可能普遍适用于所有场景，未来的趋势更可能是根据具体的医学问题、数据模态和可用资源，采用定制化的混合策略，例如，利用合成数据进行模型预训练，再结合 FSL 技术在真实的有限样本上进行微调。

此外，数据高效学习技术的发展正在重新定义“数据”本身的内涵。它已从传统的、由专家标注的真实样本，扩展到经过增强的样本、完全人工合成的样本，甚至在联邦学习和 MAML 等方法中，模型梯度或参数更新本身也成为了一种可学习的“信息”形式[1]。这种对可学习信息定义的扩展和抽象，为克服物理数据量的限制提供了新的思路，但也给验证、信任以及潜在的抽象偏倚带来了新的复杂性。

总而言之，当前研究的瓶颈表明，仅靠技术进步是不足以完全解决问题的。领域的进一步发展还需要依赖于更优的数据实践（如数据共享和标准化[30]）、更强的模型可解释性方法以建立临床信任[7]，以及针对新型数据（如合成数据）的健全伦理指南[6]。整个领域正朝着一个模型开发、数据工程和伦理考量日益交织的整体化方向发展。

四、 过往研究的里程碑

回顾医学 AI 领域数据稀缺与小样本学习的研究历程，可以看到一系列关键的理论突破和技术进展，它们共同塑造了当前的研究格局。这些里程碑往往遵循一个模式：首先在通用机器学习领域取得基础性理论进展，随后被医学 AI 领域的研究者们采纳、调整并应用于解决医学影像等具体问题。

奠基性的小样本学习算法

- **原型网络 (Prototypical Networks) (Snell et al., 2017):** 这是基于度量学习的 FSL 领域的一个重要里程碑。该工作提出了一种简单而有效的方法，通过学习一个度量空间并在其中计算与类别原型的距离来进行分类[1]。原型网络取得了优异的成果，并证明

了在数据有限的情况下，更简单的归纳偏置可能比复杂的元学习架构更为有效[14]。其在医学影像领域的应用，如图像分类和标志点检测等任务，正逐渐兴起[16]。

- **模型无关元学习 (MAML) (Finn et al., 2017):** 这是基于优化的元学习领域的一项标志性工作。MAML 提供了一个通用的框架，用于训练模型参数，使其能够通过少量梯度更新快速适应新任务[1]。该方法在小样本图像分类基准上取得了当时的最佳性能，并展示了其在回归和强化学习等问题上的适用性[19]。MAML 在医学图像分割和分类等任务中的应用也得到了关注[10]。

生成对抗网络 (GANs) 在医学影像中的发展

- **早期应用与进展：** 自 GAN 被提出以来，其在医学图像合成方面的潜力迅速得到探索。
  - **DCGAN (Deep Convolutional GAN):** 作为早期 GAN 架构之一，DCGAN 通过引入卷积层结构，显著提升了 GAN 训练的稳定性，并被应用于脑部 MRI 分类等任务[26]。
  - **StyleGAN (Karras et al.):** StyleGAN 在自然图像（尤其是人脸）的生成质量和多样性方面取得了巨大突破[13]，其生成高分辨率、逼真图像的能力，激发了研究者将其适配并评估用于医学数据生成的尝试 [[13] (例如 StyleGAN2 用于牙科 CT 重建)]。
- **应对 GAN 的挑战：** 大量研究致力于缓解 GAN 训练过程中遇到的模式崩溃、梯度消失和收敛困难等问题，并发展出如标签平滑 (label smoothing)、特征匹配 (feature matching) 和可微分增强 (differentiable augmentation) 等技术[13]。
- **系统性综述：** 如 Yi 等人 (覆盖 2016-2019 年文献) 和 Kazeminia 等人的综述指出，GAN 生成的医学图像在匿名化、重建、检测和分割等应用中的视觉真实感不断提升，但同时也强调了对其临床应用有效性验证不足的担忧[26]。

扩散模型在高保真医学图像生成中的崛起

- **作为新兴范式的出现：** 近年来，扩散模型因其生成高质量、多样化图像的能力而备受瞩目，在某些方面甚至超越了 GAN。

- **文本到图像的合成：** 出现了如 MSDM 这样的模型，用于根据临床文本描述生成医学图像，并比较了微调大型预训练模型（如 FLUX, Kandinsky）与训练小型的领域专用模型的效果 [23]。
- **高分辨率三维医学图像合成：** 针对高维度的三维医学扫描数据（如 CT、MR），研究者开发了专门的架构，如 WDM (Wavelet Diffusion Models)，能够在 1283 甚至 2563 的分辨率下生成高质量图像，并在图像保真度方面超越了以往的方法 [27]。这对于体数据医学影像分析是一个关键进展。

联邦学习在协作式医学 AI 中的应用

- **概念化与早期框架：** 医疗领域对隐私保护下的协同学习的需求，催生了联邦学习的发展和应用。
- **NVIDIA Clara Train SDK:** 为医学影像领域的联邦学习提供了一个较早的、易于使用的平台。该平台采用服务器-客户端架构，通过仅共享模型更新而非原始数据来保护隐私，从而促进了多机构间的安全协作 [12]。

医学 AI 领域关键的元学习与 FSL 综述

- 针对医学影像 FSL 的系统综述（例如 [2]）对现有技术（包括元学习、监督学习和半监督学习方法）、主要应用领域（如心脏、肺部、腹部成像）进行了全面梳理，并总结了通用的处理流程。
- 诸如《Meta-Learning With Medical Imaging and Health Informatics Applications》等专著 [21]，系统总结了元学习的理论及其在医学影像和健康信息学中的多样化应用，汇集了该领域顶尖专家的见解，并为研究人员提供了宝贵的资源。

早期应用与概念验证研究

- 最初将 FSL [22]、GANs [26] 和迁移学习 [4] 应用于特定医学图像分析任务（如皮肤病变分类 [22]、胸部 X 射线分析、肿瘤分割）的可行性研究，为后续更深入和复杂的研究奠定了基础。

这些里程碑的演进清晰地展示了研究焦点从最初的“能否用少量数据学习”，逐渐转向更深层次的实际应用考量，例如如何保护数据隐私（联邦学习的兴起）、如何提升生成数据的质量和可控性（从早期 GAN 到先进的扩散模型），以及如何满足医学领域的特定需求（如三维高分辨率图像合成）。这种发展轨迹表明，该领域正在不断成熟，从单纯追求技术可行性，迈向如何在真实世界的医疗约束条件下，实现可靠、安全且有效的 AI 应用。这也预示着未来的突破将更可能来自于能够整合不同技术优势、解决多方面挑战的综合性解决方案。

五、 产业化案例

数据稀缺和小样本学习不仅是学术研究的焦点，也是医学 AI 产业化进程中必须克服的障碍。众多企业正在积极探索和应用各种数据高效的学习策略，以期将 AI 技术成功转化为临床可用的产品和服务。

5.1 专注于医学影像 AI 的公司

这类公司通常利用深度学习技术开发辅助诊断、疾病筛查等工具。鉴于医学数据获取的普遍挑战，它们或显式或隐式地采用了数据高效的策略来训练模型。

- **PathAI:**
  - **技术与应用：** 作为一家领先的 AI 赋能病理学技术提供商[33]，PathAI 利用先进的机器学习算法，包括“基于机器学习的算法和无监督算法”[34]，来辅助病理医生进行更快速、更准确的诊断。其 AISight®平台是一个云原生的图像管理系统，集成了 PathAI 自身及第三方的 AI 算法[33]。PathAI 与生命科学公司（如 Discovery Life Sciences[33]、Precision for Medicine[34]）合作，将 AI 应用于生物样本分析、转化医学研究、生物标志物发现和临床试验服务，旨在提升生物样本分析的一致性和数据可靠性[34]。虽然公开资料未明确提及“小样本学习”，但其强调“增强多模态数据集”和“将大型复杂的生物标志物组减少为可扩展、可操作的生物标志物”[34]，暗示了其在处理数据受限或复杂数据环境时，采用了复杂的数据处理和特征提取技术。
  - **数据策略启示：** 通过合作获取多样化的数据集，并开发定制化的 AI 解决方案，

表明 PathAI 采取了将 AI 专业知识与特定数据资源相结合的策略，以应对普遍存在的数据稀缺问题。

- **Paige:**
  - **技术与应用：** Paige 致力于开发 AI 系统（如 Paige Prostate）以辅助病理医生进行癌症诊断 [35]。其发表的研究成果显示，AI 增强的病理诊断在提高前列腺癌诊断准确性和效率方面取得了显著成效 [35]。一项研究表明，Paige Prostate 的敏感性达到 99%，特异性达到 93%，并将诊断时间缩短了 65.5%，甚至识别出了一些最初被经验丰富的病理医生遗漏的癌症病例 [36]。
  - **数据策略启示：** Paige 的系统虽然依赖大规模数据集进行训练（从其临床验证的成功可以推断），但其带来的效率提升以及识别遗漏病例的能力，表明其模型具备强大的模式识别能力。这对于处理真实世界中每个患者样本呈现多样性且数据量可能有限的情况至关重要。将 AI 用作"质量控制工具" [36] 也反映了其利用 AI 从现有数据中最大化提取有价值信息的策略。
- **Aidoc:**
  - **技术与应用：** Aidoc 提供临床 AI 解决方案，能够自动分析医学影像，以优化工作流程、优先处理紧急病例，并在放射科、心脏科、神经血管科和血管科等多个专科领域激活医疗团队 [37]。其 aiOS™平台整合了多种 AI 工具 [37]。据报道，Aidoc 的解决方案带来了显著的临床效益，例如将肺栓塞 (PE) 患者的通知时间缩短了 31% [37]。该公司业务规模较大，已覆盖众多医疗中心，每月分析大量患者数据 [37]。
  - **数据策略启示：** Aidoc 在多个专科领域的成功应用，表明其拥有稳健的模型。这些模型很可能是通过多样化的数据集训练而成，并可能采用了数据高效的技术来处理不同解剖结构和病理特征的变异性。
- **Rad AI, Gleamer, Annalise.ai, Oxipit, DeepC 等:**
  - 这些公司同样致力于开发面向放射科的 AI 驱动工具，专注于自动化任务（如报告生成）、辅助诊断（如创伤、胸部 X 射线解读）以及 AI 工具的管理与集成 [38]。例如，Annalise.ai 强调使用由资深放射科医生手工标注的大规模高质量数据集来训练模型 [38]。

- **数据策略启示：** 这些公司的共同点在于利用 AI 从宝贵的、经过标注的医学影像中提取最大价值，以提升诊断效率和准确性。对“根据医疗需求定制 AI 模型” [38] 的需求，也间接反映了适应特定临床场景（这些场景的数据量可能有限）的必要性。

5.2 专注于数据解决方案的公司

这类公司提供的产品或服务直接旨在解决数据访问、隐私保护和数据量不足等问题，为医学 AI 的开发提供基础支撑。

- **NVIDIA Clara:**
  - **平台与技术：** NVIDIA Clara 是一个集计算平台、软件（如 MONAI, FLARE）和服务于一体的综合解决方案，面向医疗健康领域的 AI 应用，涵盖医学影像、数字健康和药物发现等 [31]。
  - **联邦学习 (FL)：** NVIDIA Clara Train SDK 支持联邦学习功能，允许多个机构在不共享原始患者数据的前提下协同训练 AI 模型 [12]。其采用的服务器-客户端架构通过聚合来自各客户端的模型更新来构建全局模型，同时保护数据隐私 [12]。这直接通过赋能安全的多机构数据利用，来应对数据访问受限和数据稀缺的挑战。
  - **影响：** 通过在遵守隐私法规的前提下利用更多样化的数据集，联邦学习有助于构建更稳健、泛化能力更强的 AI 模型。
- **MDClone:**
  - **技术与应用：** MDClone 提供一个动态数据探索和合成数据生成平台，能够从真实的医疗数据中创建统计学特征相似但不包含个人可识别信息 (PII) 的合成数据集 [40]。这使得研究人员可以在不泄露患者隐私的情况下进行数据分析、改进运营和促进合作 [44 (关于合成数据的一般性描述), 41]。
  - **案例：** 美国退伍军人健康管理局 (VHA) 等机构利用 MDClone 的平台来整合碎片化的数据，使非技术背景的研究人员也能进行自助式数据探索，分析如 COVID-19 的流行趋势、管理慢性疾病、利用基于合成数据的机器学习模型预测心力衰竭患者的再入院风险，并支持自杀预防等项目 [40]。
  - **影响：** MDClone 的技术显著提升了用于研究和质量改进的数据可及性，缩短了

从产生想法到获得洞察的时间，并通过使用保护隐私的合成数据，促进了与外部利益相关者的合作 [41]。

- **其他合成数据生成初创公司 (如 Mostly AI, Synthesis AI, Gretel.ai, Syntho, Datagen 等):**
  - 这些公司专注于为 AI 模型训练、软件测试和数据分析等目的创建合成数据，通常将隐私保护作为其核心价值主张之一 [42]。
  - **医疗相关性：** 其中，Gretel.ai 和开源项目 Synthea 等在医疗健康领域受到关注，它们能够生成反映真实患者数据特征的合成数据，同时遵守相关的隐私法规和数据标准（如 Synthea 支持 HL7 FHIR）[42]。
  - **潜力：** 这些公司提供的解决方案有望规模化地生成大规模、多样化的数据集，用于训练医学 AI 模型，从而克服真实数据可用性和隐私方面的限制。它们还可以帮助平衡数据集、覆盖边缘案例，提升模型鲁棒性 [44]。

5.3 产业化面临的挑战与机遇

尽管医学 AI 产业化前景广阔，但也面临诸多挑战，同时蕴藏着巨大机遇。

- **挑战：**
  - **临床验证与监管审批：** 证明 AI 产品在临床应用中的安全性、有效性和稳健性是一个漫长且耗资巨大的过程，需要通过 FDA（美国食品药品监督管理局）、CE（欧洲合格认证）等机构的严格审批。目前许多 AI 工具仍处于“仅供研究使用”(Research Use Only) 阶段 [33]。
  - **临床工作流程整合：** 将 AI 系统无缝集成到医院现有的 IT 基础设施（如电子健康记录 EHR、实验室信息系统 LIS、医学影像存档与通信系统 PACS）中，对于实际应用至关重要，但在技术上具有挑战性 [33]。
  - **建立信任与提升可解释性：** 临床医生需要信任 AI 的预测结果，这要求模型具有良好的可解释性，能够清晰地阐述其决策依据 [3]。
  - **数据治理与标准化：** 缺乏统一的数据格式、标准化的标注规范以及通畅的数据共享协议，阻碍了大规模 AI 的开发和部署。

- **成本与投资回报率 (ROI)：** 向医疗机构清晰地证明 AI 解决方案的经济效益和投资回报可能存在困难。
- **伦理考量：** 确保商业化 AI 产品的公平性、避免偏倚、以及持续保护患者隐私，是业界需要持续关注和解决的伦理问题 [6]。

- **机遇：**
  - **满足未被满足的临床需求：** 在许多临床需求尚未得到有效满足的领域，AI 技术有望通过改进诊断、治疗或提升效率来创造巨大价值。
  - **价值医疗的推动：** 能够改善患者预后、降低医疗成本的 AI 解决方案，与全球范围内向价值医疗转型的趋势高度契合。
  - **加速药物研发：** AI 技术，特别是结合合成数据和高级分析方法，有潜力显著加速新药的发现和临床试验过程 [28]。
  - **个性化医疗的实现：** 数据高效的 AI 技术能够基于个体患者数据实现更精准的个性化治疗方案。
  - **扩大优质医疗资源的可及性：** AI 工具有望增强医护人员的服务能力，尤其是在医疗资源相对匮乏的地区 [11]。

观察这些产业化案例，可以发现一个普遍现象：成功的医学 AI 企业往往采取双重策略，即不仅致力于开发先进的 AI 模型，同时也积极构建获取、管理或创造新型数据（如联邦数据、合成数据）的能力。单纯的技术优势往往不足以在竞争激烈的市场中立足，有效的数据策略才是关键。此外，医学 AI 的“产品”形态也日益从孤立的算法向集成的“解决方案”或“平台”演进 [33]。这些平台不仅包含 AI 模型，还整合了数据管理、工作流程优化以及多模态数据融合等功能，数据高效学习技术则作为底层支撑，处理平台中遇到的多样化且数据量可能有限的临床数据。这种趋势也催生了 AI 模型开发者与专业数据解决方案提供商（如合成数据公司、联邦学习平台提供商）之间日益紧密的共生关系。AI 公司可以借助第三方的数据解决方案来克服自身的数据瓶颈，从而更专注于核心算法的研发和临床应用的推广，这有望通过专业化分工加速整个行业的创新步伐。

下表总结了一些在医学 AI 领域活跃的、并采用数据高效策略的代表性企业：

表 4: 医学 AI 产业主要参与者及其数据高效策略

公司名称	核心技术/方法 (与数据效率相关)	医疗健康领域应用案例	已报告的影响/效益 (部分)	应对数据稀缺的策略
PathAI	AI 驱动的数字病理学，机器学习，无监督学习，AISight®图像管理平台 [33]	辅助病理诊断，生物标志物发现，临床试验服务，转化医学研究。	提升诊断速度和准确性，增强生物样本分析的一致性和数据可靠性 [33]。	与生命科学公司合作获取数据，增强多模态数据集分析能力。
Paige	AI 辅助癌症诊断系统 (如 Paige Prostate) [35]	前列腺癌、乳腺癌等癌症的病理诊断。	显著提升诊断准确率和效率，减少漏诊，缩短诊断时间 (如 Paige Prostate 使诊断时间减少 65.5%) [35]。	利用大规模高质量标注数据训练模型，AI 作为质量控制工具最大化数据价值。
Aidoc	aiOS™临床 AI 平台，自动医学影像分析 [37]	跨越多专科（放射、心脏、神经血管等）的影像辅助诊断，工作流程优化，紧急病例优先处理。	显著缩短紧急病例（如 PE）的通知时间 (31%) [37]。	构建统一 AI 平台处理多样化数据，模型需适应不同专科和病种的数据特点。
NVIDIA	Clara 平台 (含 MONAI, FLARE 等软件)，GPU 加速计算 [12]	提供联邦学习框架，支持医学影像 AI 模型开发、药物发现、基因	使多机构能在保护隐私前提下协同训练模型，提升模型鲁棒性和	大力推广联邦学习技术，提供软硬件一体化解决方案。

		组学分析。	泛化能力 [12]。	
MDClone	动态数据探索平台，合成数据生成引擎 [40]	临床研究，运营改进，公共卫生监测 (如 VHA 用于 COVID-19 分析、慢病管理、自杀预防)。	显著提升数据可及性，赋能非技术用户进行数据分析，支持隐私保护下的多方协作 [41]。	核心业务即生成和利用合成数据，解决真实数据使用限制。

医学 AI 产业在数据稀缺背景下的发展，正驱动着 AI 技术、数据工程、隐私增强技术和临床工作流程整合的深度融合。那些能够成功驾驭这种复杂互动，提供不仅智能而且实用、可信、可扩展并能融入现有医疗体系的解决方案的企业，更有可能在未来取得成功。与此同时，相关的伦理和监管框架也将与这些产业实践共同演进。

六、 总结与展望

核心挑战与应对策略回顾

数据稀缺以及由此引发的对小样本学习技术的需求，是当前医学人工智能领域面临的根本性且持续存在的挑战。其根源在于严格的患者隐私保护法规、高昂的数据采集与标注成本、以及医疗数据固有的异质性和复杂性。

为应对这些挑战，研究和产业界已探索并实践了多种策略，主要可以归纳为：

1. **学习算法层面：** 开发旨在从有限样本中最大化知识提取的学习算法，如小样本学习 (FSL)、元学习 (Meta-Learning) 和迁移学习 (Transfer Learning)。
2. **数据中心层面：** 采用数据增强 (Data Augmentation) 和合成数据生成 (Synthetic Data Generation) 等技术，以扩充或丰富可用的数据集。
3. **协作与隐私保护层面：** 构建如联邦学习 (Federated Learning) 这样的框架，以在保护隐私的前提下利用分布式数据资源。

尽管这些策略已取得显著进展，但每种方法都伴随着其固有的瓶颈，例如 FSL 的泛化能

力、合成数据的验证难题、迁移学习的领域漂移问题、以及普遍存在的模型可解释性不足等。

未来研究方向展望

展望未来，为进一步突破医学 AI 在数据稀缺环境下的发展瓶颈，以下研究方向值得重点关注：

- **更鲁棒和泛化能力更强的 FSL/元学习算法：** 开发能够在更少样本条件下，跨越更多样化的医学任务和数据集实现优异泛化性能的算法。
- **高保真、可控且经过充分验证的合成数据：** 持续推进生成模型（特别是扩散模型）的发展，以期生成在所有相关方面（包括细微病理特征和统计分布）均与真实数据难以区分的合成医学数据。同时，需要实现对生成过程的有效控制（例如，按需生成特定病理类型的样本），并建立一套严格的验证标准，以确保合成数据在下游任务中的有效性，并避免引入或放大偏倚 [6]。
- **提升模型的可解释性与可信度 (XAI)：** 针对数据高效学习技术和复杂的医学数据，创建定制化的 XAI 方法，以增强模型的透明度，帮助临床医生理解和信任 AI 的决策过程，并为模型的错误分析和改进提供依据 [7]。
- **有限数据下的多模态数据融合：** 发展能够有效整合来自不同来源信息（如医学影像、电子健康记录、基因组学数据等）的技术，即便某些模态的数据量非常稀缺。
- **面向医学预训练的大规模自监督学习：** 利用海量的未标注医学数据，通过自监督学习范式预训练强大的基础模型，学习普适性的医学特征表示，这些模型随后可以利用小样本学习技术针对特定下游任务进行高效微调。
- **混合策略与集成方法：** 探索将不同技术的优势相结合的混合方法，例如，利用合成数据进行大规模预训练，再结合 FSL 在少量真实数据上进行微调；或者将联邦学习与差分隐私等更强的隐私保护技术相结合，以实现更安全的数据协作。
- **伦理框架与治理机制的完善：** 针对数据高效学习技术的应用，特别是涉及合成数据（如偏倚、隐私泄露、再识别风险 [6]）和联邦学习（如安全性、公平性）的场景，制定清晰的伦理指南和健全的治理结构。
- **标准化基准与评估体系的建立：** 创建面向医学 AI 数据高效学习的综合性基准数据集

和标准化评估协议，以促进不同方法之间的公平比较，并客观追踪领域进展 [30]。

长远愿景

医学 AI 领域的长远愿景是构建能够像人类专家一样，在真实临床环境中持续学习和自适应演进的 AI 系统，即使面对稀疏或不断变化的数据，也能有效增强人类的专业能力，最终改善全球患者的健康福祉。成功克服数据稀缺的挑战，将有望解锁 AI 在更广泛疾病谱系和更多样化患者人群中的应用潜力，为实现更公平、更精准、更高效的个性化医疗奠定坚实基础 [2]。

这一过程也预示着从单纯追求"大数据"向着"智能数据"利用的范式转变。未来，数据的质量、多样性以及对其进行智能化处理的能力（即使数据量有限）将变得至关重要。这意味着需要发展出更像人类学习方式的 AI——高效、自适应，并能从多样化的经验中汲取知识。同时，伦理考量将日益成为医学 AI 技术发展和应用的核心驱动力与制约因素 [5]。那些无法满足严格伦理和安全标准的技术，无论其技术本身多么先进，都可能难以获得监管批准或临床的广泛接受。因此，成功应对医学 AI 的数据稀缺挑战，不仅是一项技术攻关，更是一项涉及 AI 研究者、临床医生、伦理学家、政策制定者和患者共同参与的社会技术工程，目标是构建一个 AI 能够安全、有效且公平地惠及所有人的未来。

七、 参考文献

1 Techscience Press. (n.d.). Dealing with data scarcity is the biggest challenge faced by Artificial Intelligence (AI). CMC, Tech Science Press. Retrieved from https://www.techscience.com/cmc/online/detail/23120/pdf

11 CBS Interactive. (n.d.). AI will end scarcity of medical expertise, Bill Gates says. Becker's Hospital Review. Retrieved from https://www.beckershospitalreview.com/disruptors/ai-will-end-scarcity-of-medical-expertise-bill-gates-says/

4 Andrearczyk, V., Oreiller, V., Błaszczyński, K., Depeursinge, A., & Müller, H. (2024). Few-shot learning with deep neural networks for inference from medical images. Frontiers in Radiology, 4. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC11540231/

2 Pachetti, G., & Colantonio, S. (2023). A Systematic Review of Few-Shot Learning in Medical Imaging. arXiv. Retrieved from https://www.researchgate.net/publication/374557473_A_Systematic_Review_of_Few-Shot_Learning_in_Medical_Imaging

3 Consensus. (2024). Limitations of Transfer Learning in Medical Imaging. Consensus.app. Retrieved from https://consensus.app/search/what-are-the-limitations-of-transfer-learning-in-d/yO2muql-ROOS9PW8p48JGg/
9 ResearchGate. (n.d.). Figure of the advantages and disadvantages of transfer learning for medical images. Retrieved from https://www.researchgate.net/figure/of-the-advantages-and-disadvantages-of-transfer-learning-for-medical-images_fig4_382902546
21 Amazon. (n.d.). Meta-Learning With Medical Imaging and Health Informatics Applications. Retrieved from https://www.amazon.com/Meta-Learning-Medical-Imaging-Informatics-Applications/dp/0323998518
7 ResearchGate. (n.d.). Meta-learning for Medical Image Segmentation: Uncertainty Quantification. Retrieved from https://www.researchgate.net/publication/362019758_Meta-learning_for_Medical_Image_Segmentation_Uncertainty_Quantification
18 Gort, P. C., Claessens, C. H. B., de With, P. H. N., & van der Sommen, F. (2025). Evaluating the interpretability of prototype networks for medical image analysis. Proceedings of SPIE, 13406. Retrieved from https://www.spiedigitallibrary.org/conference-proceedings-of-spie/13406/134061I/Evaluating-the-interpretability-of-prototype-networks-for-medical-image-analysis/10.1117/12.3046678.short
17 Liu, Y., et al. (2025). Decentralized Prototypical Contrastive Network for Medical Image Classification. Medical Physics. Retrieved from https://pubmed.ncbi.nlm.nih.gov/40089972/
24 Milvus. (n.d.). How is Data Augmentation Used in Medical Imaging? Retrieved from https://milvus.io/ai-quick-reference/how-is-data-augmentation-used-in-medical-imaging
25 Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., & Haworth, A. (2018). A review of medical image data augmentation techniques for deep learning applications. Journal of Medical Imaging and Radiation Oncology, 62(5), 597-611. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC5977656/
13 Nøhr, A. K., et al. (2023). Generative adversarial networks in medical imaging: A review. Medical Image Analysis, 84, 102706. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC10055771/
26 International Journal of Advanced Research in Medical Sciences. (2024). A Comprehensive Review of Generative Adversarial Networks for Synthetic Medical Image Generation. Retrieved from https://ijarm.com/pdfcopy/2024/jan2024/ijarm9.pdf
23 arXiv. (2025). Text-to-Image Synthesis in the Medical Domain: A Comparative Study of Latent Diffusion Models. Retrieved from https://www.arxiv.org/abs/2505.05573
27 Friedrich, P., et al. (2024). WDM: 3D Wavelet Diffusion Models for High-Resolution Medical Image Synthesis. arXiv. Retrieved from https://arxiv.org/abs/2402.19043
14 Papers With Code. (n.d.). Prototypical Networks for Few-shot Learning. Retrieved from https://paperswithcode.com/paper/prototypical-networks-for-few-shot-learning
15 Snell, J., Swersky, K., & Zemel, R. S. (2017). Prototypical Networks for Few-shot Learning. arXiv. Retrieved from https://arxiv.org/abs/1703.05175
19 Finn, C., Abbeel, P., & Levine, S. (2017). Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. SciSpace. Retrieved from https://scispace.com/papers/model-agnostic-

meta-learning-for-fast-adaptation-of-deep-1uogjkn6mb
20 Papers With Code. (n.d.). Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. Retrieved from https://paperswithcode.com/paper/model-agnostic-meta-learning-for-fast
21 Nguyen, H. V., Summers, R., & Chellappa, R. (Eds.). (2022). Meta Learning With Medical Imaging and Health Informatics Applications (The MICCAI Society book Series). Amazon.com. Retrieved from https://www.amazon.com/Meta-Learning-Medical-Imaging-Informatics-Applications/dp/0323998518
30 MICCAI Society. (n.d.). OPEN DATA 2025 - MICCAI. Retrieved from https://conferences.miccai.org/2025/en/OPEN-DATA.html
38 AI Superior. (n.d.). Best AI Radiology Companies Revolutionizing Healthcare. Retrieved from https://aisuperior.com/ai-radiology-companies/
37 Aidoc. (n.d.). Aidoc | Clinical AI Company | Rapid Responses, Smarter Care. Retrieved from https://www.aidoc.com/
42 AI Superior. (n.d.). Top Synthetic Data Generation Companies Powering AI Innovation. Retrieved from https://aisuperior.com/synthetic-data-generation-for-ai-companies/
43 Seedtable. (2025). 42 Best Synthetic Data Startups to Watch in 2025. Retrieved from https://www.seedtable.com/best-synthetic-data-startups
33 PathAI. (2025, January 23). PathAI Partners with Discovery Life Sciences to Deploy First AI-Powered Biospecimen Solutions. Retrieved from https://www.pathai.com/resources/pathai-partners-with-discovery-life-sciences-to-deploy-first-ai-powered-biospecimen-solutions/
34 PathAI. (2025, April 25). Precision for Medicine and PathAI Announce Strategic Collaboration to Advance AI-Powered Clinical Trial Services and Biospecimen Products. Retrieved from https://www.pathai.com/resources/precision-for-medicine-and-pathai-announce-strategic-collaboration-to-advance-ai-powered-clinical-trial-services-and-biospecimen-products/
35 Paige.ai. (n.d.). Publications. Retrieved from https://www.paige.ai/publications
36 Paige.ai. (n.d.). Independent real-world application of a clinical-grade automated prostate cancer detection system. Retrieved from https://www.paige.ai/publications/independent-real-world-application-of-a-clinical-grade-automated-prostate-cancer-detection-system
39 NVIDIA. (n.d.). NVIDIA Clara | AI-powered Solutions for Healthcare. Retrieved from https://www.nvidia.com/en-us/clara/
29 NVIDIA Developer. (2020, June 18). NVIDIA Clara Federated Learning [Video]. YouTube. Retrieved from https://www.youtube.com/watch?v=bVU-Ea6hc0k
22 MDPI. (2024). A Vision Transformer-Based Metric Learning Model for Brain Tumor Classification with Few-Shot Learning. Appl. Sci., 14(9), 1863. Retrieved from https://www.mdpi.com/2079-9292/14/9/1863
32 ResearchGate. (n.d.). A systematic review of few-shot learning in medical imaging. Retrieved from https://www.researchgate.net/publication/383191522_A_systematic_review_of_few-shot_learning_in_medical_imaging

8 University of Reading. (n.d.). Challenges of Deep Learning in Medical Image Analysis – Improving Explainability and Trust. Retrieved from https://centaur.reading.ac.uk/109789/1/Challenges%20of%20Deep%20Learning%20in%20Medical%20Image%20Analysis%20%E2%80%93%20Improving%20Explainability%20and%20Trust%20_%20Full%20Text.pdf
10 AlAhmad, A., et al. (2024). Medical Image Segmentation Using Model-Agnostic Meta-Learning with U-Net. Applied Sciences, 14(11), 4697. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC11202447/
5 AuxilioBits. (n.d.). Synthetic Data Generation for Healthcare AI Training: Techniques and Privacy Considerations. Retrieved from https://www.auxiliobits.com/blog/synthetic-data-generation-for-healthcare-ai-training-techniques-and-privacy-considerations/
6 Keymakr. (n.d.). Ethical and Legal Considerations of Synthetic Data Usage. Retrieved from https://keymakr.com/blog/ethical-and-legal-considerations-of-synthetic-data-usage/
21 Amazon. (n.d.). Meta-Learning With Medical Imaging and Health Informatics Applications (The MICCAI Society book Series). Retrieved from https://www.amazon.com/Meta-Learning-Medical-Imaging-Informatics-Applications/dp/0323998518 21
16 MICCAI. (2024). CeLDA: Age-Inclusive Cephalometric Landmark Detection via Holistic Prototype Learning and Relation Mining. Retrieved from https://papers.miccai.org/miccai-2024/paper/0737_paper.pdf
47 Electropages. (2025, April). How AI Works: Neural Networks in Real-World Use. Retrieved from https://www.electropages.com/blog/2025/04/how-ai-works-neural-networks-real-world-use
28 MindInventory. (n.d.). Generative AI in Healthcare: Applications, Benefits & Challenges. Retrieved from https://www.mindinventory.com/blog/generative-ai-in-healthcare/
45 Reddit. (n.d.). I don't understand why people talk about synthetic data for LLMs. Retrieved from https://www.reddit.com/r/learnmachinelearning/comments/1k2foyt/i_dont_understand_why_people_talk_about_synthetic/
44 Syntheticus. (n.d.). Democratizing Data Access with Synthetic Data [Whitepaper]. Retrieved from https://syntheticus.ai/hubfs/Resources%20-%20PDFs/Syntheticus%20Whitepaper_Democratizing%20Data%20Access%20with%20Synthetic%20Data.pdf
12 NVIDIA Developer. (2019, December 1). Federated Learning with NVIDIA Clara. Retrieved from https://developer.nvidia.com/blog/federated-learning-clara/
31 NVIDIA Developer. (n.d.). Tag: Federated Learning. Retrieved from https://developer.nvidia.com/blog/tag/federated-learning/
40 MDClone. (n.d.). Real-World Scenarios. Retrieved from https://mdclone.com/product-resources/real-world-scenarios/
41 MDClone. (2024, February). CASE STUDY: Veterans Health Administration. Retrieved from https://www.mdclone.com/wp-content/uploads/2024/02/VHA_Case_Study___MDClone.pdf
22 MDPI. (2024). A Vision Transformer-Based Metric Learning Model for Brain Tumor

Classification with Few-Shot Learning. Appl. Sci., 14(9), 1863. 22
46 PMC. (n.d.). Interpretability in Machine Learning for Medical Imaging: A Framework. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC11486155/
1 Techscience Press. (n.d.). Dealing with data scarcity is the biggest challenge faced by Artificial Intelligence (AI). CMC, Tech Science Press. Retrieved from https://www.techscience.com/cmc/online/detail/23120/pdf 1
13 PMC. (n.d.). Generative adversarial networks in medical imaging: A review. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC10055771/ 13
15 Snell, J., Swersky, K., & Zemel, R. S. (2017). Prototypical Networks for Few-shot Learning. arXiv. Retrieved from https://arxiv.org/abs/1703.05175 15
19 Finn, C., Abbeel, P., & Levine, S. (2017). Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. SciSpace. Retrieved from https://scispace.com/papers/model-agnostic-meta-learning-for-fast-adaptation-of-deep-1uogjkn6mb 19
35 Paige.ai. (n.d.). Publications. Retrieved from https://www.paige.ai/publications 35
34 PathAI. (2025, April 25). Precision for Medicine and PathAI Announce Strategic Collaboration to Advance AI-Powered Clinical Trial Services and Biospecimen Products. Retrieved from https://www.pathai.com/resources/precision-for-medicine-and-pathai-announce-strategic-collaboration-to-advance-ai-powered-clinical-trial-services-and-biospecimen-products/ 34
12 NVIDIA Developer. (2019, December 1). Federated Learning with NVIDIA Clara. Retrieved from https://developer.nvidia.com/blog/federated-learning-clara/ 12

引用的著作

1. The Future of Artificial Intelligence in the Face of Data Scarcity, 访问时间为 六月 11, 2025, https://www.techscience.com/cmc/online/detail/23120/pdf
2. A Systematic Review of Few-Shot Learning in Medical Imaging - ResearchGate, 访问时间为 六月 11, 2025, https://www.researchgate.net/publication/374557473_A_Systematic_Review_of_Few-Shot_Learning_in_Medical_Imaging
3. What are the limitations of transfer learning in disease diagnosis ..., 访问时间为 六月 11, 2025, https://consensus.app/search/what-are-the-limitations-of-transfer-learning-in-d/yO2muqI-ROOS9PW8p48JGg/
4. Few-shot learning for inference in medical imaging with subspace feature representations, 访问时间为 六月 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11540231/
5. Synthetic Data for Healthcare AI: Methods & Privacy Insights - Auxiliobits, 访问时间为 六月 11, 2025, https://www.auxiliobits.com/blog/synthetic-data-generation-for-healthcare-ai-training-techniques-and-privacy-considerations/

6. Ethical and Legal Considerations of Synthetic Data Usage - Keymakr, 访问时间为 六月 11, 2025, https://keymakr.com/blog/ethical-and-legal-considerations-of-synthetic-data-usage/
7. Meta-learning for Medical Image Segmentation Uncertainty Quantification | Request PDF, 访问时间为 六月 11, 2025, https://www.researchgate.net/publication/362019758_Meta-learning_for_Medical_Image_Segmentation_Uncertainty_Quantification
8. Challenges of deep learning in medical image analysis – improving explainability and trust - CentAUR, 访问时间为 六月 11, 2025, https://centaur.reading.ac.uk/109789/1/Challenges%20of%20Deep%20Learning%20in%20Medical%20Image%20Analysis%20%E2%80%93%20Improving%20Explainability%20and%20Trust%20_%20Full%20Text.pdf
9. of the advantages and disadvantages of transfer learning for medical images - ResearchGate, 访问时间为 六月 11, 2025, https://www.researchgate.net/figure/of-the-advantages-and-disadvantages-of-transfer-learning-for-medical-images_fig4_382902546
10. Few-Shot Learning for Medical Image Segmentation Using 3D U-Net and Model-Agnostic Meta-Learning (MAML) - PMC, 访问时间为 六月 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11202447/
11. AI will end scarcity of medical expertise, Bill Gates says - Becker's Hospital Review | Healthcare News & Analysis, 访问时间为 六月 11, 2025, https://www.beckershospitalreview.com/disruptors/ai-will-end-scarcity-of-medical-expertise-bill-gates-says/
12. Federated Learning powered by NVIDIA Clara | NVIDIA Technical ..., 访问时间为 六月 11, 2025, https://developer.nvidia.com/blog/federated-learning-clara/
13. GANs for Medical Image Synthesis: An Empirical Study - PMC, 访问时间为 六月 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10055771/
14. Prototypical Networks for Few-shot Learning - Papers With Code, 访问时间为 六月 11, 2025, https://paperswithcode.com/paper/prototypical-networks-for-few-shot-learning
15. Prototypical Networks for Few-shot Learning, 访问时间为 六月 11, 2025, https://arxiv.org/abs/1703.05175
16. Cephalometric Landmark Detection across Ages with Prototypical Network - MICCAI, 访问时间为 六月 11, 2025, https://papers.miccai.org/miccai-2024/paper/0737_paper.pdf
17. Decentralized learning for medical image classification with prototypical contrastive network, 访问时间为 六月 11, 2025, https://pubmed.ncbi.nlm.nih.gov/40089972/

18. Evaluating the interpretability of prototype networks for medical image analysis, 访问时间为 六月 11, 2025, https://www.spiedigitallibrary.org/conference-proceedings-of-spie/13406/134061I/Evaluating-the-interpretability-of-prototype-networks-for-medical-image-analysis/10.1117/12.3046678.short
19. (PDF) Model-agnostic meta-learning for fast adaptation of deep ..., 访问时间为 六月 11, 2025, https://scispace.com/papers/model-agnostic-meta-learning-for-fast-adaptation-of-deep-1uogjkn6mb
20. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks | Papers With Code, 访问时间为 六月 11, 2025, https://paperswithcode.com/paper/model-agnostic-meta-learning-for-fast
21. Meta Learning With Medical Imaging and Health Informatics Applications (The MICCAI Society book Series) - Amazon.com, 访问时间为 六月 11, 2025, https://www.amazon.com/Meta-Learning-Medical-Imaging-Informatics-Applications/dp/0323998518
22. Metric-Based Meta-Learning Approach for Few-Shot Classification of Brain Tumors Using Magnetic Resonance Images - MDPI, 访问时间为 六月 11, 2025, https://www.mdpi.com/2079-9292/14/9/1863
23. Prompt to Polyp: Medical Text-Conditioned Image Synthesis with Diffusion Models - arXiv, 访问时间为 六月 11, 2025, https://www.arxiv.org/abs/2505.05573
24. How is data augmentation used in medical imaging? - Milvus, 访问时间为 六月 11, 2025, https://milvus.io/ai-quick-reference/how-is-data-augmentation-used-in-medical-imaging
25. Differential Data Augmentation Techniques for Medical Imaging Classification Tasks - PMC, 访问时间为 六月 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC5977656/
26. Developing GANs for Synthetic Medical Imaging Data: Enhancing Training and Research, 访问时间为 六月 11, 2025, https://ijarm.com/pdfcopy/2024/jan2024/ijarm9.pdf
27. WDM: 3D Wavelet Diffusion Models for High-Resolution Medical Image Synthesis - arXiv, 访问时间为 六月 11, 2025, https://arxiv.org/abs/2402.19043
28. Generative AI in Healthcare: A Comprehensive Guide for Decision Makers - MindInventory, 访问时间为 六月 11, 2025, https://www.mindinventory.com/blog/generative-ai-in-healthcare/
29. NVIDIA Clara Federated Learning - YouTube, 访问时间为 六月 11, 2025, https://www.youtube.com/watch?v=bVU-Ea6hc0k
30. OPEN DATA 2025 - MICCAI, 访问时间为 六月 11, 2025, https://conferences.miccai.org/2025/en/OPEN-DATA.html

31. Tag: Federated Learning | NVIDIA Technical Blog, 访问时间为 六月 11, 2025, https://developer.nvidia.com/blog/tag/federated-learning/
32. A systematic review of few-shot learning in medical imaging | Request PDF - ResearchGate, 访问时间为 六月 11, 2025, https://www.researchgate.net/publication/383191522_A_systematic_review_of_few-shot_learning_in_medical_imaging
33. PathAI Partners with Discovery Life Sciences to Deploy First AI-Powered Biospecimen Solutions, 访问时间为 六月 11, 2025, https://www.pathai.com/resources/pathai-partners-with-discovery-life-sciences-to-deploy-first-ai-powered-biospecimen-solutions/
34. Precision for Medicine and PathAI Announce Strategic Collaboration ..., 访问时间为 六月 11, 2025, https://www.pathai.com/resources/precision-for-medicine-and-pathai-announce-strategic-collaboration-to-advance-ai-powered-clinical-trial-services-and-biospecimen-products/
35. Publications — Paige.ai, 访问时间为 六月 11, 2025, https://www.paige.ai/publications
36. Independent real-world application of a clinical-grade automated prostate cancer detection system - Paige.ai, 访问时间为 六月 11, 2025, https://www.paige.ai/publications/independent-real-world-application-of-a-clinical-grade-automated-prostate-cancer-detection-system
37. Aidoc | Clinical AI Company | Rapid Responses, Smarter Care, 访问时间为 六月 11, 2025, https://www.aidoc.com/
38. Best AI Radiology Companies Revolutionizing Healthcare - AI Superior, 访问时间为 六月 11, 2025, https://aisuperior.com/ai-radiology-companies/
39. NVIDIA Clara | AI-powered Solutions for Healthcare | NVIDIA, 访问时间为 六月 11, 2025, https://www.nvidia.com/en-us/clara/
40. Real-World Scenarios - MDClone, 访问时间为 六月 11, 2025, https://mdclone.com/product-resources/real-world-scenarios/
41. Veterans Health Administration - MDClone, 访问时间为 六月 11, 2025, https://www.mdclone.com/wp-content/uploads/2024/02/VHA_Case_Study___MDClone.pdf
42. Top Synthetic Data Generation Companies Powering AI Innovation, 访问时间为 六月 11, 2025, https://aisuperior.com/synthetic-data-generation-for-ai-companies/
43. 42 Best Synthetic Data Startups to Watch in 2025 - Seedtable, 访问时间为 六月 11, 2025, https://www.seedtable.com/best-synthetic-data-startups

44. Whitepaper-Democratizing Data Access with Synthetic Data - Syntheticus, 访问时间为 六月 11, 2025, https://syntheticus.ai/hubfs/Resources%20-%20PDFs/Syntheticus%20Whitepaper_Democratizing%20Data%20Access%20with%20Synthetic%20Data.pdf
45. I don't understand why people talk about synthetic data. Aren't you just looping your model's assumptions? : r/learnmachinelearning - Reddit, 访问时间为 六月 11, 2025, https://www.reddit.com/r/learnmachinelearning/comments/1k2foyt/i_dont_understand_why_people_talk_about_synthetic/
46. A Framework for Interpretability in Machine Learning for Medical Imaging - PubMed Central, 访问时间为 六月 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11486155/
47. How AI Works: From Neural Networks to Real-World Use - Electropages, 访问时间为 六月 11, 2025, https://www.electropages.com/blog/2025/04/how-ai-works-neural-networks-real-world-use