arXiv:2305.05247v1 [cs.LG] 9 May 2023

利用生成式人工智能模型在医疗保健领域进行合成数据生成：平衡研究与隐私

Aryan Jadon圣何塞州立大学美国加利福尼亚州
aryan.jadon@sjsu.edu

Shashank Kumar佛罗里达大学
美国佛罗里达州
sh.kumar@ufl.edu

摘要—电子健康记录和数字医疗数据的广泛应用催生了对数据驱动洞见的需求，以改善患者结局、诊断和治疗。然而，使用真实患者数据会带来隐私和监管挑战，包括遵守 HIPAA [1] 和 GDPR [2]。利用生成对抗网络 [3] 和变分自编码器 [4], 等生成式人工智能模型进行合成数据生成，为平衡宝贵的数据访问与患者隐私保护提供了有前景的解决方案。本文探讨了用于创建逼真匿名化患者数据的生成式人工智能模型，以供研究和训练 [5],；同时探索合成数据在医疗保健领域的应用，并讨论其优势、挑战及未来研究方向。合成数据有望通过提供匿名化患者数据，在保护隐私的同时支持多种应用，从而革新医疗保健。

索引词——**生成式人工智能模型、合成数据生成、医疗保健研究、患者隐私、数据增强、联邦学习、差分隐私、数据匿名化、迁移学习、健康信息学。**

I. 引言

医疗保健数据的快速数字化以及电子健康记录（EHR）的日益普及，为利用数据驱动的洞见来改善患者护理、诊断和治疗开辟了新的机遇 [6]。然而，使用真实患者数据往往会引发对隐私以及是否符合《健康保险流通与责任法案》（HIPAA）[1] 和《通用数据保护条例》（GDPR） [2]等数据保护法规的担忧。这些担忧为研究人员和医疗专业人员设置了障碍，他们需要访问大量数据，以开发和验证先进的算法与人工智能模型 [7]。

由生成式人工智能模型驱动的合成数据生成，为平衡数据驱动洞察的需求与患者隐私提供了一个颇具前景的解决方案。生成式人工智能模型，例如生成对抗网络 (GANs) [3] 和变分自编码器 (VAEs) [4],，能够学习真实世界数据的内在结构和分布，从而生成具有相似特征的新合成数据。通过创建逼真的匿名化患者数据，这些模型既能确保敏感患者信息得到保护，又能为研究人员提供用于分析和训练的宝贵数据[5]。

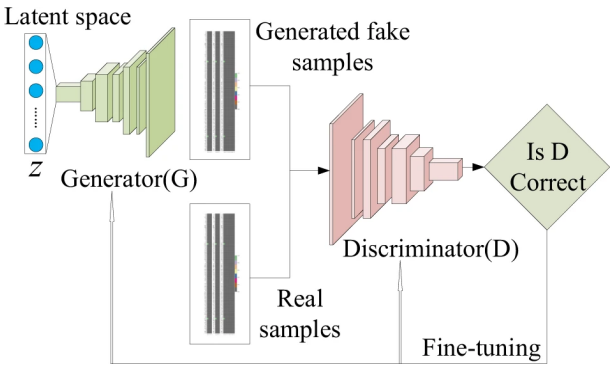

图 1. 基于生成对抗网络（GAN）的高效采样 [8]

在本文中，我们探讨生成式人工智能模型在生成合成患者数据中的作用，并讨论其在医疗保健领域的潜在应用、优势和挑战。我们还介绍未来的研究方向，以及合成数据对医疗保健研究和实践可能产生的影响，包括将其与差分隐私 [9],联邦学习 [10]和数据匿名化 [11]等其他保护隐私的技术相结合。本研究全面考察了使用生成式人工智能模型进行合成数据生成，旨在为理解这种方法在医疗保健领域的价值及其彻底革新研究、诊断和治疗的潜力奠定基础，同时保护患者隐私。我们的 GitHub 代码仓库位于
h
t
t
p
s://
github.com/aryan-jadon/Synthetic-Data-Medical-Generative-AI。

II. 用于合成数据生成的生成式人工智能模型

生成式人工智能模型（包括生成对抗网络（GANs）[3]和变分自编码器（VAEs） [4], ）旨在学习真实世界数据的内在结构和分布，并生成具有相似特征的新合成样本。这些模型已在包括医疗保健在内的各个领域成功生成高质量合成数据，表现出卓越的成果。

1)生成对抗网络（GANs）：生成对抗网络由两个相互竞争、进行极小极大博弈的神经网络组成：生成器和判别器。生成器创建合成数据样本，而判别器则区分真实样本和生成样本。生成器通过尝试欺骗判别器来改进输出，从而生成越来越逼真的合成数据。

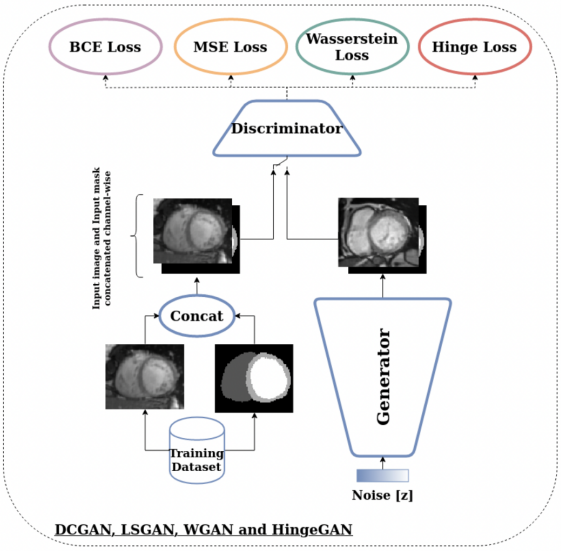

图2。各种GAN的代表性架构 [12]

2) 变分自编码器（VAEs）：VAEs是一类将自编码器与变分推断相结合的生成模型。它们学习数据与潜在空间之间的概率映射，通过从潜在空间中采样来生成新的数据样本。VAEs已被用于生成逼真的合成数据，同时在数据保真度与多样性之间保持平衡。

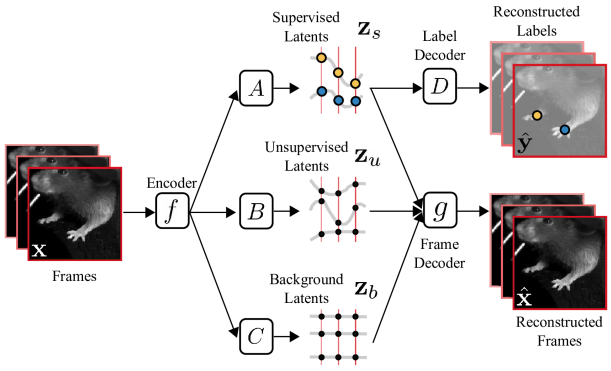

图3。变分自编码器在医学实验中的应用 [13]

A. 合成数据生成过程

使用生成式人工智能模型生成合成数据的过程涉及三个主要步骤：

1) 基于真实世界数据训练生成模型：该模型使用真实患者数据集进行训练，使其能够学习数据中存在的底层结构、关系和分布。

2)生成具有相似特征的新合成实例：训练完成后，生成模型可以创建与真实数据高度相似的合成数据样本，同时保留其底层关系和模式。此过程可确保生成的合成数据既逼真又经过匿名化处理。

3) 评估合成数据的质量和实用性：应根据生成的合成数据与真实数据的相似程度、维持底层关系和模式的能力，以及其在预期应用（如研究或人工智能模型训练）中的实用性，对其进行评估。

通过利用生成式人工智能模型生成合成数据，医疗保健研究人员和专业人员能够获取真实、匿名化的患者数据，同时应对患者隐私和法规合规方面的顾虑。

III. 合成数据在医疗保健领域的应用

利用生成式人工智能模型生成的合成数据在医疗保健领域有广泛的应用 [14]，使研究人员和专业人员能够获取真实、匿名化的患者数据，同时保护隐私并遵守数据保护法规。其主要应用包括：

A. 人工智能模型训练与验证

获取大规模、多样化的数据集对于训练和验证医疗保健领域的人工智能模型至关重要。合成数据为真实患者数据提供了一种替代方案，使研究人员能够开发和评估算法，而无需承担暴露敏感信息的风险。

- **数据增强：** 合成数据可用于扩充现有的真实世界数据集，尤其是在可用数据稀缺或类别不平衡时。这种增强能够提升人工智能模型在不同患者群体和临床场景中的性能与泛化能力 [15]。

- **模型验证：** 通过生成与真实世界数据高度相似的合成数据，研究人员可以验证人工智能模型的性能，并确保其在真实世界环境中稳健可靠 [16]。

B. 用于医学培训和决策支持的模拟

合成数据可用于创建逼真的模拟场景，支持医学培训和决策支持系统，使医疗专业人员能够在不冒伤害真实患者风险的情况下练习并提升技能。

- **医学培训：** 合成数据可用于开发虚拟患者病例和情境，供医学生、住院医师及其他医疗专业人员进行训练。这些模拟为学习和练习临床技能、诊断及治疗规划提供了安全且可控的环境 [17]。

- **决策支持系统：**合成数据可整合到临床决策支持系统中，根据对匿名化患者数据的分析提供实时指导和建议。这种方法有助于医疗专业人员做出更明智、更个性化的治疗决策，最终改善患者结局。 [18]

C. 医疗保健研究

合成数据使研究人员能够开展大规模研究和分析，而无需访问真实的患者数据，从而降低隐私泄露的风险，并确保符合数据保护法规。

- **流行病学研究：**研究人员可以利用合成数据研究健康相关疾病及病症的分布、决定因素和结果。这种方法能够为各种健康问题的风险因素和预防措施提供宝贵见解，最终为公共卫生政策和干预措施提供依据 [19]。

- **临床试验：**合成数据可用于模拟临床试验中的患者群体、治疗组和结果。这种方法有助于研究人员优化试验设计、估算干预措施的潜在疗效和安全性，并识别潜在的偏倚和混杂因素 [20]。

通过在这些不同应用中利用合成数据，医疗保健研究人员和专业人员能够获得宝贵见解、开发先进的人工智能模型，并改善临床实践，同时保护患者隐私并确保遵守数据保护法规。

IV. 合成数据在医疗保健中的优势与挑战

在医疗保健领域使用合成数据具有多项优势，但也面临一些必须考虑并解决的挑战，才能确保其得到有效且负责任的应用。

A. 优势

1) **隐私保护：**合成数据通过生成与真实世界数据高度相似、但不会泄露敏感信息的匿名化数据实例，有助于保护患者隐私。这种方法可确保符合 HIPAA [21] 和 GDPR [22]等数据保护法规。

2) **成本和时间节省：**由于需要进行数据匿名化、知情同意管理，并遵守法律和伦理要求，获取和共享真实世界的患者数据往往需要投入大量资源和时间。合成数据可以减轻这些负担，让研究人员和专业人士能够更高效地获取和共享数据。

3) **数据效用与质量：**合成数据可以保留真实世界数据中存在的底层关系和模式，确保其在研究、人工智能模型训练及其他应用中的效用与质量。此外，合成数据还可用于数据增强，以提升人工智能模型的性能和泛化能力。

B. 挑战

1) **保持数据保真度和多样性：**生成能够准确反映真实世界数据，同时保留多样性的合成数据是一项复杂的任务。过拟合或生成不切实际的数据可能会限制合成数据在研究和人工智能模型训练中的实用性 [23]。

2) **潜在偏倚：**根据真实世界数据生成的合成数据可能会无意间再现或放大现有偏倚，从而对人工智能模型的公平性和有效性以及研究成果产生负面影响。为确保结果公平且准确，识别并缓解真实数据和合成数据中的潜在偏倚至关重要 [24]。

3) **模型复杂度和计算资源：**生成式人工智能模型（例如生成对抗网络和变分自编码器）可能计算成本高、结构复杂，需要大量资源进行训练和优化。研究人员和专业人士必须仔细权衡模型复杂度、数据质量与计算资源之间的取舍。

通过应对这些挑战并充分利用合成数据的优势，医疗保健研究人员和专业人士可以有效运用生成式人工智能模型，获取逼真的匿名化患者数据，推动研究、诊断和治疗的发展，同时保护患者隐私并遵守数据保护法规。

V. 未来研究方向与影响

合成数据生成在医疗保健领域有几个前景广阔的未来研究方向，也可能产生重要影响：

1) **生成式人工智能模型的进展：**随着生成式人工智能模型不断发展，合成数据的质量、多样性和保真度有望提升。这将推动医疗保健研究、人工智能模型训练和临床实践中的应用更加准确、灵活。

2)**与隐私保护技术相结合：**将合成数据生成与差分隐私、联邦学习等其他隐私保护技术相结合，可以进一步提升医疗保健应用中数据的隐私保护水平和实用性，同时最大限度地降低隐私泄露风险。

3) **拓展医疗保健领域的应用：**合成数据生成可应用于广泛的医疗保健领域，包括个性化医疗、远程医疗和公共卫生监测。通过提供真实可信的匿名化患者数据，合成数据有助于加快这些领域的研究并改善患者结局。

合成数据在医疗保健领域的成功开发与应用，有望彻底改变研究人员和专业人士获取及使用患者数据的方式。

数据，最终推动诊断、治疗和整体患者护理取得重大进展，同时保护隐私并确保遵守数据保护法规。

VI. 结论

本文探讨了生成式人工智能模型（如生成对抗网络和变分自编码器）在医疗保健领域中的作用，即生成逼真且匿名化的合成患者数据，用于研究和训练。通过应对与隐私和法规合规相关的挑战，合成数据能够推动人工智能模型开发、医学培训、医疗保健研究和决策支持系统的发展。随着生成式人工智能模型不断演进，未来研究方向包括提升合成数据的质量和多样性、整合隐私保护技术，以及拓展其在各类医疗保健领域中的应用。

合成数据生成对医疗保健领域的潜在影响巨大，有望在维护患者隐私并遵守数据保护法规的同时，彻底改变研究、诊断和治疗。合成数据的成功应用最终能够改善患者结局、提升医疗保健系统的效率，并增进我们对影响人类健康的复杂因素的理解。

参考文献

[1] G. J. Annas，《HIPAA法规：医疗记录隐私新时代？》新英格兰医学杂志，第348卷，第1486页，2003年。[2] B. K. Beaulieu-Jones、W. Yuan、S. G. Finlayson、Z. Wu、P. Avillach和I. S. Kohane，《保护隐私的生成式深度神经网络支持临床数据共享》*Circulation: Cardiovascular Quality and Outcomes*，第13卷，第7期，第e006246页，2020年。[3] I. Goodfellow、J. Pouget-Abadie、M. Mirza、B. Xu、D. Warde-Farley、S. Ozair、A. Courville和Y. Bengio，《生成对抗网络》*arXiv*预印本 *arXiv:1406.2661*，2014年。[4] D. P. Kingma和M. Welling，《自动编码变分贝叶斯》*arXiv*预印本 *arXiv:1312.6114*，2013年。[5] E. Choi、S. Biswal、B. Malin、J. Duke、W. F. Stewart和J. Sun，《使用生成对抗网络生成多标签离散患者记录》*arXiv*预印本 *arXiv:1703.06490*，2017年。[6] R. K. Sevakula、W.-T. M. Au-Yeung、J. P. Singh、E. K. Heist、E. M. Isselbacher和A. A. Armoundas，《旨在改善心血管系统相关患者结局的最先进机器学习技术》美国心脏协会杂志，第9卷，第4期，第e013924页，2020年。[7] S. Jadon和A. Jadon，《少样本学习领域深度学习架构概述》，2023年。[8] Y. Dan、Y. Zhao、X. Li、S. Li、M. Hu和J. Hu，《基于生成对抗网络的高效化学成分空间采样：无机材料逆向设计》*npj*计算材料学，第6卷，第1期，第84页，2020年。[9] M. Abadi、A. Chu、I. Goodfellow、H. B. McMahan、I. Mironov、K. Talwar和L. Zhang，《采用差分隐私的深度学习》，载于*2016*年*ACM SIGSAC*计算机与通信安全会议论文集，2016年，第308–318页。[10] N. Rieke、J. Hancox、W. Li、F. Milletari、H. R. Roth、S. Albarqouni、S. Bakas、M. N. Galtier、B. A. Landman、K. Maier-Hein 等，《联邦学习与数字医疗的未来》*NPJ*数字医学，第3卷，第1期，第119页，2020年。[11] N. Mohammed、B. C. Fung、P. C. Hung和C.-K. Lee，《面向高维医疗保健数据的集中式和分布式匿名化》*ACM*知识发现与数据挖掘汇刊（*TKDD*），第4卷，第4期，第1–33页，2010年。[12] Y. Skandarani、P.-M. Jodoin和A. Lalande，《用于医学图像合成的生成对抗网络：一项实证研究》，2021年。[13] M. R. Whiteway、D. Biderman、Y. Friedman、M. Dipoppa、E. K. Buchanan、A. Wu、J. Zhou、N. Bonacchi、N. J. Miska、J.-P. Noel 等，《使用半监督变分自编码器对动物行为视频中的变异性进行分区》*PLoS*计算生物学，第17卷，第9期，第e1009439页，2021年。[14] J. Dahmen和D. Cook，《Synsys：面向医疗保健应用的合成数据生成系统》*Sensors*，第19卷，第5期，第1181页，2019年。[15] M. Frid-Adar、I. Diamant、E. Klang、M. Amitai、J. Goldberger和H. Greenspan，《基于生成对抗网络的合成医学图像增强可提升卷积神经网络在肝脏病灶分类中的性能》*Neurocomputing*，第321卷，第321–331页，2018年。[16] J. Chen、D. Chun、M. Patel、E. Chiang和J. James，《合成临床数据的有效性：使用临床质量指标对领先的合成数据生成器（Synthea）进行验证研究》*BMC*医学信息学与决策制定，第19卷，第1期，第1–9页，2019年。[17] O. Mendelevitch和M. D. Lesh，《合成医疗数据的保真度与隐私》*arXiv*预印本 *arXiv:2101.08658*，2021年。[18] K. Cresswell、A. Majeed、D. W. Bates和A. Sheikh，《面向医疗专业人员的计算机化决策支持系统：诠释性综述。》*Informatics in Primary Care*，第20卷，第2期，2012年。[19] S. I. R. Bhagyashree、K. Nagaraj、M. Prince、C. H. Fall和M. Krishna，《流行病学研究中使用机器学习方法诊断痴呆症：一项来自印度南部的探索性试点研究》*Social Psychiatry and Psychiatric Epidemiology*，第53卷，第77–86页，2018年。[20] E. H. Weissler、T. Naumann、T. Andersson、R. Ranganath、O. Elemento、Y. Luo、D. F. Freitag、J. Benoit、M. C. Hughes、F. Khan 等，《机器学习在临床研究中的作用：变革证据生成的未来》*Trials*，第22卷，第1期，第1–15页，2021年。[21] I. G. Cohen和M. M. Mello，《HIPAA与21世纪健康信息保护》*JAMA*，第320卷，第3期，第231–232页，2018年。[22] X. Larrucea、M. Moffie、S. Asaf和I. Santamaria，《探索符合GDPR要求的欧洲跨境医疗保健行业4.0安全方案》*Computer Standards & Interfaces*，第69卷，第103408页，2020年。[23] M. F. Naeem、S. J. Oh、Y. Uh、Y. Choi和J. Yoo，《生成模型可靠的保真度与多样性指标》，载于国际机器学习大会。PMLR，2020年，第7176–7185页。[24] M. A. Gianfrancesco、S. Tamang、J. Yazdany和G. Schmajuk，《使用电子健康记录数据的机器学习算法可能存在的偏倚》*JAMA*内科学，第178卷，第11期，第1544–1547页，2018年。[25] P. Voigt和A. Von dem Bussche，《欧盟通用数据保护条例（GDPR）》实用指南，第*1*版，瑞士查姆：*Springer International Publishing*，第10卷，第3152676期，第10–5555页，2017年。[26] D. Rankin、M. Black、R. Bond、J. Wallace、M. Mulvenna、G. Epelde 等，《在医疗保健中使用合成数据进行监督式机器学习的可靠性：用于保护数据共享隐私的模型》*JMIR Medical Informatics*，第8卷，第7期，第e18910页，2020年。