arXiv:2601.00787v1 [cs.CL] 2 Jan 2026

跨司法管辖区调整自然语言处理模型：加拿大癌症登记处的一项试点研究

Jonathan Simkin
不列颠哥伦比亚大学人口与公共卫生学院，温哥华，BC
jonathan.simkin@bccancer.bc.ca

Lovedeep Gondara
不列颠哥伦比亚大学人口与公共卫生学院，温哥华，BC
lgondara@mail.ubc.ca

Zeeshan Rizvi
纽芬兰与拉布拉多卫生服务 圣约翰斯，NL Zeeshan.Rizvi@nlhealthservices.ca

Gregory Doyle
纽芬兰与拉布拉多卫生服务 圣约翰斯，NL Gregory.Doyle@nlhealthservices.ca

Jeff Dowden
纽芬兰与拉布拉多卫生服务 圣约翰斯，NL Jeff.Dowden@nlhealthservices.ca

Dan Bond
纽芬兰与拉布拉多卫生服务 圣约翰斯，纽芬兰与拉布拉多省 Dan.Bond@nlhealthservices.ca

Desmond Martin
纽芬兰与拉布拉多卫生服务 圣约翰斯，纽芬兰与拉布拉多省
Desmond.Martin@nlhealthservices.ca

Raymond Ng
数据科学研究所 ¡学研究所¡ tute 不列颠哥伦比亚大学
rng@cs.ubc.ca

2026年1月5日

摘要

基于人群的癌症登记处（PBCRs）依赖病理报告作为主要诊断来源，但人工抽取耗费大量资源，也会延误癌症监测。尽管基于Transformer的自然语言处理系统改善了地方登记处的工作流程，但其能否推广至报告规范各异的不同司法辖区，仍未得到充分了解。我们首次开展跨省评估，研究如何调整在不列颠哥伦比亚癌症登记处（BCCR）开发的领域适配型Transformer模型BCCRTron，以及生物医学Transformer模型GatorTron，用于加拿大癌症监测。我们的训练数据集包括纽芬兰与拉布拉多癌症登记处（NLCR）约104,000份和22,000份去标识化病理报告，分别用于一级（癌症与非癌症）和二级（需报告与无需报告）任务。两个模型均采用互补的、侧重结构化摘要和诊断的报告章节输入管道进行微调。在NLCR测试集上，适配后的模型保持了较高性能，表明在一个司法辖区预训练的Transformer模型只需少量微调，即可高效地适配至另一个司法辖区。为进一步提高敏感度，我们采用保守的或集成策略组合了这两个模型。该集成模型在一级任务中达到0.99的召回率，漏诊癌症数降至24例，而单个模型分别为48例和54例。在二级任务中，集成模型达到0.99的召回率，将漏掉的需报告癌症数降至33例，而单个模型分别为54例和46例，同时保持较高的精确率（0.99）和F1值（0.99）。这些发现表明，结合互补文本表示的集成模型可显著减少漏诊癌症

并提高癌症登记处自然语言处理中的错误覆盖能力。我们实施了一种隐私保护工作流程，仅在各省之间共享模型权重，而不共享患者级数据，从而支持构建互操作的自然语言处理基础设施，并为未来开发用于癌症病理学和登记处工作流程的泛加拿大基础模型奠定基础。

1 引言

癌症是全球发病和死亡的首要原因，仅2020年就造成近1,000万人死亡 [1]。癌症预防和控制依赖于基于人群的癌症登记处提供的及时、高质量数据，这些数据是监测、政策与项目评估以及人群肿瘤学研究的基础 [2]。登记处从病理报告中收集数据，病理报告是确认癌症的主要原始记录 [3, 2]。然而，人工抽取流程耗时且耗费大量资源，并会导致报告和数据获取延迟数月；在许多情况下，延迟时间可达确诊后24个月 [4]。

自然语言处理（自然语言处理）和大型语言模型（大型语言模型）已成为从病理文本中自动进行信息抽取的可行方案 [4, 5, 6], 。Santos等人证实，利用深度学习自然语言处理进行癌症病理学分析正日益受到关注[4]。不列颠哥伦比亚癌症登记处（BCCR）开发了基于深度学习自然语言处理的流水线，采用了BlueBERT等领域适配的变换器模型[7] 以及GatorTron[8], ，这些模型均在本地病理报告上进行了微调。BCCR在需上报肿瘤分类中的召回率超过98%，显著优于传统的基于规则的系统[5]。此外，BCCR还开发了由小型和大型语言模型组成的集成，用于分类病理报告中记录的肿瘤解剖部位[6]。

此类模型能否在不同司法辖区间泛化，目前尚不明确。加拿大各省使用不同的实验室信息系统，叙述惯例也可能有所差异，尽管核心医学术语基本已实现标准化。规模较小或资源较少的登记处，若能重复利用现有且经过充分验证的模型，而不是从头构建，将大有裨益。此前的研究展示了迁移学习的潜力，即将病理信息提取模型从一个癌症登记处迁移至另一个登记处[9]。此外，新兴研究提出，联邦学习可作为一种策略，让各机构在保护数据隐私的同时协作训练模型[10]。然而，鲜有研究在省级癌症登记处和真实环境中测试这些方法。

本研究介绍了一项试点项目，旨在评估将 BCCR 的自然语言处理流水线调整后用于纽芬兰与拉布拉多癌症登记处（NLCR）的可行性。我们的目标是复现 BCCR 用于判定肿瘤上报属性的双层分类架构[5]:

- 一级（T1）：将病理报告分类为“癌症”或“非癌症”；
- 二级（T2）：在癌症报告中，识别应向登记处“上报”的报告。

我们评估由两个微调语言模型 BCCRTron 和 GatorTron 组成的保守集成，能否在 NLCR 数据上准确执行 T1 和 T2 分类。

2 方法

2.1 数据集描述

我们的训练数据集分别包含约 104,000 份和 22,000 份来自 NLCR 的去标识化电子病理报告，用于 T1 和 T2 分类任务。这些报告由 NLCR 肿瘤数据专员整理、标注并准备。所有记录在处理和分析前均已去标识化。每份报告都包含多个部分，例如“结构化摘要”“诊断”和“标本”，不同医院和时期的报告结构与内容各不相同。这些报告从诊断年份为 2022 年和 2023 年的记录中随机抽取。T1 数据集中，癌症病例占 21%，非癌症病例占 79%。T2 数据集中，可报告病例占 80%，不可报告病例占 20%。

2.2 训练数据

为了优先提高癌症和可报告病例的召回率，使目标与癌症登记处的目标保持一致，并解决类别不平衡问题，我们采用有放回的随机欠采样构建训练数据集。对于 T1，非癌症病例按照以下公式进行欠采样：

$$N(noncancer) = N(cancer) * 0.8 \tag{1}$$

对于 T2，采用以下方法对不可报告病例进行欠采样：

$$N(reportable) = N(nonreportable) * 1.2. \tag{2}$$

这种方法使 T1 的训练数据集规模达到 n=31,694，T2 的训练数据集规模达到 n=7,821。

2.3 训练数据

为确保评估稳健，我们使用了涵盖不同时期的多个测试数据集，分别评估 T1 和 T2 模型。初始测试数据集采用完整数据集（欠采样前）按 80-20 的比例进行训练/测试划分，T1 的样本量为 n=20,777，T2 的样本量为 n=4,409。这样可以确保初始测试集反映真实世界数据的实际分布。此外，我们还使用了不属于原始数据集的其他测试集进行进一步评估；这些测试集取自诊断年份为 2024年的数据，T1 的样本量为 n=18,549，T2 的样本量为 n=3,078。T1 的额外测试集中，癌症/无癌症分布为 17%/83%；T2 的额外测试集中，可报告项/不可报告项分布为 100%/0%。

2.4 模型与微调

我们采用了两个预训练的基于变换器的语言模型：

- GatorTron[8], 一种大型生物医学语言模型，基于大量临床和电子健康记录语料库训练而成；
- BCCRTron[5, 11], 在 BC癌症登记处的病理文本上使用 ∼11 百万份病理报告对 Gatortron[8] 进一步预训练。

每个模型均分别使用各预处理流程的输出在 NLCR数据上进行微调（详见下文），最终共得到四个微调模型：每个层级（T1 和 T2）各两个，分别基于两种输入管道。我们遵循既定最佳实践选择超参数以减少过拟合，包括限制训练轮次[12]；完整详情见补充材料（表 S1）。微调在一台台式机上完成，配备英特尔至强处理器、64 GB RAM 和一块 NVIDIA RTX 4090图形处理器（24 GB 显存）。我们使用其原始名称（即 BCCRTron 和 Gatortron）指代这些微调模型。
此外，我们采用了不同的文本预处理方法，以确保模型输出具有多样性，从而实现高效的集成，具体如下。

2.4.1 预处理与输入流程

我们通过转换为小写、移除标点符号，并使用特定于模型的分词器进行分词，对文本进行了规范化处理。为支持集成策略，我们设计了两种不同的预处理流程，使每个模型在处理同一份病理报告时采用不同方式。其目标是训练出能够关注输入不同方面的模型，从而提高组合使用时的稳健性。

- 流程A提取并优先处理结构化概要部分，该部分通常包含符合标准化癌症报告格式的半结构化清单式数据。该流程被用作 BCCRTron 的输入。
- 流水线 B 提取并优先处理诊断部分，该部分由自由文本形式的组织病理学描述和临床印象组成。该流程被用作 GatorTron 的输入。

这种双流水线方法确保每个模型在微调过程中都能接触到互补信息。其依据是：如果一份报告确实属于负类（例如，T1 中的非癌症病例或 T2 中的不可报告病例），那么尽管两个模型是在报告的不同视角上训练的，它们仍会分别将其判定为负标签。相反，如果只有一个模型检测到相关信号，集成仍可将该报告标记为阳性，这与我们最大化敏感度的目标一致。

2.4.2 分类任务与集成策略

我们针对两个分类层级采用了相同的架构和建模方法：

- T1：将每份报告分类为“癌症”或“非癌症”；
- T2：在被分类为癌症的报告中，判断该报告所述癌症是否属于登记处的“可报告”癌症。

对于这两个层级，我们均采用了结合 BCCRTron 和 GatorTron 的集成策略。如果给定流程中的两个模型至少有一个预测为正类，则将报告标记为阳性（根据层级不同，表示癌症或可报告）。

我们选择了 Gondara等人（5）提出的“或”集成策略，以优先提高敏感度并尽量减少假阴性，这与登记处的目标一致。

2.4.3 评估

我们使用留出测试集和额外测试集评估模型性能。对于 T1，癌症与非癌症的金标准标签由经过培训的 NLCR 工作人员人工抽取确定，这也是实际操作中的金标准。对于 T2，真实标签取自一部分癌症阳性报告，这些报告经人工标注为可报告或不可报告。我们报告两个层级的标准性能指标：准确率、敏感度、特异度、精确率以及 F1 分数（微观和宏观）。

3 结果

3.0.1 T1：癌症与非癌症分类

所有模型的性能都很高；以两位小数表示时，在我们优先关注的召回率方面，各模型表现 indistinguishable。对于癌症标签，集成模型的召回率为 [99%], 精确率为 [88%], F1 分数为 [93%]。详细结果见表1。

在 NLCR 数据上进行微调后，BCCRTron 的召回率为 [99%]，精确率为 [91%] 。Gatortron 模型（在 NLCR 数据上微调的 GatorTron）召回率为 [99%]，精确率为 [89%] 。
从漏诊癌症的绝对数量来看，集成模型表现最佳，其漏诊病例数仅为其他模型的一半；组合模型漏诊 24 例，而 Gatortron 和 BCCRTron 分别漏诊 48 例和 54 例。请注意，在其他测试集上也观察到了类似的表现。

表1：单个模型和集成模型在 T1 任务上的性能指标。

模型（类别）	召回率	精确率	F1 分数
GatorTron（非癌症）	0.97	1.00	0.98
GatorTron（癌症）	0.99	0.89	0.94
BCCRTron（非癌症）	0.97	1.00	0.99
BCCRTron（癌症）	0.99	0.91	0.95
组合（非癌症）	0.96	1.00	0.98
组合（癌症）	0.99	0.88	0.93

3.0.2 T2：需报告与无需报告分类

对于 T2 分类，在癌症阳性的初始 NLCR 测试集的可报告类别上，集成模型的召回率为 [99%] ，精确率为 [99%] ， F1 分数为 [99%]。而微调后的 BCCRTron 和 Gatortron 的召回率、精确率和 F1 分数分别为 [99%、100%、99%] 和 [98%、100%、99%] 。详细结果见表 2。

与T1一样，集成模型在平衡精确率的同时，实现了最高的召回率，优于任何单个模型。组合模型漏检了33条需报告记录，而GatorTron和BCCRTron分别漏检了54条和46条。这再次体现了登记处工作流程中对召回率的预期偏重。值得注意的是，在其他测试集上也观察到了类似的性能。

表 2：单个模型和集成模型的 T2 性能指标。

模型（类别）	召回率	精确率	F1 分数
GatorTron（不可报告）	0.99	0.94	0.96
GatorTron（可报告）	0.98	1.00	0.99
BCCRTron（不可报告）	0.99	0.95	0.97
BCCRTron（可报告）	0.99	1.00	0.99
组合（不可报告）	0.98	0.96	0.97
组合（可报告）	0.99	0.99	0.99

4 讨论

本研究证明了在加拿大各省级癌症登记处之间共享和调整自然语言处理模型的可行性。据我们所知，这是首个涉及直接复用并本地调整领域专用语言模型、用于癌症监测的省际合作案例。通过部署由BCCR开发的即插即用模型架构，并在NLCR数据上进行微调，我们在T1（癌症与非癌症）和T2（需报告与无需报告）分类任务中均取得了良好性能。该集成模型由两个经过领域调整且输入管道互补的Transformer模型构建，在敏感度方面优于单个模型，同时保持了可接受的特异度。这一性能和我们采用的方法表明，在各癌症登记处大规模部署自然语言处理技术，有望采用一种实用、低成本且稳健的路径。

我们的研究发现进一步印证了越来越多的研究成果，这些研究表明迁移学习在临床自然语言处理中具有重要价值，尤其是在病理信息抽取方面。Alawad等人评估了多种跨登记处迁移策略，发现使用第二个登记处的数据对共享模型进行微调，可显著提升信息抽取性能，尤其是对罕见实体的抽取效果。这一做法是在分布式隐私保护设置下实现的，其准确率接近集中汇总模型[9]。在本研究中，我们发现，基于不列颠哥伦比亚省癌症病理文本进行预训练的BCCRTron，在经过微调并应用于纽芬兰和拉布拉多省数据时表现良好。这支持了一个更广泛的观点：在某一机构环境中训练的Transformer模型，可以通过微调有效地适配至另一机构，从而在不同司法辖区高效复用高性能基础设施。

我们研究的另一项关键设计是采用集成架构，将两个分别基于不同文本输入训练的语言模型结合起来。BCCRTron优先处理结构化摘要部分，而GatorTron则侧重于诊断文本，从而提供对同一份报告的不同视角。这种设计提高了表征多样性，使集成模型能够识别单个模型可能遗漏的病例。为契合登记处的工作重点，我们采用“或”逻辑以最大限度提高敏感度，即便这会在一定程度上降低特异度。这与Gondara等人的既往研究一致：他们将BlueBERT和GatorTron集成后应用于不列颠哥伦比亚省病理报告，在肿瘤报告性分类中实现了超过98%的敏感度[5]。其他研究表明，集成方法能够提升临床自然语言处理任务的稳健性、性能和泛化能力[13, 14, 15]。然而，集成方法也存在重要局限，包括计算成本增加和可解释性降低[16, 17]。此外，只有当组成模型具备足够的多样性时，集成方法才能发挥最佳效果；这也支持了我们的做法，即基于不同文本视角训练模型，以充分发挥互补优势并提升整体稳健性[14]。

保护数据隐私并尊重司法辖区数据边界，是医疗领域面临的核心挑战。在本研究中，所有病理报告均在训练前完成彻底去标识化，模型微调也完全在省级辖区内进行。共享和更新的仅有模型权重，因此无需跨司法辖区传输数据即可开展协作。这体现了一种务实且保护隐私的替代方案，可适用于分布式学习场景[9, 18]。联邦学习（FL）常被视为一种规范化方案，使机构能够在不共享数据的情况下开展协作式模型开发[19, 20]。尽管前景可期，联邦学习在临床自然语言处理中的实际应用仍面临多项挑战，例如协调和管理跨站点的实时训练、统一模型架构，以及降低通信开销[21, 20] 。我们采用的方法仅共享模型权重，而训练完全在省级辖区内进行，为实现这些目标提供了一条更简单、务实的途径。目前，研究人员仍在积极探索减轻联邦学习实际应用负担的方法[21, 22]。

4.1 局限性

本研究存在若干局限性。首先，研究仅关注将模型适配到单一目标登记系统，仍需在不同司法辖区开展进一步验证，以评估其泛化能力。其次，尽管集成方法提高了敏感度，但也增加了计算复杂度，可能影响其在资源受限环境中的部署。随着我们探索更广泛的实施方式，解决这些局限性将是未来工作的重点。

4.2 未来方向

本研究指出了一条切实可行的途径，可推动加拿大癌症登记处及整个医疗保健领域开展协作式自然语言处理。虽然本研究侧重于将现有模型调整至一个新的司法辖区，但同一框架也可扩展应用，支持通过分布式或联邦方法共同开发泛加拿大基础模型。要大规模实现这一目标，不仅需要开发模型，还需要共享数据标准、统一的评估方案，以及关于模型共享和隐私保护的正式协议。此外，将这些模型纳入运营工作流程，还需要与登记系统信息系统相衔接，并由领域专家持续验证。重要的是，这项工作符合《泛加拿大癌症数据战略》的目标[23], , 该战略倡导创新、互操作性和及时获取癌症数据，以支持人群健康。尽管如此，我们的研究结果表明

跨司法辖区复用语言模型不仅在技术上可行，也与国家层面推动卫生数据基础设施现代化、提高癌症监测及时性的工作方向一致。

4.3 结论

总之，本研究提供了切实证据，表明预训练自然语言处理模型可以跨司法辖区调整，以支持及时、自动化的癌症监测。通过在保留数据本地化的同时，支持复用高性能基础设施，我们的方法为加拿大及其他地区的登记处工作流程现代化提供了一条可行且可扩展的途径。

4.4 致谢

我们感谢纽芬兰与拉布拉多癌症登记处的 Jennifer Gushue 及其团队，感谢他们提供专业知识、数据准备支持，并为本研究作出重要贡献。我们还感谢 BC癌症登记处的领域专家和登记处领导层给予的支持。本研究由加拿大抗癌伙伴关系通过泛加拿大癌症数据战略提供支持，其资助促成了跨司法辖区合作，并推动了癌症登记处现代化方法的发展。

参考文献

[1] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, 等。2020年全球癌症统计：Globocan 对全球 185 个国家 36 种癌症的发病率和死亡率估算。《*CA*：临床医师癌症杂志》，71(3):209–249，2021。[2] Donald Maxwell Parkin。基于人群的癌症登记处的发展历程。《自然评论：癌症》，6(8):603–612，2006。[3] William Blumenthal, Taiwo O Alimi, Sophia F Jones, Deborah E Jones, Julie D Rogers, Vicki B Benard, 等。利用信息学改进癌症监测。《美国医学信息学会杂志》，27(9):1488–1495，2020年。[4] Tiago Santos, Adeel Tariq, Judy W Gichoya, Hitanshu Trivedi, 和 Imon Banerjee。癌症病理报告的自动分类：系统综述。《病理学信息学杂志》，13:100003，2022。[5] Lovedeep Gondara, Jonathan Simkin, Greg Arbour, Shamila Devji, 和 Ryan Ng。在基于人群的癌症登记环境中，使用语言模型对非结构化电子病理报告中的肿瘤报告状态进行分类。《*JCO*临床癌症信息学》，8:e2400110，2024年。[6] Lovedeep Gondara, Jonathan Simkin, Shamila Devji, Greg Arbour, 和 Ryan Ng。ELM：用于从病理报告预测肿瘤组别的语言模型集成，2025年。[7] Yifan Peng, Shankai Yan, 和 Zhiyong Lu。生物医学自然语言处理中的迁移学习：在十个基准数据集上评估 BERT 和 ELMo。载于第*18*届生物自然语言处理研讨会与共享任务论文集，第58–65页，意大利佛罗伦萨，2019年。计算语言学协会。[8] Xiang Yang, An Chen, Nima PourNejatian, Hoo-Chang Shin, Kirk E Smith, Christopher Parisien, 等。GatorTron：用于从非结构化电子健康记录中提取患者信息的大型临床语言模型，2022年。[9] Mansur Alawad, Song Gao, Jiayi Qiu, Noah Schaefferkoetter, Jacob D Hinkle, Hong-Jun Yoon, 等。跨癌症登记处的深度迁移学习：从病理报告中进行信息抽取。载于*2019 IEEE EMBS*生物医学与健康信息学国际会议（*BHI*），第1–4页。IEEE，2019年。[10] Sebastian Festag 和 Cord Spreckelsen。用于在真实世界数据中检测受保护健康信息的隐私保护深度学习：比较评估。《*JMIR*形成性研究》，4(5):e14064，2020年。[11] Lovedeep Gondara, Jonathan Simkin, Geoffrey Sayle, Shamila Devji, Greg Arbour, 和 Ryan Ng。小型还是大型？零样本还是微调？为医疗保健专业应用选择语言模型提供指导，2025年。[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, 和 Kristina Toutanova。BERT：用于语言理解的深度双向变换器预训练，2019年。[13] Zhen Li, Qian Wei, Li-Ching Huang, Jiaqi Li, Yusheng Hu, Yung-Szu Chuang, 等。集成预训练语言模型，从文献中抽取生物医学知识。《美国医学信息学会杂志》，31(9):1904–1911，2024年。[14] Bahareh Naderalvojoud 和 Tina Hernandez-Boussard。利用观察性医疗保健数据上的集成学习改进机器学习。载于*AMIA*年度研讨会论文集，第521–529页，2024年。

[15] Hao Yang, Mingxin Li, Hua Zhou, Yan Xiao, Qiang Fang, 和 Rui Zhang。一个大语言模型还不够：发挥集成学习的优势以实现医学问答，2023年。medRxiv。[16] Muslim A Ganaie, Meng Hu, Arif K Malik, Mohsin Tanveer, 和 Ponnuthurai N Suganthan。集成深度学习：综述。《人工智能工程应用》，115:105151，2022年。[17] Abdul Mohammed 和 R Kora。集成深度学习全面综述：机遇与挑战。《沙特国王大学学报——计算机与信息科学》，35(2):757–774，2023年。[18] Filip Zerka, Soha Barakat, Sean Walsh, Marta Bogowicz, Ralph TH Leijenaar, Agnes Jochems, 等。医疗保健领域基于联邦数据库的隐私保护分布式机器学习系统综述。《*JCO*临床癌症信息学》，(4):184–200，2020年。[19] Lina Peng, Guozheng Luo, Shun Zhou, Jia Chen, Zheng Xu, Jun Sun, 等。对联邦学习在生物医学自然语言处理信息抽取中的应用进行深入评估。《*npj*数字医学》，7(1):127，2024年。[20] Fan Zhang, Daniel Kreuter, Yichen Chen, Sören Dittmer, Samuel Tull, Tolou Shadbahr, Jacobus Preller, James H. F. Rudd, John A. D. Aston, Carola-Bibiane Schönlieb, Nicholas Gleadall, 和 Michael Roberts。联邦学习在医疗保健领域的最新方法学进展。《*Patterns*》，5(6):101006，2024年。[21] Min Li, Peng Xu, Jing Hu, Zhi Tang, 和 Guang Yang。从挑战与陷阱到建议与机遇：在医疗保健领域实施联邦学习。《医学图像分析》，101:103497，2025年。[22] Yang Li 和 Lei Zhang。选择性注意力联邦学习：提升临床文本分类的隐私性与效率。*arXiv*，2025年。[23]加拿大抗癌伙伴关系。泛加拿大癌症数据战略。https://www.partnershipagainstcancer.ca/about-us/corporate-resources-publications/pan-canadian-cancer-data-strategy/，无日期。访问日期：2024-04-19。