卫生信息化国际发展动态

北京市卫生健康大数据与政策研究中心 2023.12.29

（十一）合成数据

1. **标题：我们能相信医学合成数据吗?隐私和效用的度量综述**

来源： medRxiv.

时间： 2023 年 11 月.

链接： https://doi.org/10.1101/2023.11.28.23299124.

概要： 共享和重用超出其初始收集范围的健康相关数据对于加速研究、开发可转化为临床应用的强大且值得信赖的机器学习算法至关重要。越来越多的人认为，共享人工生成的类似于真实患者数据的合成数据是一种很有前途的方法，可以在解决与个人医疗数据相关的隐私问题的同时实现此类重用，但是目前还没有一致的系统和定量评估合成数据的实际隐私收益和剩余效用的标准方法。这阻碍了其实际采用。本研究将从隐私和效用两方面系统介绍合成健康数据评估领域的当前认识，并对当前状况提出建议和批判性分析，为进一步研究提出具体方向和步骤。研究是通过文献研究和创建包含用于评估综合数据的所有方法和度量的公共本体来评估和语境化该领域现有知识，再遵循 PRISMA-ScR 方法进行数据收集和知识合成。本研究共纳入 92 项研究，通过分析和分类，共发现 48 种不同的方法来评估合成数据的统计效用和 9 种不同的方法来评估隐私风险。研究结果表明，目前合成数据隐私还存在一种令人担忧的趋势，即在没有适当评估的情况下相信合成数据的安全性。尽管在医疗保健中使用合成数据可以提供一种简单而无忧的替代真实数据的方法，但是在评估方面还缺乏共识，这阻碍了这种新技术的广泛采用。研究者认为，通过提高认识、提供能考虑到文献现状评估的综合方法，可以促进评估标准的统一和采用，从而促进在医学领域使用合成数据。

2. 标题：通过扩散模型生成数字病理学中的合成数据：一种多方面的评估方法

来源： medRxiv.

时间： 2023 年 11 月.

链接： https://doi.org/10.1101/2023.11.21.23298808

概要： 合成数据现已成为计算病理学家工具箱中的新贵，可以支持多项任务，例如解决数据稀缺问题或增强深度学习的训练集。然而使用这些新资源是需要仔细规划的构建和评估，以避免产生临床上毫无意义的伪影等陷阱。本研究对合成数据当前主要应用成果进行了综合描述，并引入一种新的全栈方法，生成和评估由扩散模型驱动的合成病理学数据。该工作流程最大特点是一种新的多面评估策略，具有嵌入式可解释性程序，有效地解决了在健康相关领域使用合成数据的两大关键问题。采用相似集成策略来评估生成的数据，通过一组完善的指标评估真实数据和合成数据的相似性比较，评估生成图像在深度学习模型中的实际可用性，并辅以可解释的人工智能方法，并通过由三位专业病理学家回答的专用问卷来验证其组织病理学真实性。该方法在包含 650 个 WSI 的公共 GTEx 数据集上进行了演示，包括五个不同的组织，并调整了底层扩散模型的训练步骤，然后从这五个组织中的每一个生成相同数量的切片，最后，使用建议的评估方法来评估生成数据的可靠性，并取得了显著效果。研究证明，这些评估步骤中的每一个都是必要的，因为它们提供了有关生成数据质量的补充信息。上述所有功能都将所提出的工作流程描述为数字病理学中生成式人工智能的成熟解决方案，成为数字病理学界向数字化和数据驱动建模过渡的潜在有用工具。

（徐健编辑）

译文一：

我们能相信医学合成数据吗?隐私和效用的度量综述

Bayrem Kaabachi, Jérémie Despraz, Thierry Meurers, Karen Otte, Mehmed Halilovic, Fabian Prasser, Jean Louis Raisaro, 徐健（译）

介绍

获取高质量数据对于有影响力的医学研究和实践至关重要，尤其是随着人工智能(AI)和机器学习(ML)的兴起，因为它推动了精准医疗等领域的进步和创新，在这些领域，建立安全、快速和可靠的程序来访问数据以供二次使用已变得至关重要。

然而，由于隐私问题，对医疗数据的访问通常受到高度限制并受数据保护法（如美国健康保险流通与责任法案（HIPAA）中规定的保障措施的约束以及欧盟《通用数据保护条例》（GDPR）在这些监管框架下，用于共享高度敏感数据的常用方法是低于接受阈值的数据匿名化。这种方法是采用数据屏蔽和转换技术来降低重新识别风险的。

尽管如此，即使在可以达到足够保护级别的情况下，匿名化高维数据也往往会受到严重冲击，而且匿名数据集的效用可以使它几乎无法用于研究。

对于这一数据共享问题，合成数据是一个很有效的解决方案，Chen 等人将其描述为一种“肯定会很快用于解决医疗保健领域紧迫问题”的技术。它背后的主要思想是生成模拟真实患者数据统计特性的人工数据。这种数据合成过程可以使用多种算法来实现，但最近几年的主要突破是使用生成对抗网络(GANs)。gan 通过使用两种神经网络来实现:一种是创建假样本，另一种是评估它们与真实数据的接近程度。用这些网络协作来改进生成的样本，直到它们与真实数据非常相似，使其成为生成真实人工信息的有价值的工具。由于所有样本都是人工生成的，因此合成样本与真实样本匹配的概率通常很小。

因此，合成数据已经获得了相当大的覆盖面，甚至超出了专业来源，这种更广泛的认可产生了大胆预测，声称“到 2024 年用于人工智能和分析项目的数据的 60%将是合成的”。

特别是在医学领域，有几项研究使用合成数据来复制最初对健康相关数据进行的案例研究。这些结果凸显了合成数据在医疗环境中的潜在好处，并为使用合成数

据作为严格监管的个人数据的替代方案提供了强有力的论据。

然而，虽然这些结果似乎对医疗环境中保护隐私的数据共享的未来充满希望，但最近的研究指出，过度依赖合成数据作为“银弹”解决方案存在一些风险。例如，作为合成数据一部分的单个记录可能会对生成的合成数据产生强烈影响，从而允许恶意攻击者推断原始数据集中存在的单个记录。这尤其涉及到机器学习模型在训练数据上的过拟合趋势，以及记忆数据集中个体的泄漏。一般来说，最接近原始数据集的合成数据集可能也最有用，但同时提供的隐私保护较少。另一方面，与原始数据非常不同的合成数据集将提供强大的保护，但可能不太实用。由于 gan 的黑箱特性，很难预测在训练和生成过程中丢失了哪些数据效用，以及生成的数据中可能包含哪些敏感信息。因此，Stadler 等人认为，在生成和共享合成数据时必须采取谨慎的方法。

最近的研究中强调了与合成数据使用相关的潜在风险提出了一个问题，即合成数据领域的研究重点是否更强调实用性而不是隐私考虑。与匿名数据相比，我们可以在其中找到广泛文献来描述可以应用的各种攻击和隐私保护机制，合成数据尚未得到全面审查。这促使我们进行这次审查，希望我们能为这个问题提供一个可靠、无偏见的答案。

该领域的一些调查研究了合成数据生成的各个方面。Figueira 等人提供了对多代方法的广泛描述，而 Hernandez 等人探索了评估方法并对其进行了比较，以确定表现最佳的方法。与这些先前的研究相比，我们的方法不同之处在于我们如何识别阻碍采用合成数据的障碍，因为我们更加强调评估过程和隐私效用权衡困境，通过系统地研究如何在 92 项研究中评估合成数据。

同时，开源解决方案，如 Synthetic Data Vault、Table Evaluator 和 TAPAS 已开发并公开发布，以帮助研究人员创建和衡量合成数据的质量。这些平台提供了一系列评估指标和方法来评估效用和隐私，从而简化了评估过程。然而，这些开源工具也带来了自己的挑战，因为它们各自使用自己的命名法和术语，这增加了在医疗保健领域内实现合成数据的统一视角的复杂性。这一点，再加上在文献中存在相互矛盾的观点，对医疗保健中合成数据的统一理解的发展变得复杂。

因此，为了更好地了解与医疗保健相关的合成数据生成的现状，我们启动了这项审查，专门针对评估方法，旨在对合成数据评估方法的适用性进行严格的定量分

析。为此，我们遵循以下研究问题：

1. 社区内是否就如何评估合成数据的隐私和效用达成共识？

2. 在评估合成数据时，隐私和实用性是否同样重要？

方法

对于本次范围综述，我们采用了系统综述和荟萃分析的首选报告项目（PRISMA)。PRISMA 是公认的指南，通常用于制定系统评价和荟萃分析。其框架旨在使流程更加清晰和一致。具体而言，PRISMA 强调明确定义研究问题的重要性，设置明确的纳入和排除参数，并详细说明从所选文献中搜索、选择和收集数据的方法。

为了确定本次范围综述的合成数据评估方法相关研究，我们在 2018 年 1 月到 2022 月 12 月期间对多个书目数据库和存储库进行了全面检索。图 1 描述了计算机科学文献的主要数据库和存储库 IEE Explore 和 ACM Digital Library 和专注医疗保健和医学研究的 PubMed 和 Embase。符合表 1 所述的纳入标准的文章进入全文审查。

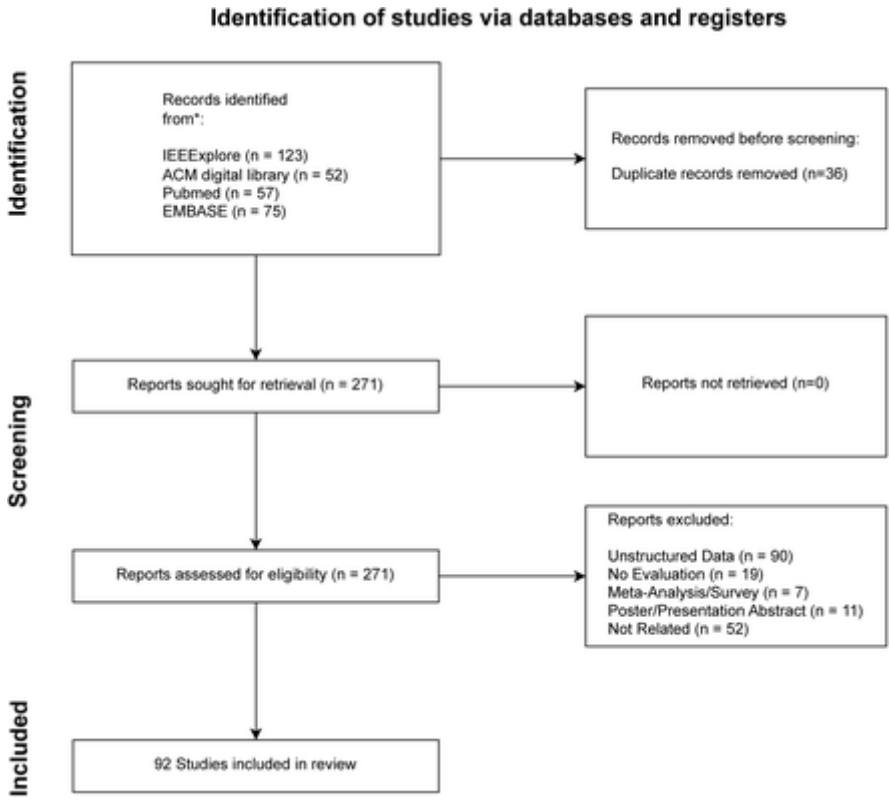

图 1. 用于审查过程的 PRISMA 流程图

表 1.文献纳入和排除标准

Inclusion criteria	Exclusion criteria
Publications describing research that uses synthetic data generation methods and evaluates their outputs.	Surveys and systematic/scoping reviews.
Papers published between 01.01.2018 and 31.12.2022.	Documents in languages other than English.
Publications describing work that focuses on structured data i.e. no images/text problems.	No assessment of the generated output i.e. no look at the utility/privacy aspect of the generated data.
–	Do not contain structured data.
–	Poster abstracts.

包括 IEE Explore 和 ACM Digital Library 等计算机科学数据库的原因是为了捕获更先进和创新的合成数据生成技术和评估方法，因为这些数据库通常包含可能推动该领域向前发展的新技术文章。

相反，纳入 PubMed 和 Embase 等医疗保健特定数据库则是为了确定可能提供更扎实、更实用和临床相关评估方法的研究。这些数据库更有可能包括考虑了医疗保健环境的独特限制和要求的研究，从而确保所审查的合成数据方法适用于现实世界的医疗环境。

每个数据库的检索策略都是在研究的早期阶段制定的，然后通过团队讨论和对结果的初步分析进行改进。我们专门设计了查询条件，以专注于评估合成数据的效用或隐私方面的出版物。这样做是为了找到与合成数据方法的质量和安全性相关的文章，提出可操作的建议，而非只描述新技术。**表 4** 附录 **A** 中列出了用于每个数据库的查询，最新检索时间是 2023 年 8 月月 14 日。

查询设计中的另一个考虑因素是避免误报，例如讨论合成化合物或材料而不是合成数据的出版物。为此，我们将“标题”和“摘要”作为查询字段，确保已确定出版物的主要关注点确实是合成数据及其效用或隐私的评估指标。如果这些文章仍出现在最终的论文选择中，我们将手动删除它们。

研究选择中的任何差异都通过两位作者之间的讨论和共识得到解决。研究团队合作设计了一个数据图表表格，如**表 5 附录 A** 所示，以描述从所选出版物中提取的特定变量。选择后，从每项研究中提取数据图表表格中描述的信息。

数据制图过程中的挑战是属性的标准化，因为它确保了从选定出版物中提取信息的一致性，并能够对所考虑的研究进行定量评估。

为了满足标准化的要求，我们创建了评估方法的分类法。下一节将介绍这种分类法，重点介绍隐私和效用的各个方面，以及它们在现有文献中的表示。

分类 – 合成数据实用程序

图 2 所示的效用方法分为三个统计类：“单变量相似性”、“双变量相似性”和“多变量相似性”。结构化的评估方法简化了对多个维度的合成数据和真实数据之间相似性的理解，因为它可以直接比较各种生成方法。我们还纳入了与纵向数据相关的方法类别，因为它们具有分析时间模式和趋势的独特性质。另一个包括的类别是“特定领域相似性”，它评估合成数据在特定研究领域的表现，例如复制研究结果或使用特定于该领域的指标。

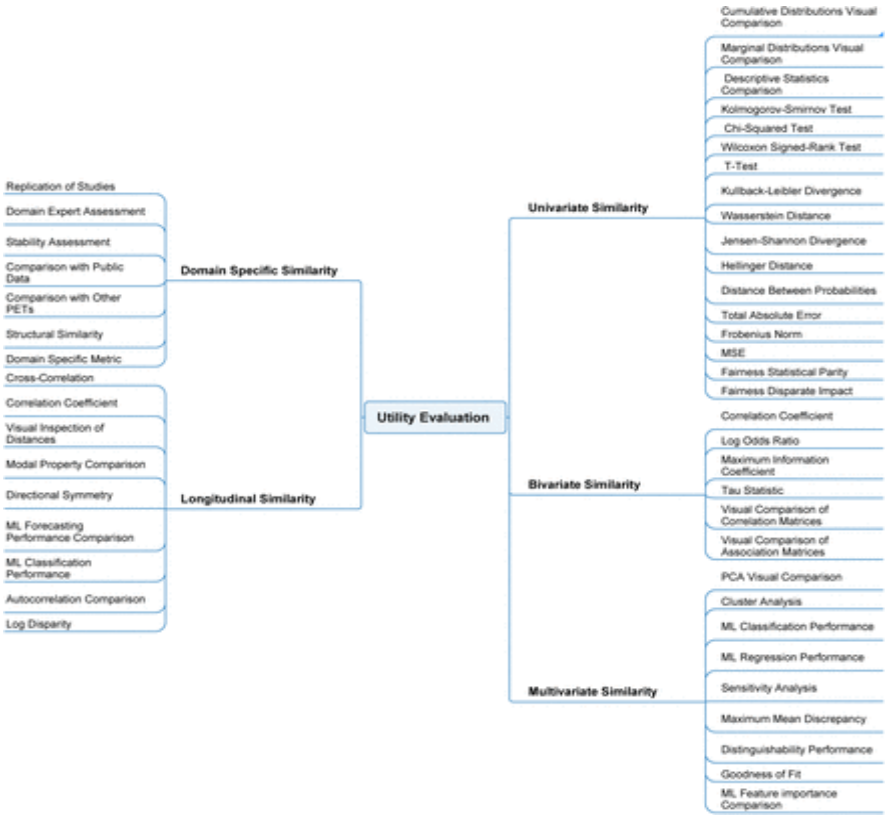

图 2. 合成数据效用评估的分类

图 3. 合成数据隐私评估的分类

图 4. 合成数据生成方法

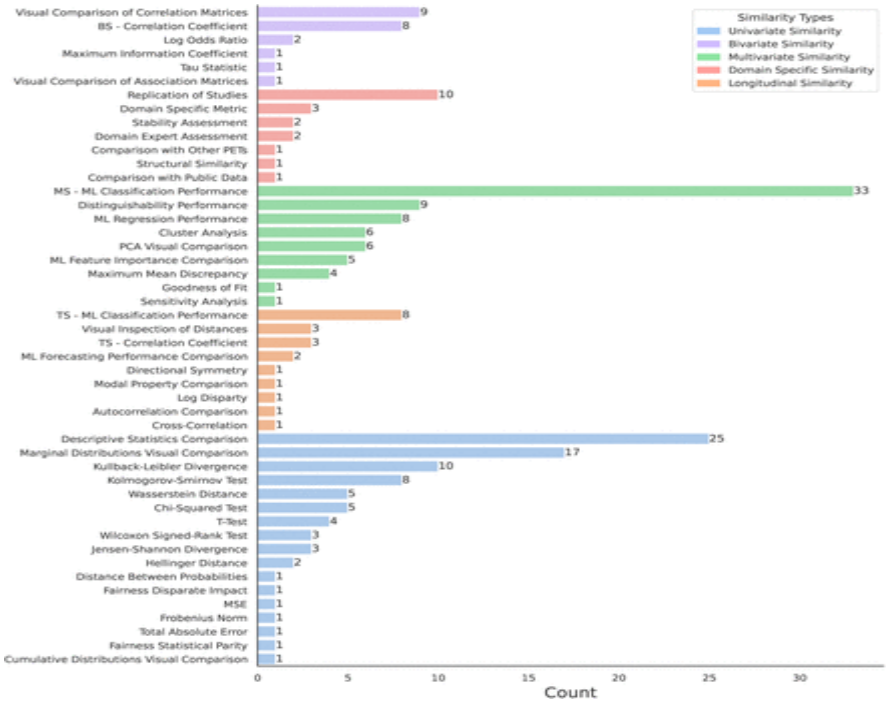

图 5. 合成数据实用程序

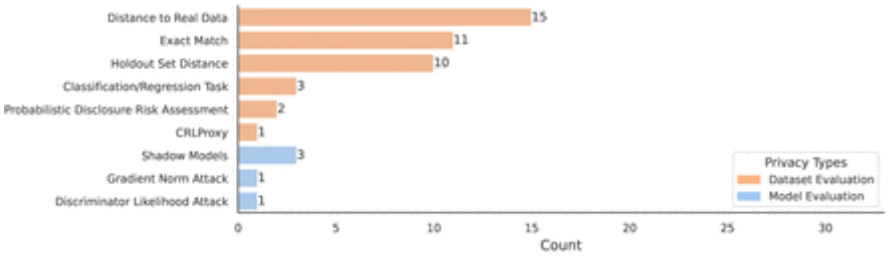

图 6. 合成数据隐私

这种方法的局限性涉及公用事业类别的潜在重叠。例如，对合成数据和真实数据执行机器学习分类任务可以同时属于“领域特定相似性”作为“研究的复制”和“多变量相似性”作为“分类性能”。由于很难辨别我们研究的出版物作者的初衷，因此当出现任何歧义以避免冲突时，我们选择将方法归类为“研究复制”。

附录 B 包含本分类法中每个项目的全面描述。

分类法 – 合成数据隐私

图 2 所示的隐私方法分为两类：“数据集评估方法”和“模型评估方法”。

数据集评估方法评估合成数据本身提供的隐私保护。主要目标是确定合成数据集在多大程度上保护敏感信息并保护隐私，尤其是与原始真实数据相比。当主要关注的是数据本身的隐私时，例如当医院想要向公众发布数据集时，或者当重点是发布单个数据集时，这种评估非常重要。通常，这些方法利用距离指标进行比较。它们要么将生成的合成数据集直接与生成过程中涉及的原始数据集进行对比，要么与从同一群体中提取的维持数据集进行对比。

模型评估方法将重点转移到评估用于生成合成数据的隐私保护算法或方法上，这使得了解所选机制在各种情况下保护隐私的程度成为可能，并且通常涉及计算合成数据生成机制带来的隐私风险的估计上限。一种常见的用法是评估生成机制的多个输出，以建立可以描述为“最坏情况”的情况。当重点放在隐私保护机制本身的性能和稳健性上时，这种评估至关重要。它有助于公平地比较彼此之间的技术，因为两者都是根据隐私风险的上限进行评估的。这方面的一个例子是使用阴影模型 15. 这些模型涉及创建多个副本，这些副本反映了主要合成数据模型的行为。尽管这种方法可能占用大量资源，但它建立了一个强大的评估框架，模拟黑盒攻击场景，确保进行全面的隐私风险评估。

附录 C 包含本分类法中每个项目的全面描述。

结果

一般结果

在这篇综述中，我们发现，在调整了语义相同但命名不同的方法后，有 48 种方法用于评估效用，9 种方法用于评估隐私。图 7 概述了我们选择的所有出版物中使用的效用和隐私评估方法的所有情况。审查的完整结果见表 2 和表 3。

	Utility					Privacy	
	Univariate Similarity	Bivariate Similarity	Multivariate Similarity	Longitudinal Similarity	Domain Specific Similarity	Model Evaluation	Dataset Evaluation
34	X	X	X				
35	X	X		X			
36				X			
37	X	X	X				
38							X
39					X		
40	X	X	X				
41	X						X
42			X		X		
12	X	X	X		X		
43					X		
44				X			
45	X		X			X	X
46				X			X
47	X			X			X
48	X			X			
49	X		X				
50			X				
51		X	X				X
52		X	X				X
53	X	X	X				X
54	X		X				
55							X
56			X		X		X
57	X		X				
58			X				
59			X				
60	X	X					
10	X	X	X				X
61		X					
62	X	X					X
63	X		X				X
64	X				X		
65	X		X	X			
66			X				
13					X		
67	X				X		X
68	X						
69	X		X		X		X
32	X		X			X	
70	X	X	X				X
71			X	X			
72				X			
73	X						
74	X	X	X				
75	X						X

表 2.审查结果

	Utility					Privacy	
	Univariate Similarity	Bivariate Similarity	Multivariate Similarity	Longitudinal Similarity	Domain Specific Similarity	Model Evaluation	Dataset Evaluation
76	X		X				
77			X		X		
78				X	X		
79	X		X				
80							X
81							X
31			X			X	
82					X		
83				X			
84			X				
85	X	X	X				X
86	X	X	X				X
87				X			
88	X		X				
89			X				
90	X		X				
91	X						
92			X				X
30						X	X
93					X		
94			X				X
95		X	X				
96	X						
97					X		
11	X		X				
98	X		X				
99			X				
100	X		X				X
101	X						
102			X				
103		X	X				
104			X				X
105		X					
106	X						
107	X		X				
108		X		X			
109	X		X	X			
110	X		X				
111					X		
112			X				
113			X				X
114	X				X		
115		X	X				X
116	X		X				
117							X
118	X		X				

表 3.审查结果

我们回顾了 2018 年至 2022 年发表的文章，这段时间涵盖了生成式 AI 技术的激

增和崛起，包括对 GAN 的早期热情和大型语言模型（LLM）的出现。根据图 9 附录 D，我们发现只有 4.35%（4/92）来自 2018 年，10.87%（9/92）来自 2019 年。到 2022 年，这一比例已跃升至 43.48%（40/92），这表明随着时间的推移，人们对合成数据的兴趣日益浓厚。

此外，我们发现大多数文章使用横断面数据，占总数的 70%（64/92）。只有 26% 24/92）使用时间纵向数据，可能是因为它通常更难合成。对于这种类型的表格数据，困难在于不仅要维护反映在变量之间相关性的列之间的关系，还要维护表示数据时间一致性的行之间的关系。如表 1 所示，本综述未考虑非结构化数据。

使用不同的方法来创建合成数据。大约 35%（17/49）的文章使用了 GAN。其余的 65%（32/92）使用了其他方法的组合，包括统计建模和 Synthpop 等专业软件 28R 包或 MDClone29 平台。

合成数据实用程序

在我们的资格标准中，我们特别关注评估其合成数据生成方法输出的作品。其中，94%（86/92）评估了合成数据的效用。

在 48 种效用评估方法中，我们确定了 17 种单变量相似性、9 种纵向相似性、另外 9 种专门针对多变量相似性、8 种特定于领域相关相似性和 6 种特定于双变量相似性。

有三种方法最常用。多变量相似性：ML 分类性能是主要方法，在 33 个实例中应用。单变量相似性：描述性统计比较在 25 例中使用，单变量相似性：边际分布视觉比较在 17 例中使用。

合成数据隐私

图 8 显示，合成数据的隐私方面是大多数入选论文背后的主要动机，因为其中 80%（74/92）打算将合成数据用于私人数据共享场景。另外 16%（15/92）将其用于数据增强目的，并回答数据稀缺问题或阶级不平衡。其余 4%（3/92）研究了合成数据在两种情况下的潜力。

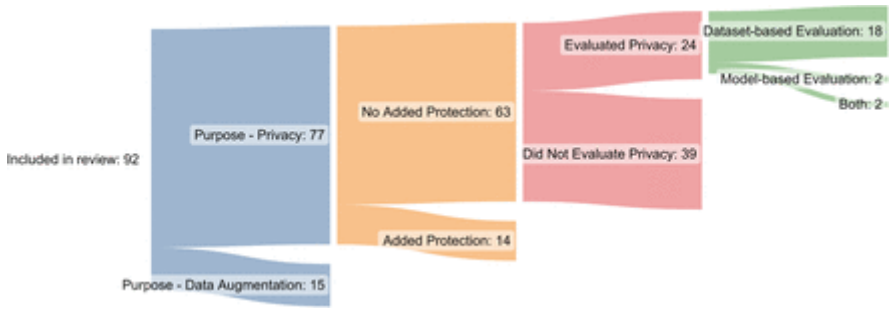

图 8. 说明如何以及是否评估合成数据的隐私

在这 77 项旨在使用合成数据保护隐私的研究中，有 15 项应用了差分隐私技术或为数据引入了额外的掩蔽层。这种额外的掩码层通过添加二级数据更改来发挥作用，这进一步隐藏了原始记录，确保实际数据受到保护，并且不太容易追溯到单个来源在剩下的 63 项研究中，只有 38%（24/63）包括至少一种隐私评估方法。这意味着，虽然被评估作品对隐私的需求是显而易见的，但等式的评估部分并没有遵循，合成数据的隐私往往被盲目信任。

在包含隐私评估的论文中，84%（20/24）主要依赖于基于数据集的评估。较小的数字，8%（2/24），专注于基于模型或机制本身，例如那些利用 GAN 架构的模型或机制 **30** 或涉及阴影建模过程的那些 **31-33** 另外 8& （2/24） 进行了两次评估。

讨论

RQ1：社区内部是否就如何评估合成数据的隐私和效用达成共识？

我们在上一节中显示的结果表明，研究人员目前尚未就评估合成数据隐私和效用的标准化指标达成共识。在各研究中使用各种各样的指标使得比较和综合现有证据具有挑战性。

随着该领域研究的不断发展，选择最佳“解决方案”来生成高保真度和高隐私的合成数据变得越来越困难，因为无法直接和公平地比较可用的解决方案。关于如何知道新兴的新合成数据生成方法是否真正适合采用的整体混乱使得这些最先进的技术在现实世界环境中使用具有挑战性，这凸显了对一套标准化评估指标的需求，以促进有意义的这在隐私评估方面更加明显，因为这些方法也受到法律约束。据我们所知，关于如何评估合成数据隐私风险，没有明确的法律文本。尽管最近有人尝试

将合成数据指标映射到现有的 GDPR 定义，例如 Giomi 等人的“挑出”、“链接”和“推理”。目前尚无关于合规性的确认。

综上所述，我们得出结论，合成数据评估领域仍处于萌芽阶段。我们预计，随着技术的成熟和法律框架的适应，评估合成数据的方法应该会融合成一种更加标准化和值得信赖的方法。这一点很重要，尤其是在医疗保健等关键领域，利益相关者的信任是必不可少的。

RQ2：在评估合成数据时，隐私和实用性是否同等重要？

我们在图 7 和图 8 中的发现清楚地表明，隐私评估通常不如公用事业评估那么彻底。

虽然合成数据的效用一直是一个主要关注点，但隐私评估通常非常有限且不完整，因为与评估机制的方法相比，评估数据集的方法在文献中应用的次数之间存在明显差异。

在本次综述中，使用合成数据时对隐私的低估尤为明显。超过一半的研究声称使用合成数据来保护隐私，并且“应该”评估隐私没有进行任何正式的隐私评估。取而代之的是，他们“按原样”使用合成数据，假设固有的隐私优势，而没有经验验证。

这种疏忽引起了重大关注，尤其是在生成合成数据的软件解决方案领域。用户可能会无意中认为他们生成的合成数据默认是保护隐私的。这可能会导致合成数据在不知情的情况下共享，除了道德和法律上的复杂性外，还可能导致个人数据泄露。

对隐私评估的宽松方法，加上对合成数据隐私保护能力的假设，暴露了当前研究和实践中的重大差距。它强调了在评估合成数据生成方法的效用和隐私方面需要一种更平衡的方法。

影响评价方法选择的因素

在评估合成数据的效用和隐私时，有多种方法是显而易见的。文献揭示了 48 种不同的效用评估方法，而隐私评估方法相比之下较少。这些评估技术的选择可归因于多个相互交织的因素：

研究目标在方法选择中起着举足轻重的作用。对于效用评估，当合成数据用于数据增强时，指标主要倾向于机器学习任务和传统基准测试。相反，当合成数据充当真实数据的“代理”时，则指标更侧重于特定属性，而不是更广泛的评估。在隐私考虑中，选择通常落在成员推断和基于研究目标的属性推断之间。例如，选择成员推理是因为其对数据泄漏的直接评估，并作为检查复杂推理攻击可行性的前兆。

实施的复杂性是另一个关键的决定因素。更简单的方法，例如用于效用的单变量相似性比较或相关矩阵分析，以及用于隐私的基于距离的指标，由于它们易于使用标准软件包而受到青睐。相反，更复杂的方法（如日志集群指标或影子模型）需要额外的考虑因素，例如无监督学习或多个模型的训练。

可解释性也是方法选择的核心。允许进行视觉比较的评估技术通常更具吸引力，尤其是在向利益相关者展示时。虽然一些方法，如隐私评估中的精确匹配攻击，提供了明确的可解释性，但其他方法由于其复杂性而需要更详细的解释。

最后，数据的结构和类型以及模型的泛化性会影响选择过程。例如，时间序列数据需要与横截面数据不同的效用指标。此外，一些攻击方法是为特定的合成数据生成技术（例如 GAN）定制设计的，其中鉴别器可用于量化风险因素 30，限制了它们在各种数据生成技术中的通用性。

局限性

这种范围界定审查虽然全面，但并非没有局限性，因为我们的分析中可能没有捕捉到一些相关的研究或方法。在制图过程和分类法的发展过程中，必须做出某些决定，这些决定可能会引入主观性或限制我们评估的粒度。在解释多篇论文中的不同指标并试图将它们整合到一个统一的术语下时，这一点尤其明显。

例如，一个限制与“特定领域相似性”指标的分类有关，因为很明显，该类别下的方法通常具有与其他指标不同的范围或含义。这个总括性术语可能包含各种方法，这些方法在粒度和具体目标方面有很大差异。决定将这些不同的指标归入“特定领域的相似性”是为了简化分类法，但我们承认它可能不是每种情况的最精确拟合。

此外，我们发现术语“保真度”和“效用”有时可以互换使用，但一些研究 120 认

为它们应该被视为不同的指标。保真度主要与合成数据和原始数据之间的统计相似性有关，而效用则侧重于合成数据集对特定任务的功能有用性。这种区别虽然在本次综述中没有直接涉及，但仍然反映在“特定领域相似性”一词下的分类法的构建中。

结论

这篇综述详细介绍了合成健康数据的效用和隐私性研究现状，揭示了其潜力和缺陷。对标准化评估措施的迫切需求是一个主要观点，我们认为拥有统一的指标可以提供一个公平的竞争环境，允许以一致和有意义的方式比较不同的合成数据生成方法。

在这项工作中提出的一个重要问题是需要进行强有力的隐私评估。由于医疗保健行业包含敏感信息，因此确保合成数据不会无意中导致数据泄露或失去信任至关重要。当涉及到 GAN 时尤其如此，因为它们固有的复杂性和缺乏透明度可能会阻止许多人采用它们，从而成为障碍，或者由于缺乏意识而导致误导的使用。当 IEEE 的 Industry Connections 活动等国际倡议时，医疗保健领域对标准化和安全合成数据的迫切需求越来越大和地平线欧洲对合成数据的呼吁证实了为该领域的安全和开发可靠框架制定明确指南的紧迫性。因此，我们进行本综述的目的不仅是阐明这些挑战，而且还要激发合作努力，制定最佳实践，使这些技术更容易获得和理解。

在将合成数据集成到医疗保健环境中的旅程中应谨慎行事。其能力的诱惑力应该用平衡的观点来缓和，避免过度宣传。任何评估或实施都应有条不紊地进行，确保结果既有效又公正。

译文二：

通过扩散模型生成数字病理学中的合成数据：一种多方面的评估方法

Matteo Pozzi, Shahryar Noei, Erich Robbi, Luca Cima, Monica Moroni, Enrico Munari, Evelin Torresani, Giuseppe Jurman, 徐健（译）

介绍

近年来，扫描技术的快速发展彻底改变了传统病理学，催生出数字病理学。从传统载玻片到高分辨率全载玻片图像（WSI）的过渡为利用人工智能（尤其是深度学习）彻底改变病理学提供了令人兴奋的机会，为计算病理学铺平了道路。深度学习算法在自动执行关键任务方面表现出令人印象深刻的能力，例如癌症检测或分级以及细胞分割。然而，这些深度学习算法的成功很大程度上取决于大规模注释数据集的可用性和质量，这在数据稀缺和隐私问题突出的医学领域可能具有挑战性。为了解决这个问题，合成数据生成已经获得了动力，通过生成保持与真实数据相似的统计特性的数据，使其成为在处理敏感数据（如医学图像）时解决隐私问题的有希望的途径。

合成数据生成的核心是生成模型，例如生成对抗网络（GAN）。GANs 在生成各种医学领域的合成图像方面得到了广泛的应用，包括数字组织病理学图像。然而，最近的进展表明，扩散模型在生成自然和医学图像方面都超过了 GAN。扩散模型是一种受非平衡热力学启发的生成模型。这些模型采用两步过程来学习数据分布，然后生成新数据。在第一步中，定义了扩散步骤的马尔可夫链，其中噪声逐渐添加到输入中。然后，该模型通过从噪声输入开始生成新数据来学习逆转这一过程。扩散模型在医学图像生成中的应用包括（但不限于）脑磁共振图像、显微血细胞图像、正电子发射断层扫描心脏图像和胸部 X 射线。最近发表了一篇关于该主题的综合综述。

扩散模型在生成组织病理学图像中的首次应用可以追溯到，作者成功采用了去噪扩散模型生成了三种不同神经胶质瘤亚型的苏木精和伊红（H&E）图块。他们的结果表明，基于图像生成质量评估中使用的常见定量评估指标，扩散模型在这项任务中优于 GAN。在他们成功之后，作者在文章中还证明了扩散模型在生成包含六种

不同细胞核类型及其相应掩膜的结肠组织病理学图像方面的能力和优于 GANs。此外，另一篇文章的作者使用基于视觉转换器的扩散模型来生成结肠组织病理学图像，其质量高于使用 GAN 生成的图像。

在利用生成的数据之前，对其进行评估至关重要。然而，对生成数据的评估仍然是一个有争议的话题，各种研究采用了不同的方法。缺乏标准化的评估管道来全面评估生成的数据是一个重大警告。在这项研究中，我们旨在通过引入一个通用框架来解决这个问题，该框架用于生成，更重要的是，从各种角度彻底评估合成病理图像。我们将说明这些多方面的评估如何产生互补的建议，而这些建议无法单独通过任何单一方法获得。

如图 1 所示，拟议的合成病理学数据生成和评估管道从预处理阶段开始，从该阶段提取 WSI 中的切片。

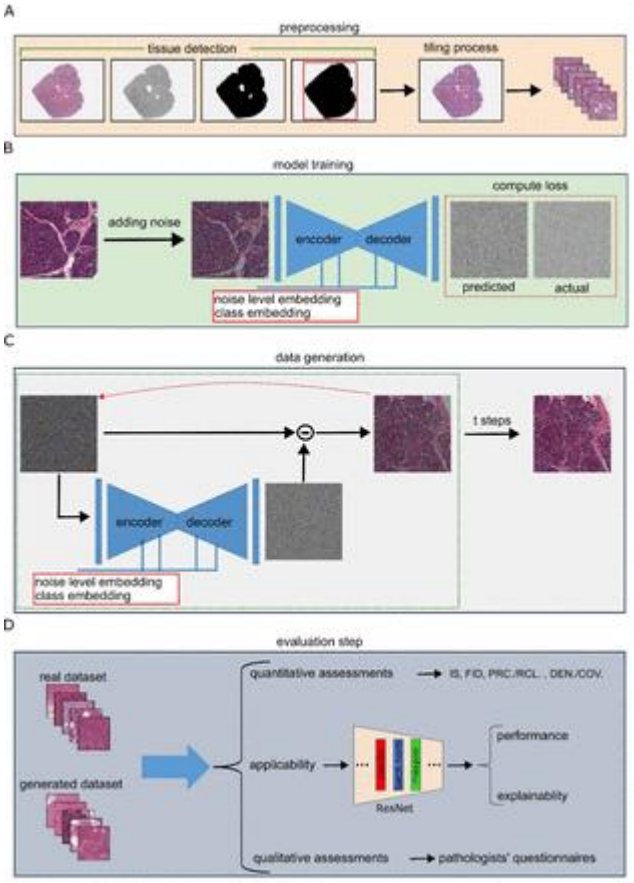

图 1 建议渠道。

演示了我们提议的管道的图形化工作流程。（A）演示预处理步骤，首先，通过一组步骤在载玻片中检测组织，然后进行平铺步骤以提取成批的切片。（B）演

示模型训练步骤，其中提取的切片被输入到去噪扩散模型中，其中噪声最初被添加到图像中。这些受扰动的图像，以及噪声水平嵌入和类嵌入（在我们的例子中是组织类型），随后被输入到 U-Net 模型中，该模型预测添加到图像中的噪声。初始噪声和预测噪声都用于计算损失并驱动训练过程。（C）描述数据生成步骤。在这里，纯噪声以及所需的类嵌入被输入到训练的模型中以预测噪声图像，然后从原始噪声中减去噪声图像以创建噪声较小的图像。生成的图像用作新的噪声输入。在重复此过程“t”步骤后，将生成最终图像。（D）随后，对生成的数据集进行综合评估，分为三个不同的类别。定量评估需要诸如初始分数 （IS）、Frechet 初始距离 （FID）、改进的精确召回率 （PRC、RCL） 和密度覆盖率（DEN、COV）等指标来衡量生成的数据集与真实数据集的相似性。

实际评估包括训练用于组织检测的 ResNet 分类器，评估和比较生成数据与真实数据集在性能和可解释性方面的可用性。最后，通过对专家病理学家进行的一系列问卷调查来评估生成数据的生物学真实性。

随后，训练扩散模型以生成基于不同类别的数据。在训练过程之后，将执行采样阶段，为每个类生成特定数量的合成数据。最后，评估步骤过程包括三组不同的评估。第一组涉及定量指标，用于衡量生成的图像和真实图像之间的相似性。对于我们的定量评估，我们采用了合成图像领域常用的成熟指标，包括初始得分（IS）、Fréchet 初始距离 （FID）、改进的精确召回率和密度覆盖率。

第二组评估生成图像在深度学习模型中的实际可用性。为了补充这组数据，我们研究了分类器学习的卷积滤波器，这些过滤器使用概念相关性传播（CRP）算法在真实数据和生成的数据上进行训练，这是一种可解释的人工智能方法。

最后，进行定性评估以验证生成数据的生物学真实性，这是以前评估中没有针对的关键方面。这种全面的方法确保了为我们选择的任务生成的数据集的可靠性和有用性。我们将该管道应用于源自基因型组织表达 （GTEx）的 650 种不同组织的 39 个 WSI。我们特别选择包括多个组织，因为大多数以前的研究只关注一种组织类型：这种保守的方法未能证明对于具有不同形态复杂性的其他组织类型是否可以预期类似的成功。

据我们所知，这项研究代表了一种新颖且可重复的组织病理学图像生成方法，它覆盖了多个组织并从各个角度对生成的数据进行了全面评估。通过考虑不同的组

织和评估指标，我们旨在更全面地了解所提出的生成模型的功能和局限性。这种多方面的评估使我们能够深入了解生成的数据集对潜在各种任务的有效性和适用性。

材料与方法

数据

图像来自 GTEx 研究）。GTEx 数据集包含使用 Aperio 载玻片扫描仪使用 H&E 染色收集的 WSI。

在 53 种可用的人体组织中，我们保留了仅与从我们指定的感兴趣组织（即大脑、肾脏、肺、胰腺和子宫）获得的样本相关的数据。此过滤过程产生了大约 5000 个样本的元数据。随后，我们从过滤后的元数据中随机选择了一组 650 个样本，以获得它们对应的 WSI。在无法下载 WSI 的情况下，该样本被排除在最终数据集之外。

在下载的 648 个 WSI 中，共有 589 个被用作深度学习模型的训练集，而其余的则专门用于测试目的。

预处理流水线

WSI 展示了相当高的空间分辨率:在我们的案例中，在 20 倍放大系数下，空间基础分辨率为每像素 0.494µm。这种粒度使得它们无法作为神经网络的直接输入。

通常建议将 WSI 划分为大小一致的块，然后将其用作深度学习模型的输入。因此，预处理阶段被认为是必要的，以充分准备收集的数据进行模型训练。

在我们的研究中，我们使用了 histab1 库来促进玻片的平铺过程。这个库提供了工具来生成组织蒙版和执行平铺过程。

WSI 不一定只包含组织信息。因此，在给定载玻片的情况下创建组织蒙版是关键的一步，因为它确保生成的图块主要包含与组织相关的信息，同时最大限度地减少无关的背景区域。我们观察到以下有序步骤（如图 **1A** 所示）定义了足够的组织掩膜：

1. **RgbTo** 灰度滤镜：此过程涉及将 3 通道图像转换为单通道灰度图像。

2. **OtsuThreshold** 过滤器：给定灰度图像，返回前景和背景分离的二进制阈值图像。

3. **二元膨胀：**对二进制图像执行膨胀，即通过为其相邻像素中的最大值分配输出图像的像素值来确定输出图像的像素值。

4. **BinaryFillHoles** 过滤器：此过程涉及填充二进制图像中所有区域的任务。

除了掩码生成之外，还需要其他参数才能平铺 WSI：

- **切片尺寸：**指定提取的切片的最终分辨率（以像素为单位）。
- **级别：**指定幻灯片中的切片提取级别，该级别越低（最小值 0），放大倍数就越高。
- **像素重叠：**以像素为单位指定图块之间的重叠程度。

我们决定以 512 倍的放大倍数从幻灯片中提取分辨率为 512×5 像素的非重叠图块（因此分辨率为每像素 1.976μm）。这些参数是在与病理学家协商后选择的，旨在优化组织分类任务，为组织区分提供足够的上下文信息，并避免可能导致切片无法识别的过度缩放。

这导致 63247 块切片被分为两组:从 589 个训练 WSI 中，提取了 56990 块切片用于训练生成模型和分类器，而从 59 个测试 WSI 中提取了剩余的 6257 块切片用于测试目的。然而，即使使用精确定义的蒙版，也不是所有生成的贴片都能一致地捕获原始幻灯片的信息属性。如补充图 1(左面板)所示，一些切片缺乏可识别的组织形态学特征，这表明它们可能不会为模型提供有价值的信息。为了过滤这些贴图，我们利用了文章所启发的复杂性度量(C)，它被定义为:

其中水平(x)和垂直(y)方向的逐像素梯度(G)计算每个贴图 t。然后，梯度的总和乘以依赖于贴图的宽度 N 和高度 M 的项。具有非均匀组织的贴图 t 将表现出较高的像素梯度值，从而导致相应的高 Ct 值。

我们基于视觉检查应用了一个选择的阈值(即 0.4)，以仅保留显示出足够可区分特征的切片。这一步移除了 948 个无信息的贴图，让我们保留了总共 62299 个贴图，范围从 10891 个 Brain 到 14173 个胰腺。补充表 2 总结了其他细节。

去噪扩散概率模型

对于我们的生成模型，我们使用了研究文章提出的去噪扩散概率模型（DDPM）。DDPM 是扩散模型的一个特定变体，它专注于通过学习能够生成高质量输出的概率模型来对输入数据进行去噪。为了创建一个涵盖五种不同组织类型的数据集，我们遵循了研究文章中概述的无分类器指导方法。简而言之，管道的工作方式如下。

首先，在输入图像中引入高斯噪声。该噪声的大小由调度程序随机确定，从而导致对输入图像施加不同级别的噪声。这种方法允许模型探索图像噪声的频谱，范围从细微失真到各向同性高斯分布。噪声图像以及表示噪声水平的嵌入向量和表示类别（在我们的示例中为组织类型）的可训练嵌入向量被馈送到基于 U-Net 的神经网络，该神经网络旨在学习输出噪声。然后将该预测与实际添加的噪声进行比较，形成计算损失和指导模型训练过程的基础。该过程如图 **1B** 所示。

训练阶段完成后，图像生成阶段开始。从纯各向同性噪声的基线开始，经过训练的网络预测相应的噪声水平。该预测噪声与初始噪声之间的差异导致创建噪声较小的图像，然后再次成为网络的输入。对 t 个步骤重复此迭代过程，从而生成最终图像，如图 1**C** 所示。

模型实现可以在网上找到 2。图像被调整为 256 像素，因为不同尺寸的实验表明，这产生了最有利的结果。对于模型训练，我们使用了 26 个批的大小。学习率设定为 1e-5，训练过程包含 215000 步。我们通过监测每 1000 步后的采样数据来经验地发现学习率。数据生成的步数(t)设置为 250。每 1000 步对图像生成和模型检查点进行视觉检查。一直训练到训练损失出现收敛，并且在训练过程中进行采样提供了令人信服的结果，因为观察到即使在训练损失稳定的情况下，生成的图像的质量也可以得到提高。类调节是使用可训练的 torch.nn.Embeddings 变量完成的。最后，使用锐化滤波器对生成的图像进行进一步增强，锐化系数为 3。

评估

为了从各个角度全面评估生成图像的质量，我们设计了多组实验，从广泛认可的评估生成图像质量的指标到专家人工评估。在下文中，我们将概述每个实验的细节。

规范的定量评估指标

我们使用合成图像领域公认的指标评估了生成图像的质量，包括初始得分（IS）、Fréchet 初始距离 （FID）、改进的精确召回率和密度覆盖率。在这里，我们将简要解释每个指标。

IS 使用以下公式计算：　　其中 y 是一组类（在我们的例子中是组织），x 是图像，p（x|y） 是条件概率，p（y） 是边际类概率，D 吉隆坡代表 Kullback Leibler 背离和 Cx~PG 系列表示所有生成图像的平均值。

较高的 IS 值表明生成的图像在不同类别中具有高度的多样性，同时在类别内保持良好的相似性。我们利用了火炬计量 **3** 在我们的研究中，实现计算 IS，将图像拆分 5 次，并使用 Inception_v3 网络无偏对数作为特征来计算条件概率和边际概率。

Fréchet Inception Distance （FID） 是一种量化生成图像分布与真实图像分布之间相似性的度量。它通过将生成的真实图像传递到原始 Inception_v3 网络，从指定层中提取输出，并计算以下公式来实现：在这里，μr 和 Σr 表示真实图像的估计平均值和协方差矩阵，而 μg 和 Σg 表示 Inception\textunderscore{v3} 特征空间中生成的图像相同。跟踪操作用 Tr 表示。为了计算 FID，我们利用了 torchmetric 实现，将特征参数设置为 2048，这对应于最后一个池化层。FID 值越低，表明真实图像和生成图像的分布更相似。

FID 和 IS 都依赖于 imagenet 预训练的 Inception_v3 网络，对于那些分布与自然图像明显不同的图像来说，这可能不是理想的。为了解释这一点，我们还采用了通过检查真实数据和生成数据之间流形的相似性来评估生成图像质量的方法。文章作者提出了通过测量生成数据和真实数据的流形重叠来提高精度和召回率的方法。简而言之，它们是通过将数据投影到特征空间，并通过将每个数据点周围环绕一个到达其第 k 近邻的超球来估计真实数据和生成数据的流形来计算的。精确度定义为生成的样本在真实数据流形内的百分比，召回率定义为真实图像在生成的数据流形内的百分比。两者的值越高，表示实际数据和生成数据之间的重叠越大。在异常值存在的情况下，准确性和召回率受到影响。为了克服这个问题，文章作者提出了密度和覆盖度量，它们提供了改进的流形估计。

在本研究中，我们使用研究文章提供的代码计算精度、召回率、密度和覆盖率。我们将邻居的数量设置为 10。为了防止在特定数据集上训练网络可能产生的

任何潜在偏差，通过随机权重的 ResNet-50 网络将数据投影到 100 维空间中来创建真实数据和生成数据的特征空间(为了结果的可比性，我们还使用 ImageNet 预训练网络进行了分析)。我们修改了网络的最后一个完全连接层，以输出一个 100 维向量。由于真实图像和生成图像数量之间的不平衡可能会使估计产生偏差，我们通过从真实数据中随机抽取每个组织生成的图像数量相等来解决这个问题。为了获得可靠的估计，我们将整个过程重复了五次，并报告了平均值和平均值的标准误差。

基于深度学习的实践评估

虽然上述评估措施为生成图像的质量及其与真实图像的相似性提供了有价值的建议，但它们可能无法提供有关更复杂的形态学特征（例如组织病理学合理性）或生成图像的适用性以用作机器学习研究中真实图像的替代或补充的足够信息。这些方面至关重要，也是我们当前研究的主要重点之一。因此，需要额外的评估标准。为此，我们设计了一组使用深度卷积神经网络（CNN）的实验。

我们使用具有随机初始化权重的 ResNet-50 架构在两种不同的场景中训练组织分类器。在第一个场景中，我们使用用于训练扩散模型的真实切片从头开始训练网络。根据研究文章中显示的结果，我们留出 30%的载玻片专门用于内部验证。这种方法减少了模型评估中与数据泄漏相关的偏差的可能性。为了确保网络学习与组织类型相关的相关形态特征，我们在训练期间应用了一组以下增强：空间转换，包括随机旋转和随机水平垂直翻转，色度转换，如抖动和随机通道交换，以减轻染色的影响，以及鲁棒性增强增强，例如使用高斯滤波器进行模糊处理和随机擦除部分图像。在保留用于测试的真实图块（检查**“材料和方法”，“数据集”**部分）和生成的图块上评估了性能。

该网络训练了 100 个 epoch，批次大小为 32。我们采用了 AdamW 优化器，初始学习率为 0.0002。使用余弦退火调度程序（tmax 参数为 100）降低了学习率。与验证最高精度相对应的模型被保存以供后续步骤使用。模型和训练是使用 pytorch 实现的和 pytorchlightning.

在第二个场景中，采用与第一个场景完全相同的训练设置，我们使用生成的切片从头开始训练网络。唯一的区别是，由于生成的图像位于平铺级别而不是幻灯片级别，因此我们保留了 30%的平铺用于内部验证。随后，在两个不同的测试集上对

训练好的网络进行了评估。第一个测试集包括随机选择的 15%的载玻片子集，用于训练扩散模型（称为内部测试集）。相比之下，第二个测试集由扩散模型看不到的真实幻灯片组成（用于评估在真实数据上训练的网络性能的相同测试集）。

我们使用多类精度（ACC）和马修斯相关系数 （MCC）评估了模型在测试集上的性能。

可解释性分析

我们采用概念相关传播 （CRP）作为评估 ResNet-50 是否利用类似的概念（由 CNN 中的卷积滤波器表示）对属于同一类的图像进行分类，而不管它是在哪个数据集上训练的。这种分析对于确保分类器在识别切片组织时采用一致的策略至关重要，无论其来源如何（真实的或生成的）。从该分析中获得的建议为评估分类器本身和扩散模型生成的合成切片的质量提供了有价值的信息。

CRP 结合了局部和全局方法，为模型的预测提供可靠而全面的解释。它基于逐层相关性传播（LRP），但扩展到根据网络学习的概念解开解释。CRP 计算概念条件热图以定位输入空间中概念的激活，并提供数据集级别的统计数据来研究不同类或图像如何激活模型中的特定通道。

我们从生成的数据集上训练的网络中提取了每个类的重要概念，方法是随机选择每个类一个生成的图像，然后选择卷积层，在 ResNet-50 架构中，卷积层是输出前的最后一个卷积层。这种选择是合理的，因为我们的目标是识别更高层次的概念，这些概念通常被编码在更靠近网络输出的更深层中。对在真实数据上训练的网络也做了同样的过程。

之后，我们通过在类和层上调节相关性传播来限制相关性传播。这使我们能够确定所选图层中每个通道的重要性，以便准确地将给定的图块分类为属于其相应类别。

在我们的分析中，我们提取了每个模型每个类别的前 5 个通道列表，按相关性排序。这些通道被进一步用于特征可视化，我们从合成和真实数据集中获得了每个概念最具代表性的样本，以及归因热图。此外，我们评估了与这些概念相关的流行类别，评估了概念提取程序的完整性和有效性。该分析还有助于评估研究结果的普

遍性，通过检查在不同数据集上使用相同模型时相同的关联是否成立。此分析是使用 CRP python 包完成的。

定性评估

众所周知，人类专家评估仍然是最终基准，尤其是在医学图像方面。出于这个原因，我们请了三位专业病理学家（LC、EM、ET）来帮助我们评估生成图像的质量。

我们通过设计两种不同的问卷来做到这一点。第一份问卷的前半部分包括 100 个问题。病理学家从单个组织中获得了三个切片，并负责识别相应的组织。一半的问题以从真实数据中提取的图块为特色，而剩下的一半则由生成的图像组成。病理学家不知道切片的“类型”（真实与生成的），问题是随机的，并且包括相同数量的组织。目的是评估病理学家使用有限数量的切片检测组织的能力，并检查真实数据和生成数据之间的潜在性能差异。第一份问卷的后半部分包括另外 100 个问题。这一次，病理学家需要根据单个图块确定给定的图块是真实的还是生成的。每种组织和数据类型（真实或生成）的问题数量相等。与上一节类似，这些问题在组织中随机分配，并区分真实数据和生成数据。第一份问卷中的问题示例见**补充图 2**。

后续问卷是在收到对第一份问卷的答复一周后发送的。病理学家被要求从 0（不重要）到 5（非常重要）对他们在区分真实和生成的切片时使用某些特定图像特征的程度进行评分。这一阶段的目的是弄清楚病理学家的方法是否主要依赖于“视觉属性”，例如模糊、色度特征或伪影的存在，或者他们是否在区分真实图像和生成的图像时也发现了生物学或形态学上的不规则性。在七个问题中，四个对应于视觉属性，三个对应于生物属性。完整的问题清单见**附表 3**。

实现

所有数据处理和模型训练代码均在 Python（版本 3.7.16 或更高版本）中实现。使用的任何特定软件包都在相应的材料和方法部分中指定。对于评估问卷，我们使用了 JavaScript 和 Google Apps 脚本 API。

预处理、分类器训练和评估是在一台配备了 2 个 Nvidia GTX-1080Ti GPU、Intel i7-6700 CPU 和 32GB RAM 的机器上进行的。扩散模型训练是在具有 2 个 NVIDIA V-

100 GPU 和 32GB VRAM 的集群上进行的。

所有代码都可以在 Git-Hub 存储库 {https://github.com/Mat-Po/diffusion_digital_pathology} 中找到。

结果

我们使用提取的 56990 个真实切片集作为扩散模型的输入，并执行了超过 215000 个步骤的训练。确认收敛损失和对图像质量的初步视觉评估，我们为每个组织制作了 3001 张合成图像。这一选择是出于我们可用的资源和时间限制。图 2 描绘了每个组织的真实图像和生成的图像对样本。

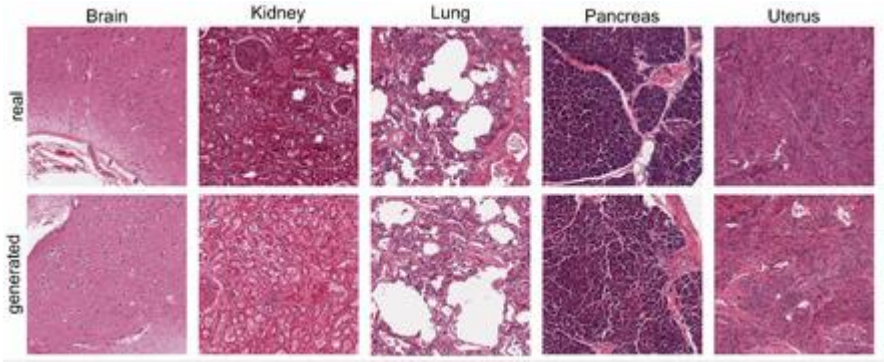

图 2. 每个组织的真实图像对和生成的图像对示例。

顶行显示每个组织的真实数据集中的代表性图像，而底行显示来自同一组织的生成图像。

经过初步的目视检查，确认生成的图像与实际数据非常相似，我们继续进行评估。

测量生成图像和真实图像之间的相似性会发现生成的数据与真实图像相似

生成合成数据集后，我们使用各种合成图像评估指标对其质量进行了定量评估。这些结果总结在表 1 中。

IS (↑)	FID (↓)	Precision (↑)	Recall (↑)	Density (↑)	Coverage (↑)
2.9115 ± 0.0339	35.1140	0.9620 ± 0.0033	0.9385 ± 0.0091	0.7877 ± 0.0186	0.8516 ± 0.0101

表 1. **量化评价指标结果。**

该表显示了定量指标的值，包括初始分数（IS）、精确率、召回率、密度、覆盖率和 Fréchet 初始距离（FID）。IS、精度、召回率、密度和覆盖率的值报告为

基于重复计算的均值±均值的标准误。FID 是使用整个数据集计算的。箭头指示更高（↑）或更低的（↓）值更理想。有关这些指标的计算和解释的更多详细信息，请参阅“材料和方法”部分。

总结果表明，生成的图像与真实图像非常相似。我们生成的图像获得了很高的 Inception 分数（2.9115 ± 0.0339），超过了以前生成合成数字病理图像的尝试。事实上，使用我们生成的图像获得的初始分数与我们的真实数据集（3.2892±0.0149）所获得的分数相当。FID 值虽然略高，但仍与其他研究报告的值相当。需要注意的是，这些研究利用了较小的视野，这可能导致较低的 FID。例如，在研究文章中，将视野增加两倍（相当于我们的缩放级别）使 FID 增加了一倍多（达到 38.1 的值），这超过了我们报告的 FID。此外，根据研究报告的结果，更改数据集会导致实现的 FID 发生重大变化。精确率和召回率值接近最大值 1，优于以前的工作。然而，重要的是要考虑到这些研究使用 ImageNet 预训练网络计算精度和召回率，这可能会影响结果。为了确保更公平的比较，我们使用 ImageNet 预训练的 ResNet 重新计算了这些值。结果表明，我们的精确度优于其他研究，而我们的回忆仍然与他们的结果相当。这些比较的摘要见补充表 4。据我们所知，之前没有研究报告密度和覆盖率指标，但我们的值也接近最大可实现值，表明生成的图像与真实图像之间的高度相似性。然而，需要注意的是，由于数据集、图块大小、缩放级别和组织的差异，研究之间的直接比较可能并不完全公平和可行。我们在本文中的主要关注点不仅仅是提高生成图像的质量。尽管如此，我们提出了这些定量指标，以证明我们生成的图像表现出与以前发表的其他使用扩散模型的研究相似的质量。

计算出的指标提供对生成数据集的整体质量的评估。但是，为了进行更详细的分析，我们还单独计算了每个类别的 FID。结果如表 2 所示。虽然在单独计算每个类别时预计 FID 总体增加，但我们观察到不同组织中这种增加的变化。具体来说，与子宫和肺组织相比，大脑和胰腺组织的 FID 增加幅度更大，其 FID 值保持在接近总体 FID 评分的水平。这些结果凸显了在类别水平上检查生成模型性能的重要性，因为它为不同组织类型中生成数据质量的差异提供了有价值的建议。

Brain	Kidney	Lung	Pancreas	Uterus
58.92	45.89	39.19	64.98	33.97

表 2.类 FID。

为了更好地了解每种组织生成数据的质量，我们分别计算了每种组织类型的 Fréchet 起始距离（FID），如材料和方法中所述。报告每个组织类别的 FID 分数，表明生成的图像与真实图像之间的差异。

在基于深度学习的任务中，生成的数据表现出与真实数据相当的性能

表 5 总结了经过训练的 ResNet-3 在各种场景下用于组织分类的性能。该表概述了从不同实验设置中获得的测试结果。此外，为了更深入地了解分类结果，相应的混淆矩阵在补充图 3 中可视化。

Train set – Test set	ACC	MCC
Real data – External real data	98%	0.9705
Real data – Generated data	98%	0.9796
Generated data – Internal real data	95%	0.9331
Generated data – External real data	93%	0.9133

表 3. **用于组织分类任务的训练神经网络的性能。**

报告各种训练和测试方案的准确度（ACC）和马修斯相关系数（MCC）值，如材料和方法部分所述。

结果表明，在看不见的真实数据或生成的数据上进行测试时，在真实数据上训练的网络的性能没有显着差异。这一发现表明，生成的图像已经成功地捕获了原始组织中存在的各种形态模式，因为网络对它们进行分类的能力保持一致。相反，在评估在生成数据上训练的网络时，观察到看得见和看不见的真实数据的性能下降。值得注意的是，早期的研究还记录了当仅依赖生成数据而不是真实数据时性能下降。这种性能下降可归因于几个因素。首先，与真实数据相比，在生成数据上训练的网络可用的训练样本数量较少，这可能会导致准确性降低。事实上，当我们使用与生成图像切片相同数量的真实图像切片重新训练网络进行实验时，我们观察到精度下降了约 1.5%（数据未显示）。此外，从混淆矩阵中可以明显看出，某些组织（如胰腺和肺）的性能下降明显。这一观察结果表明，这些组织呈现出更加错综复杂的特征，可能给扩散模型生成多样化和准确的切片带来挑战。有趣的是，这种现象与我们的定量指标形成鲜明对比，例如按类别划分的 FID，后者可能预测了大脑等组织的更大困难。然而，深度学习评估的结果与这一预期不一致，这凸显了多方面评估的重要性。

尽管存在这些缺点，但总体结果是有希望的，因为它们表明生成的图像表现出与真实切片中的图像相似的模式。此外，通过利用少量生成的图像，我们可以达到与使用真实数据相当的性能水平。

可解释性介导的分析进一步揭示了有关从两个数据集中学习到的特征的建议

我们使用 CRP 来识别两个 ResNet 中每个类激活的前 5 个通道或概念，一个在真实数据上训练，另一个在生成的数据上训练。这些清单列于附表 5。一旦我们确定了特定于类的概念，我们就会使用它们来可视化哪些输入像素激活了这些概念。

例如，在图 3A 中，通道 428 与基于真实数据训练的 ResNet 中与肾脏相关概念相关联，由来自生成数据和真实数据的肾脏样本激活。考虑到我们如何提取这个通道，这在某种程度上是从真实数据中预料到的。但是，对于生成的数据，这并不是理所当然的。值得注意的是，激活通道 428 的输入像素是同一解剖结构的一部分，即肾脏的肾小球。简单来说，在真实数据上训练的 ResNet 中的通道 428 在真实数据和生成数据中都被肾脏样本中的肾小球激活。

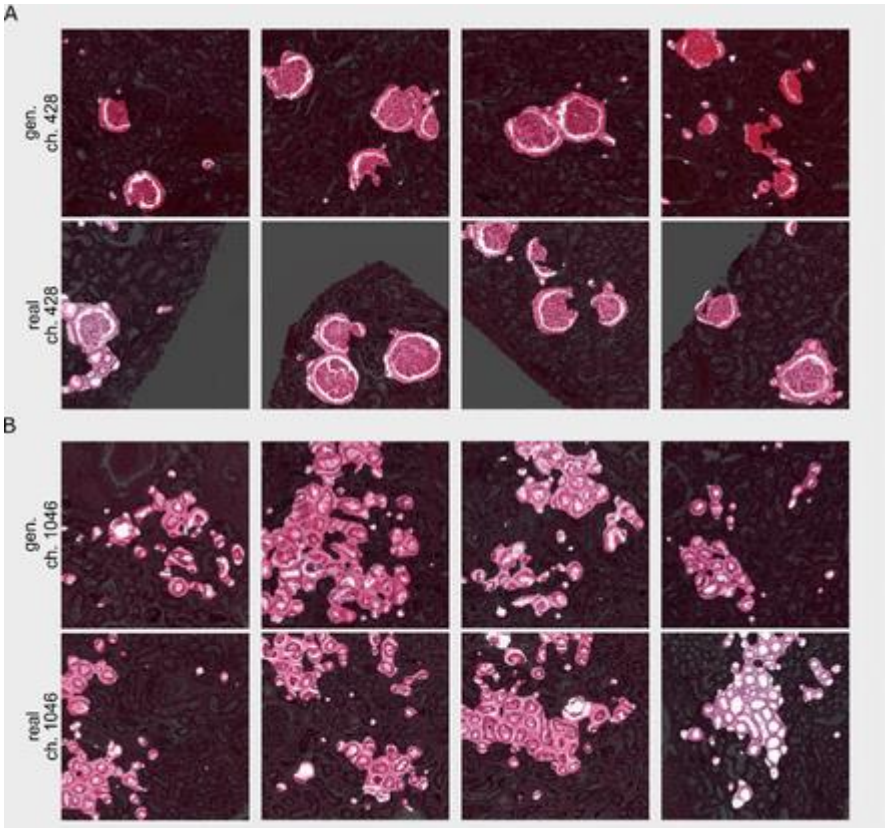

图 3. 概念相关性传播可视化。

激活指定通道的输入像素的可视化。每个小组探索一个网络，无论是根据真实数据还是合成数据进行训练，以及一个概念。每一行检查通道在合成数据集（gen）或真实数据集（real）上的相关性。（A）在真实数据集上训练的 ResNet。在最后一个卷积层中激活通道 428 的输入像素属于同一组织类别，并且在两个数据集中具有一致的形态元素。（B）在合成数据集上训练的 ResNet。在这里，激活通道的切片在两个数据集中也属于同一类，输入像素属于一致的形态结构。

至关重要的是，当我们检查在生成数据训练的网络中肾脏样本激活最多的概念时，如图 3B 所示，出现了类似的模式。在真实数据和生成的数据中，激活该通道的输入像素对应于肾脏图像的肾小球。

我们还使用了 CRP 中的全局方法来确定哪些类被同一组选定通道激活得最多。补充表 5 的上半部分包含来自在真实图像上训练的网络的通道，而下半部分包含在生成的切片上训练的网络通道。在检查两个数据集中每个通道激活的类后，我们发现有很强的一致性，这表明网络在涉及两个数据集的切片时表现相似。此证据支持用于分析分类器性能的性能指标。

此外，这些结果揭示了网络犯错误的情况，特别是在胰腺和肺之间混淆的情况下。补充表 890 下部的通道 5 说明了这一点，因为尽管该通道激活的最常见类别是胰腺，但第二常见的类别是肺类，根据数据集的不同，其实质性归一化相关性为 0.95 和 0.94。

为了更好地理解，我们为每个组织选择了最相关的概念，并使用在真实数据集上训练的网络（补充图 4）和在生成的数据集上训练的网络（补充图 5）在一些示例真实图像和生成的图像中可视化激活该概念的像素).虽然在某种程度上，大脑、肾脏、肺和子宫等组织最重要的概念在真实数据和生成数据训练的网络之间保持一致，但胰腺的情况并非如此。对于在生成数据上训练的网络，胰腺组织最重要的概念虽然具有足够的描述性（如分类器的整体良好性能所示），但似乎类似于肺和子宫组织最重要的概念，这解释了性能的下降。

在专家眼中，生成的数据在生物学上是现实的

病理学家的问卷结果汇总在表 4 中，如图 4 所示。

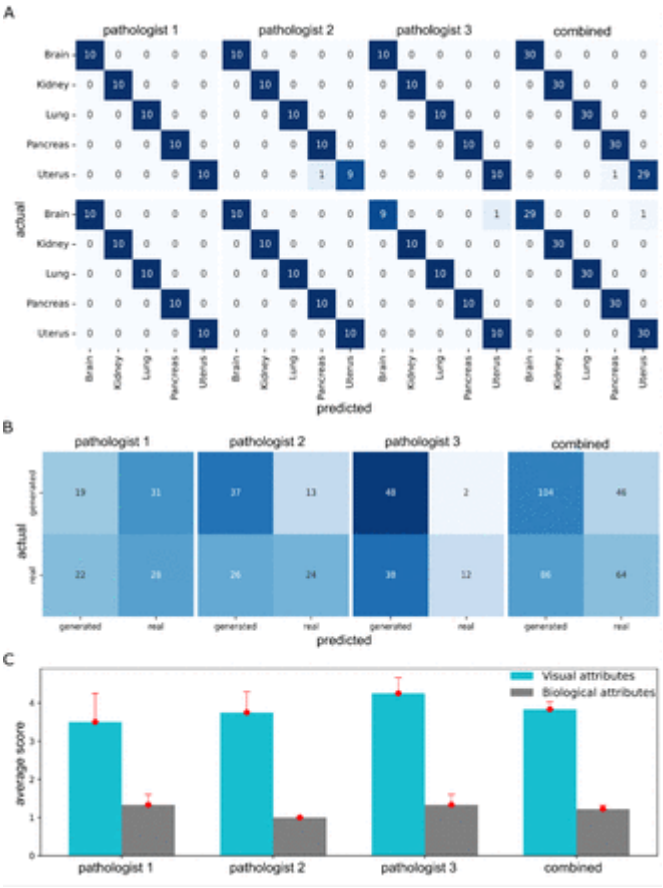

图 4. 病理学家的问卷结果。

（一）在第一份问卷中，使用真实（顶行）和生成的（底行）图像，展示了病理学家在组织检测任务中的表现的混淆矩阵。这些矩阵对应于表 3 的前两行。（B）显示描述病理学家在第一份问卷中区分真实图像和生成图像的任务中表现的混淆矩阵。这些矩阵对应于表 3 中结果的最后一行。（C）说明病理学家对第二份问卷的回答平均值的平均±标准误差。青色条表示平均分数，表示病理学家依靠视觉属性来区分真实图像和生成图像的分数，而灰色条表示他们利用材料和方法中描述的生物属性的程度。

Task	Pathologist 1		Pathologist 2		Pathologist 3		Combined	
	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
Tissue detection with real images	100%	1	98%	0.9757	100%	1	98.33%	0.9917
Tissue detection with generated images	100%	1	100%	1	98%	0.9757	98.33%	0.9917
Real or generated	47%	-0.061	61%	0.2278	60%	0.2881	56%	0.1245

表 4.第一份问卷结果。

使用真实和生成的数据报告组织检测任务的准确性（ACC）和马修斯相关系数（MCC）值，以及真实与生成的鉴别任务。每个病理学家的结果单独呈现，并通过汇总所有病理学家的反应获得组合结果。

组织检测的第一个任务的结果表明，病理学家非常熟练地在提供的缩放级别下根据给定的三个图块来识别组织类型。重要的是，真实图像和生成的图像在性能上没有显着差异，表明生成的数据准确地捕捉了原始组织的形态。病理学家在检测真实切片与生成的切片方面的综合表现略高于机会水平（准确率为 56%），这表明生成的切片具有生物学上逼真的外观。当我们检查生成的图块（生成的图块被误认为是真实的）的错误分布时，如**补充图 6** 所示，大脑图像的平均错误数略高，这表明它们在病理学家的眼中看起来更逼真，相反，肾脏图像的平均错误数最低，表明病理学家发现更容易检测生成的肾脏图像。然而，各组织的总体差异很小且不显着。值得注意的是，第二份问卷的结果表明，病理学家主要依靠色彩特征和模糊度等视觉属性来确定切片是真实的还是生成的，因为他们没有在生成的图像中检测到任何明显的生物学不规则性。当组合时，视觉属性的平均得分为 3.83±0.19，与生物属性的平均得分 1.22±0.01 有统计学差异（双尾配对 t 检验，p 值=0.0075）。这一发现凸显了生成的数据集在教育目的中的潜在价值，因为它表明生成的图像在生物学上是逼真的，即使是专家病理学家也几乎无法区分它们，除了图像视觉质量水平上的一些细微差异。

讨论

与许多其他学科一样，数字病理学领域可以从生成逼真的合成数据的能力中受益匪浅。这些数据可以通过提供高质量的数据来帮助训练过程，从而帮助解决隐私问题并实现人工智能辅助诊断。这在需要像素/切片级注释的场景中很有希望，这些注释在由人类执行时可能非常耗时且容易出错。因此，在本文中，我们介绍了一个全面的管道，用于生成和评估数字病理学中的合成数据。我们采用该管道从五个具有宽视场的不同组织生成图像，并采用去噪扩散模型。通过三种不同的评估方法评估数据质量：采用成熟指标的定量评估，通过基于深度学习的组织分类和可解释性进行实际适用性评估，以及涉及对病理学家进行问卷调查的定性评估。我们的研究结果强调了这样一个事实，即采用这些不同的评估为我们提供了对数据质量的互

补建议，并强调仅依靠单一方法是不够的。

鉴于目前在整个载玻片水平上分析数字病理学图像的趋势，也希望在同一水平上生成合成数据。在我们的研究中，我们专注于生成宽视场的切片，以检查扩散模型在这种分辨率下生成图像的能力，并作为更接近 WSI 图像生成的早期尝试。未来的研究可以在此基础上，从多个视野生成图块，然后将它们集成在一起，以生成完整的整个幻灯片图像。此外，在生成模型中加入切片的邻接关系可以帮助生成相邻切片，以便能够生成 WSI。研究文章中用 GAN 完成了这方面的一个例子。

我们探索了扩散模型从各种组织生成切片的能力，每个组织都具有不同的形态复杂性。我们的研究结果表明，尽管整体性能令人满意，但该模型的有效性在不同组织中有所不同。一个有趣的观察结果强调了在单独检查每个组织时不同评估的重要性。

特别是，在比较各类 FID 评分时，大脑和胰腺显示出相似的结果，均明显高于肺。然而，胰腺和肺的分类器性能明显较低，而大脑则表现出近乎完美的性能。此外，病理学家问卷的结果表明，病理学家最难区分真实和生成的脑图像，这意味着生成的脑图像具有最高质量。

可解释性分析进一步阐明了这些发现。他们发现，当网络在生成的数据上训练时，检测胰腺的最相关的组织病理学特征与在真实数据上训练时不同。有趣的是，这些特征与肺组织最相关的概念非常相似。这些差异是否归因于扩散模型的无分类器引导管道在区分肺和胰腺组织方面的局限性，或者它们是否源于胰腺组织本身固有的复杂性，还有待确定。

尽管组织分类不是数字病理学中最重要的优先事项之一，但我们利用其实用性来评估生成数据的质量及其在基于深度学习的任务中替换真实图块的可行性。正如预期的那样，仅根据生成数据训练的模型的性能与在真实数据上训练的网络的性能不匹配，但结果仍然有希望和令人满意。在处理需要像素级注释的场景时，生成数据以促进网络训练变得特别有价值，这些场景是劳动密集型的。最近的研究表明，生成模型在生成带注释的组织图像对以及特定组织的相应细胞注释掩模方面的潜力。这为扩展我们的工作奠定了基础，可能涉及从各种组织生成数据，并在每个组

织特有的正常和各种病理条件下附有它们各自的细胞类型注释。

可解释的人工智能是机器学习社区中一个快速增长的子领域。它是为了应对解释黑盒模型预测的挑战而出现的，其中潜在的数学函数通常非常复杂且难以理解，即使对于人工智能专家也是如此。深度神经网络虽然在处理非结构化数据（如图像）时性能非常有效，但属于黑盒模型。为了利用深度学习方法的力量，同时满足对可解释性的需求，已经开发了事后方法。在这项研究中，我们展示了可解释性技术的应用，以增强我们对数据质量的理解，并衡量对于生成的数据，模型是否学习了与使用真实数据训练时相当的特征。这具有特殊的意义，特别是考虑到最近的一项研究强调了生成的组织病理学图像如何提供超出真实数据所能达到的解释。这在医学领域尤为重要，因为对可解释模型的需求仍然至关重要。

最后，病理学家在区分真实图像和生成图像方面相对较低的表现增强了我们对生成图像真实性的信心，从而促进了它们在教育目的中的实用性。值得注意的是，从随后的调查问卷中获得的反馈很有意思。它强调，病理学家成功识别生成的切片的例子通常归因于可辨别的视觉线索，如模糊或色度特征。鉴于组织病理学图像的分辨率非常高，生成的图像无法始终如一地匹配该分辨率水平，这是意料之中的。在未来的研究中，可以通过集成超分辨率网络来解决这一局限性，以提高生成图像的质量。

总体而言，尽管数字病理学中的合成数据可能还没有准备好完全补充用于研究目的的真实图像，但我们的研究结果与其他研究一样，强调了它们在未来的巨大潜力。需要考虑的重要因素是需要从不同角度检查这些生成的数据，以充分了解它们的优势和劣势。我们的研究结果强调了使用各种方法彻底评估生成的数据集的重要性，这应该成为该领域的普遍做法，以更好地有效地开发和使用合成数据集。

***注：原文和译文版权分属作者和译者所有，若转载、引用或发表，请标明出处。**