人工智能驱动的癌症研究中的可解释链接预测：揭示共同署名模式

Shahab Mosallaie[1], Andrea Schiffauerova[1] 和 Ashkan Ebadi[1,2,*]

[1] CIISE，康考迪亚大学，加拿大[2] 数字技术部，加拿大国家研究委员会，加拿大。

* 电子邮箱：ashkan.ebadi@nrc-cnrc.gc.ca

摘要

人工智能（AI）正在改变癌症诊断与治疗。癌症的复杂性要求具备不同专业知识的各方通力合作，以确保癌症研究的成效。尽管组建有效的跨学科研究团队至关重要，但这仍是一项挑战。了解并预测合作模式，有助于研究人员、组织和政策制定者优化资源配置，推动产生重要影响的研究。我们以合著网络作为人工智能驱动的癌症研究中合作情况的替代指标。我们使用 Scopus 收录的 7,738 篇出版物（2000年—2017年），构建了 36 个相互重叠的合著网络，分别代表新建、持续和中断的合作关系。我们设计了基于属性和基于结构的特征，并构建了四种机器学习分类器。我们使用 Shapley 加性解释（SHAP）进行模型可解释性分析。随机森林在所考察的三类合作关系中均取得了最高召回率。学科相似度得分是一个关键因素，对新建和持续的共同署名模式有正向影响，对中断的合作关系则有负向影响。此外，高生产力和高资历与合作关系中断呈正相关。我们的研究结果有助于组建有效的研究团队、加强跨学科合作，并为战略性政策决策提供参考。

关键词 协作模式，链接预测，机器学习，可解释性，癌症

1. 引言

癌症仍是全球主要死因之一，2020年导致近1,000万人死亡（Ferlay 等，2021），约占全球记录在案死亡人数的六分之一（世界卫生组织，2022）。在加拿大，癌症仍是首要死因，预计约40%的加拿大人一生中会被诊断出患有癌症，约25%可能死于该疾病（加拿大政府，2021）。尽管癌症具有致命性，但早期检测和有效治疗能够成功治愈许多患者及多种癌症（世界卫生组织，2022）。事实上，及时诊断和干预能够遏制癌症进展并改善预后（Crosby 等，2022）。这凸显了采取有效癌症诊断策略和流程的迫切必要性，以拯救无数受这种危及生命的疾病威胁的人（Okoli 等，2021）。

先进癌症治疗技术的快速发展，进一步凸显了跨学科协作的重要性。例如，基于纳米颗粒的药物递送系统已成为一种颇具前景的方法，有望提高肿瘤学治疗的精确性并降低全身毒性。这些系统可以经过设计，精准靶向特定肿瘤部位，从而实现治疗药物的可控释放，并有望改善患者结局（Cheng 等，2021）。此类技术的开发需要肿瘤科医生、材料科学家、生物医学工程师和数据科学家密切合作，尤其是在设计、测试和优化阶段。

医疗专业人员之间的合作可以改善健康结果（Ebadi 等，2017），而知识共享和决策改善则有助于促进这种合作。癌症护理是一项复杂且碎片化的工作（Ullgren 2021），需要医疗服务提供者与患者携手合作，形成紧密协作的团队，以提供高质量的护理（Amafah 等，2023）。此外，有效的癌症护理涵盖诊断和治疗，需要来自不同学科的研究人员和专业人员通力合作，各自贡献独特的专业知识（Knoop、Wujcik 和 Wujcik 2017）。许多人工智能驱动的医疗研究项目本质上具有国际性，汇聚多国团队，整合多元的专业知识、资源和观点，以应对复杂的科学与临床挑战。

人们投入了大量精力来理解研究合作的动态，并确定促进研究合作的因素。一些研究将创新和科学发现归功于个别研究人员之间的合作努力（例如，Beck 等，2022年）。研究还强调了合作对科研生产力的积极影响（例如，Ebadi 和 Schiffauerova，2016年）。共同署名常被用作科学合作的替代指标（Ebadi 和 Schiffauerova，2015年；Essers、Grigoli 和 Pugacheva，2022年），大规模文献书目数据的可获取性进一步推动了这一趋势。关于共同署名决定因素的研究通常聚焦于两个视角：一个基于社会动态（例如，Lalli、Howey 和 Wintergrün，2020年），另一个则源自合著网络（例如，Ebadi 和 Schiffauerova，2015年）。这两个视角彼此关联，表明合著网络可能源自潜在的社会动态。

在宏观层面，合作研究无论是否形成共同署名，均由三个主要因素驱动：知识、经济和社会因素（Thakur、Wang 和 Cozzens，2011年）。知识驱动因素通常表现为专业化和多学科合作（Thakur 等人，2011年）。经济因素（如充足的研究经费）能够促进更多合作与资源共享（Davies 等人，2022年）。社会互动对于知识获取和职业发展都至关重要。例如，Thakur 等人（2011年）通过说明初级研究人员如何与资深同事合作来获取资源（如设备和数据），以及如何通过与知名科学家建立合作关系来获得声望，凸显了社会因素的影响。此外，其他研究表明，高生产力的研究人员往往倾向于与生产力水平相近的同行合作（Moed、Glänzel 和 Schmoch，2005年）。

尽管 Yu 等人（2014年）探讨了医学领域中的共同署名链接预测，但目前仍缺乏专门针对从事人工智能应用于癌症诊断与治疗研究的研究人员共同署名网络的研究。鉴于癌症的复杂性，以及来自不同学科的研究人员开展合作的必要性，这一研究领域至关重要。本研究旨在揭示从事人工智能驱动的癌症研究的研究人员之间形成不同共同署名模式的影响因素。我们分析了三种具体模式：

- 新合著模式：此前从未共同署名发表过出版物的作者之间建立合作关系。
- 持续合著模式：曾共同署名发表过一篇或多篇论文的作者之间继续保持合作关系。
- 终止合著模式：曾共同撰写过一篇或多篇论文的作者之间，合作关系终止。

本研究的主要目标是开发可解释的机器学习模型，用于预测人工智能驱动的癌症研究中的三种不同共同署名模式，即新建、持续和中断合作。此外，本研究还旨在利用可解释人工智能技术，识别并解释每种共同署名模式背后的关键驱动因素，为研究人员、机构和政策制定者提供可行的见解。本研究对于加深我们对合著网络内部动态的理解至关重要，尤其是在人工智能驱动的癌症研究这一跨学科领域。这样的理解对于应对癌症诊断与治疗的复杂性至关重要。通过识别并解释影响不同共同署名模式的因素，我们为建立有效的研究合作、优化资源分配以及促进不同科学领域的创新提供了宝贵见解。本研究作出以下贡献：（1）提出一种方法框架，整合基于属性和基于结构的特征，以预测多种共同署名模式；（2）通过应用可解释性技术，实现对影响每种共同署名模式最显著的特征进行识别和排序；（3）不同于主要关注预测新合作的方法，我们的方案同时对新建、持续和中断合作进行建模，从而更全面地呈现合作动态；（4）将该框架应用于人工智能驱动的癌症研究领域的大规模出版物数据集，该领域具有高度跨学科性和显著的社会影响。

本文其余部分的结构如下：第2节概述相关工作。第3节详细介绍数据流程。第4节 प्रस्तुत方法论。第5节展示并讨论研究结果，结合支持性数据呈现对新建、持续和终止合著模式的分析。第6节总结全文。第7节介绍研究的局限性及未来研究方向。

2. 相关工作

相关文献采用了访谈、调查和社会网络分析（社会网络分析）等多种方法，以识别影响共同署名的因素。在合著网络中，节点代表作者，两个作者之间的链接表示一篇联合发表的出版物。这些网络是动态的，会随着新作者和新链接的加入而不断演变，同时也可能出现合作关系的终止。链接预测是根据网络中的现有连接和网络特征，估计节点之间未来建立连接可能性的过程（Hasin 和 Hassan，2022年）。在合著网络中，具有相似性的研究人员更有可能在未来展开合作。预测得到的共同署名链接可为组建稳健的研究团队提供指导，并为合作提供战略方向（Pavlov 和 Ichise，2007年）。

在另一项研究中（Liben-Nowell 和 Kleinberg，2003年），研究人员分析了 arXiv 上涵盖 1994 年至 1999 年的五个合著网络，并采用基于相似度的指标预测未来的共同署名链接。研究结果表明，在预测共同署名链接方面，大多数基于相似度的指标都优于各种基线方法。Pavlov 和 Ichise（Pavlov 和 Ichise，2007年）将多种基于相似度的指标整合为一个特征向量，并运用监督式机器学习（机器学习）来预测新合著，从而提出了一种改进的链接预测方法。他们的研究聚焦于 1993 年至 2006 年日本研究人员的合著网络。Hasan 等人（2006年）进一步改进了这一方法，将基于属性相似度的指标纳入特征向量，以提升机器学习模型在共同署名链接预测中的性能。例如，他们考察了由一对作者的出版物关键词得出的交集大小，

并据此得出结论：该值越高，作者之间越相似，因此未来合作的可能性也越大。在另一项研究中，Yu 等人（2014年）从冠状动脉疾病研究的合著网络中提取了各种基于相似度的指标，并将这些特征作为两种不同机器学习模型的输入。他们证明，基于网络结构的特征能够有效预测潜在的未来合作。

在 Chuan 等人（2018年）的一项研究中，研究人员评估了各种基于相似度的链接预测指标，并提出了一种新颖的混合属性特征，以提升合著链接预测的效果。该创新特征采用潜在狄利克雷分配（LDA）主题模型（Blei，2003年）提取出版物主题，并计算作者对的学科相似度得分，从而纳入作者出版物的内容。研究表明，研究人员更倾向于与从事相似领域研究的同行合作。通过分析物理学不同领域的三个合著网络，该研究强调，评估出版物内容对于确定作者和论文之间的相似度得分至关重要，有助于提升共同署名链接预测的准确性。

若干研究采用链接预测和机器学习技术来预测共同署名链接，重点通常是预测此前未曾合作的研究人员之间的合作关系。然而，若忽视曾有合著经历的作者之间合作关系的延续或终止，可能会限制我们对驱动各种共同署名模式之因素的理解。为了让推荐系统有效利用预测模型，这些模型应全面呈现所有可能的共同署名链接，从而促进强有力的研究团队的组建。此外，机器学习分类器如何作出这些预测仍不透明，因此影响共同署名链接预测的关键指标或因素尚未明确。可解释性至关重要，尤其是在将这些模型用于决策过程时，例如战略规划、帮助研究人员识别潜在合作者，以及在组织内组建凝聚力强的研究团队。

3. 数据

为考察参与人工智能驱动的癌症研究的研究人员之间的共同署名模式，数据取自爱思唯尔的 Scopus，涵盖 2000年至2017年的出版物。该数据集包括期刊文章、会议论文、书籍章节和书籍，共计 7,738 篇出版物、46,644 位作者和 123,054 条共同署名链接。如图1所示，出版物数量呈明显上升趋势，且近年来增长速度不断加快。

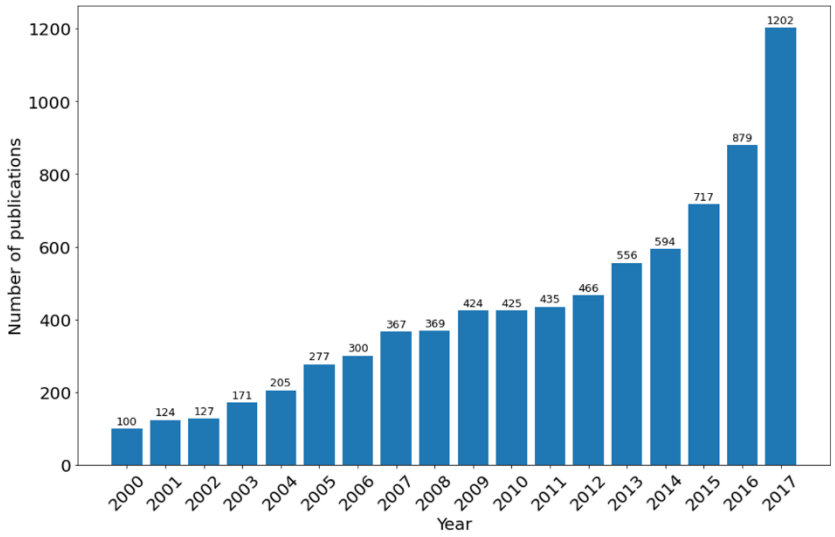


图1. 随时间变化的出版物。

4. 方法论

数据分析工作流由四个关键部分组成。首先，对出版物数据进行预处理，为分析做好准备。其次，针对不同类型构建各种特征。第三部分是开发机器学习模型，以预测共同署名链接。最后，第四部分专门用于解读这些模型，以确定最重要的预测因素。图2展示了分析的高层次概念流程。以下各节将详细说明每个部分。

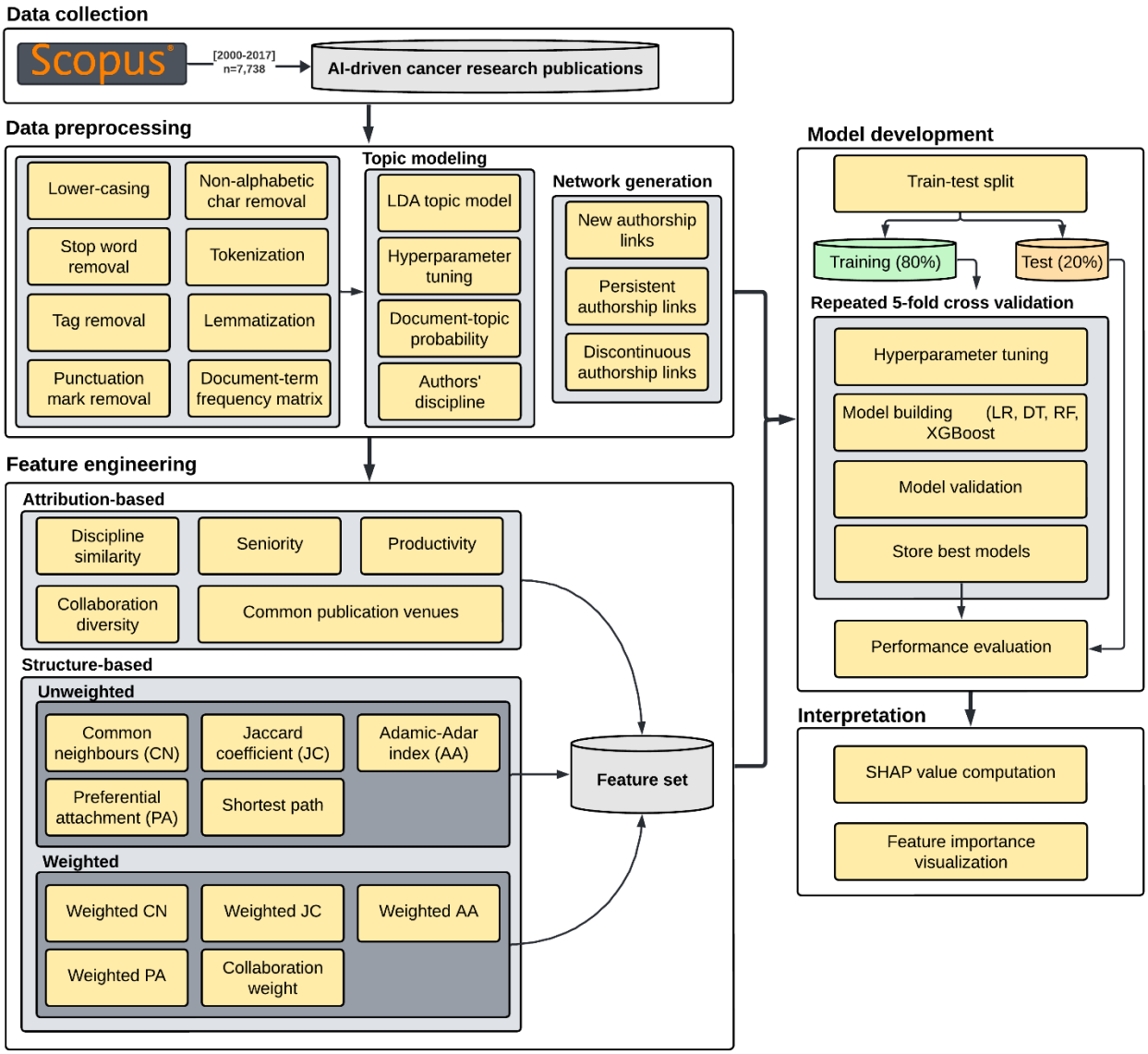


图2。 分析的高层次概念流程。工作流程包括四个主要组成部分：1）数据预处理，2）特征工程，3）模型开发，以及 4）解释。

4.1. 数据预处理我们应用了若干

预处理步骤处理文本数据（即出版物的标题和摘要），包括将文本转换为小写，以及移除特殊字符、停用词、标签和标点符号。随后，对文本数据进行分词和词形还原，生成文档-词项频率矩阵。

在数据准备阶段，我们构建了总计 36 个相互重叠的合著网络，针对新建、持续和中断这三种共同署名模式，各构建了 12 个网络。这一过程采用了类似于（Chuan 等人（2018）；Pavlov 和 Ichise，2007）中所述的方法论。每个网络都使用四年的输入时间窗口，从过去的合著活动中提取特征，并使用三年的预测窗口，为共同署名链接分配标签，以标明其为新建、持续或中断。这种方法确保数据

得到了充分准备，可用于机器学习模型；每个数据集都包含由基于属性和基于结构的指标组成的特征向量，以及用于标识特定共同署名链接的标签。我们关注的是在输入时间窗口和预测窗口期间都活跃的作者，这既全面呈现了合著动态，也确保了数据集中的标签准确完整。

对于新合著模式，标签的分配方式如下：类别1表示未来会出现新的共同署名关系，类别0表示不会出现。在共同署名网络中预测链接时，一个常见的挑战是类别不平衡，即一个类别的数量远远超过另一个类别——例如，在预测时间窗口内，不合作的作者对数量多于合作的作者对。为解决这种不平衡，在测试了多种方法后，我们排除了网络中不存在连通路径的作者对之间的链接，并采用随机欠采样来平衡数据集。对于持续和终止合著模式，我们分别构建了12个网络。在持续模式中，类别1表示合作持续，类别0表示合作终止。相反，在终止模式中，类别1表示合作结束，类别0表示合作继续，这与持续模式的情形正好相反。

4.2. 特征工程我们从预处理数据中提取了两组

特征：(1) 基于属性的特征，以及 (2) 基于结构的特征。

4.2.1. 基于归因的特征这些特征侧重于个体作

者特征，涵盖学科、资历、生产力、合作多样性和出版场所等方面。出版场所指作者发表其作品的具体会议和期刊。此特征集包括：学科相似度得分、资历相似度得分、作者资历水平、生产力相似度得分、作者生产力水平、合作多样性相似度得分、作者合作多样性水平，以及共同发表场所的数量。本节将详细阐述每项特征。

学科相似度：利用文档-词项频率矩阵（详见第4.1节），我们建立了一个 LDA 模型（Blei 2003年），并进行了超参数调优；根据一致性得分（Röder、Both 和 Hinneburg，2015年）以及专家验证，确定六个为最佳主题数量。该模型生成了文档-主题概率数据集，使我们能够通过对作者各篇出版物的特征向量取平均值来确定其学科。随后，我们使用余弦相似度计算作者对的学科相似度得分，以评估其研究兴趣的契合程度。这一做法基于如下观点：作者更倾向于与研究领域相近的同行合作。学科相似度得分介于0到1之间，数值越高表示作者对的学科越相似。

资历：我们将作者的学术生涯年限作为其资历水平的指标，该年限由其首次出版物与最近出版物之间的时长确定。基于此，我们为作者对构建了两个指标：资历相似度得分和资历水平。资历相似度得分评估一对作者在学术生涯年限方面的匹配程度，计算方式为两位作者学术生涯年限之差的绝对值；数值越低，表示资历越相似。相比之下，资历水平指标反映作者对的综合资历，数值越高，表示总体资历越高。该指标的计算方式为作者对中两位作者学术生涯年限之和。

生产力：作者的出版物数量被用作衡量其生产力的指标。基于个体生产力水平，我们为作者对推导出两个指标：生产力相似度得分和生产力水平。生产力相似度得分评估一对作者在生产力方面的一致程度，计算方式为两位作者出版物数量之差的绝对值；数值越低，表示相似度越高。相反，生产力水平指标反映作者对的综合生产力，数值越高，表示总体生产力越高。该指标的计算方式为两位作者出版物数量之和。

合作多样性：度中心性可作为衡量作者合作多样性水平的指标，反映作者参与过的不同合作关系的数量。基于此，我们设计了两个特征：合作多样性相似度得分和合作多样性水平。作者对的合作多样性相似度得分由双方度中心性数值的绝对差值决定，得分越低，表明其合作多样性越相似。相反，合作多样性水平指标反映作者对合作关系的综合多样性，数值越高，表明整体合作多样性越高。该指标通过将双方的度中心性数值相加计算得出。

共同发表场所：出版场所是指作者发表其研究成果的会议和期刊。虽然作者所属学科已明确，但这一特征能更全面地呈现作者的专业领域。它统计作者对共同发表过成果的期刊和会议数量，为了解他们共同的学术兴趣以及未来合作的可能性提供了有价值的见解。

4.2.2. 基于结构的特征这组特征源自共同署名网络的结构。

网络（G）定义为一个元组（$G=(V,E)$），其中（V）表示节点（或顶点）集合，（E）表示连接（V）中节点对的边集合。节点（x）与（y）之间的边记为（$e_{(x,y)}$）。术语（$|V|$）表示网络中的节点总数，而（$|E|$）表示边的总数。节点（x）的邻域记为（Γ_x），指与（x）相连的节点集合；节点的度（$|\Gamma_x|$）则是与其相连的边数。我们的网络是无向的，这意味着所有边都是双向的。

我们的基于结构的指标旨在为每对作者计算一个得分，以反映他们在网络中的相似程度。我们采用了无权重和加权的局部基于相似度的链接预测指标，以全面了解影响各种共同署名模式的因素。为简便起见，我们将这些指标称为无权重和加权的基于相似度的指标。无权重的基于相似度的指标评估网络中两个节点之间的相似度，而不考虑边权重；例如，在共同署名网络中，它们将两位作者之间的多次合作视为一次合作。相反，加权的基于相似度的指标会考虑共同署名链接的强度，并认识到有多次合作经历的作者之间的联系，比合作次数较少的作者之间更紧密。在本研究中，我们采用了若干无权重的基于相似度的指标，包括：1）共同邻居（CN）；2）杰卡德系数（JC）；3）Adamic-Adar指数（AA）；4）优先连接（PA）；以及5）最短路径（SP）。研究人员已有效运用这些指标预测新的共同署名链接（Chuan 等人，2018；Hasan 等，2006；Pavlov 和 Ichise，2007；Yu 等，2014）。

共同邻居（CN）：该指标基于这样的假设：两个节点共享的邻居越多，它们在社交网络中形成链接的可能性就越大。观察到共同邻居的数量与链接形成的可能性之间存在相关性，为这一假设提供了支持（Newman 2001）。共同邻居指标通过统计共享邻居的数量来量化相似度得分，详见方程（1）（Liben-Nowell 和 Kleinberg，2007）：

$$CN = \left|\Gamma_x \cap \Gamma_y\right| \quad (1)$$

其中，Γ_x 和 Γ_y 分别是与作者 x 和 y 合著过的作者集合，$|\Gamma_x \cap \Gamma_y|$ 是 x 和 y 之间共享的合著者数量。如果作者 x 和 y 有许多共同合作者，那么他们很可能属于同一个研究界，这会提高未来合作的可能性。

杰卡德系数（JC）：该指标源自信息检索系统（Jaccard 1901），将相似度得分定义为两个节点的共享邻居数量与其邻居总集合之比，如方程（2）所示：

$$JC = \frac{\left|\Gamma_x \cap \Gamma_y\right|}{\left|\Gamma_x \cup \Gamma_y\right|} \quad (2)$$

在方程中，$\Gamma_x \cup \Gamma_y$ 是 x 和 y 的所有唯一合著者合并后的集合，$|\Gamma_x \cup \Gamma_y|$ 是 x 和 y 之间唯一合著者的总数。在合著网络中，杰卡德系数通过计算一对作者的共享合著者数量与他们合并后的唯一合著者总数之比，来评估这对作者之间的相似度。

Adamic-Adar指数：该指标通过考量一对节点的共享邻居属性，为其计算相似度得分。与共同邻居指标不同，AA指数采用加权方式，根据每个共享邻居 (z) 的度对其进行惩罚，具体如方程 (3) 所示。当节点随着邻居数量增加而降低其在各连接之间分配资源的能力时，这种调整尤其适用（Martínez、Berzal 和 Cubero，2016年）。

$$AA = \sum_{z \in \Gamma_x \cap \Gamma_y} \frac{1}{\log\left|\Gamma_z\right|} \quad (3)$$

其中，Γ_z 是作者 z 的合著者集合，$|\Gamma_z|$ 是 z 的合著者数量。求和遍历所有同时为 x 和 y 的合著者的 z。在共同署名网络中，AA 指数会为总体合著者较少的共享合著者赋予更高权重，凸显了排他性更强的合作关系的重要性（Yu 等，2014年）。

优先连接：PA 指标基于幂律分布和无标度网络原理，这些原理由 Barabási-Albert 网络模型提出（Barabási 和 Albert，1999年）。根据该模型，两个此前未连接的节点之间形成连接的概率会随着各自的度增加而上升。本研究根据方程 (4) 定义 PA 得分，并通过将相关两位作者的度相乘来计算。

$$PA = \left|\Gamma_x\right|\left|\Gamma_y\right| \quad (4)$$

最短路径：该指标用于计算网络中节点对之间所需的最少边数。它基于这样的假设：路径越长（数值越高），两个尚未连接的节点未来形成链接的可能性就越低。

本质上，在网络中相距较远的作者，比彼此距离较近的作者更不容易合作。

与非加权的基于相似度的特征相比，加权的基于相似度的指标通过将合作权重纳入计算，提供了更丰富的细节。合作权重记为 (ω(a,b))，用于量化节点 (a) 和 (b) 之间联系的强度，其定义为两位作者合作发表的合著论文数量（Chuan 等人（2018））。例如，加权共同邻居（WCN）通过纳入先前合作的强度，对共同邻居指标进行了扩展。设 x 和 y 为两位作者，Γ_x 和 Γ_y 分别表示他们的合著者集合。交集 $\Gamma_x \cap \Gamma_y$ 是同时与 x 和 y 合作过的作者集合。WCN 得分的计算方法如下：

$$WCN = \sum_{z \in \Gamma_x \cap \Gamma_y} \frac{\omega(x,z) + \omega(y,z)}{2} \quad (5)$$

此处，分子对每位共同合著者 (z) 与目标作者 (x 和 y) 之间的平均合作权重求和。WCN 得分表明，共同合著者数量较多、且与这些合著者共同发表过更多论文的作者对，未来更有可能开展合作（Murata 和 Moriyasu（2007））。类似地，其他加权的基于相似度的指标，包括加权杰卡德系数（WJC，方程（6））、加权亚当ic-阿达尔指标（WAA，方程（7））和加权优先连接指标（WPA，方程（8）），也基于类似的假设。

$$WJC = \frac{\sum_{z \in \Gamma_x \cap \Gamma_y} \frac{\omega(x,z) + \omega(y,z)}{2}}{\sum_{u \in \Gamma_x} \omega(x,u) + \sum_{v \in \Gamma_y} \omega(y,v)} \quad (6)$$

$$WAA = \sum_{z \in \Gamma_x \cap \Gamma_y} \frac{\omega(x,z) + \omega(y,z)}{2} \frac{1}{\log(\sum_{t \in \Gamma_x} \omega(z,t))} \quad (7)$$

$$WPA = \sum_{u \in \Gamma_x} \omega(x,u) \sum_{v \in \Gamma_y} \omega(y,v) \quad (8)$$

合作权重：我们将合作权重（CW）作为一项独立指标，用于衡量网络中两位作者之间合作关系的强度（Parimi 和 Caragea（2011））。该指标不仅考虑作者之间联合合作的次数，还将其共同发表的论文中涉及的作者数量纳入考量。如方程（9）所述，作者对（x,y）的合作权重会根据参与合作发表的出版物的作者数量进行调整，从而有效惩罚较大的作者群体。

$$CW = \sum_{i=1}^{p} \frac{\sigma_x^i \sigma_y^i}{n_i - 1} \quad (9)$$

在该方程中，p 是合作发表的出版物总数；如果 x 是论文 i 的作者，则 σ_x^i 等于 1，否则为 0。此外，n_i 表示参与论文 i 的作者数量。

4.3. 模型开发我们开发了四个

m

机器学习分类器：逻辑回归（LR）、决策树（DT）、随机森林（RF）和极端梯度提升（XGBoost）。LR 是一种基线线性

模型，而 DT、RF 和 XGBoost 则是能够建模复杂特征交互的基于树的方法。我们评估了这些模型彼此之间以及相对于基线的表现，基线以患病率指标表示。患病率指标表示数据集中正样本所占的比例，并作为随机分类的基线。它反映了完全随机进行预测的分类器的预期准确率。我们将数据集划分为训练集和测试集，其中 80% 的总数据分配给训练集，其余 20% 分配给测试集。我们使用随机搜索对分类器进行超参数调优。为了验证机器学习模型并确定其最优超参数，我们进行了三次重复的分层五折交叉验证。选择这种方法是为了确保性能估计稳健。该问题被设定为二分类任务。我们使用多种指标评估模型，包括精确率、召回率、F1分数、曲线下面积（AUC）和平均精确率（AP）；后者用于计算精确率-召回率曲线下面积。对于二分类任务，尤其是在应对不平衡分类问题时，特别推荐使用 AP（Davis 和 Goadrich，2006年；Saito 和 Rehmsmeier，2015年）。我们通过对多数类进行随机欠采样来处理类别不平衡。

4.4. 解释我们使用了 Shapl

ey 加性解释（SHAP）（Lundberg 和 Lee，2017年）来解释各项特征对模型预测结果的贡献。SHAP 值基于合作博弈论，其中每项特征都被视为一个“参与者”，为“收益”（预测结果）作出贡献。某项特征的 SHAP 值表示该特征在所有可能的特征组合中的平均边际贡献。正 SHAP 值表示该特征会提高形成链接的概率，而负值则表明该特征会降低这一概率。例如，在预测新合作时，较高的学科相似度得分可能会产生正 SHAP 值，从而提高模型对形成链接的信心。针对每种共同署名模式（新建、持续、中断），我们分别应用 SHAP，并选用该模式下性能最佳的模型。

5. 结果

5.1. 新合著

5.1.1. 性能评估表1呈现了性能

机器学习模型预测新合作关系的表现。逻辑回归、决策树、随机森林和 XGBoost 模型的平均召回率分别为 0.77、0.81、0.88 和 0.87。随机森林分类器取得了最高的平均召回率，平均成功预测了 88% 的新共同署名链接。LR、DT、RF 和 XGBoost 分类器的平均 AUC 值分别为 0.86、0.84、0.90 和 0.88，表明所开发模型在预测新共同署名链接方面具有较高性能。正如其他研究（例如 Chuan 等人（2018））所观察到的，在链接预测任务中，较低的精确率较为常见。然而，这些分类器允许调整召回率或精确率，从而使预测模型能够根据具体需求进行定制，但会以牺牲另一项指标为代价。

表1. 预测新合作的模型性能。

网络	模型	召回率	AUC	AP
	LR	1.00	0.94	0.44
1	DT	1.00	0.94	0.23
	RF	1.00	0.96	0.52

	XGB	1.00	0.85	0.30
2	LR	0.68	0.91	0.76
	DT	0.90	0.88	0.55
	RF	0.87	0.92	0.73
	XGB	0.81	0.88	0.69
3	LR	0.41	0.74	0.54
	DT	0.76	0.76	0.38
	RF	0.94	0.87	0.61
	XGB	0.88	0.83	0.52
4	LR	0.63	0.79	0.58
	DT	0.79	0.81	0.38
	RF	0.89	0.89	0.59
	XGB	0.74	0.88	0.52
5	LR	1.00	0.96	0.49
	DT	0.95	0.92	0.25
	RF	0.95	0.97	0.74
	XGB	1.00	0.96	0.45
6	LR	0.76	0.91	0.44
	DT	0.80	0.84	0.24
	RF	0.84	0.92	0.49
	XGB	0.80	0.92	0.33
7	LR	0.71	0.86	0.29
	DT	0.88	0.85	0.22
	RF	0.79	0.89	0.36
	XGB	0.83	0.89	0.31
8	LR	0.85	0.80	0.22
	DT	0.60	0.80	0.17
	RF	0.75	0.85	0.42
	XGB	0.90	0.87	0.40
9	LR	0.83	0.81	0.30
	DT	0.88	0.84	0.21
	RF	0.88	0.90	0.53
	XGB	0.88	0.85	0.29
10	LR	0.95	0.87	0.35
	DT	0.76	0.82	0.24
	RF	1.00	0.91	0.49
	XGB	0.95	0.89	0.41
11	LR	0.94	0.91	0.38
	DT	0.78	0.85	0.25
	RF	0.94	0.89	0.42
	XGB	1.00	0.89	0.34
12	LR	0.50	0.81	0.23
	DT	0.67	0.77	0.09
	RF	0.67	0.82	0.33
	XGB	0.61	0.79	0.28

注：极端梯度提升：XGBoost。

图3比较了各分类器在预测新合著模式时的精确率-召回率曲线下面积（平均精度）与基线的表现。所有分类器均超过了基线，展现出其在预测新共同署名链接方面的有效性。除第二个合著网络外，RF 分类器在大多数合著网络中均表现出更优性能；在第二个网络中，LR 的表现略优于 RF。值得注意的是，RF 模型在第五个合著网络中取得了最高的平均精度值 0.74，显著超过了 0.04 的基线值。

紧随随机森林之后，逻辑回归和XGBoost并列第二，成为预测新合著链接表现第二好的算法。

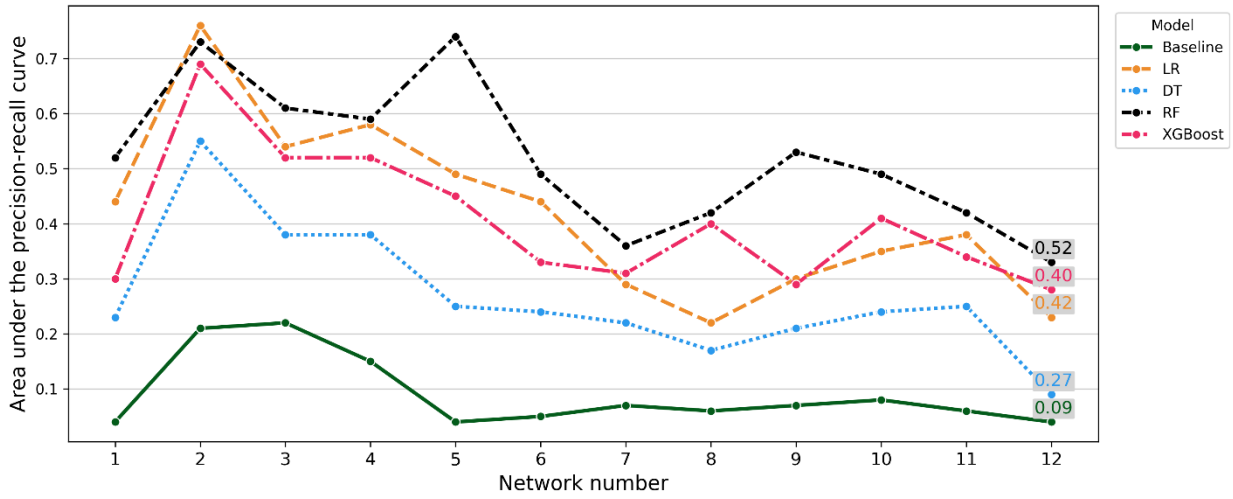


图3。 基于精确率-召回率曲线下面积对预测新合作关系的模型进行性能评估。每条线末端的数值表示平均性能。

5.1.2. 驱动因素我们进行了
S

HAP 分析（参见第4.4节），以确定推动新合著链接预测的关键因素。图4-a和图4-b展示了合著网络2和5中影响最大的因素；由于这些模型在预测新合著链接方面表现出色，因此选取了这两个网络。在这些图中，两种颜色表示各因素的影响：红色表示正向影响，蓝色表示对新合著链接可能性的负向影响。横轴显示各特征的贡献值，特征按贡献值从高到低排列，贡献最大的特征位于最前。图中突出显示了以下关键预测特征：加权/非加权杰卡德系数、加权/非加权Adamic-Adar指数、加权/非加权共同邻居、加权/非加权优先连接、学科相似度得分、作者生产力水平以及生产力相似度得分。

上述识别出的特征分为基于结构的因素和基于属性的因素。基于结构的因素包括杰卡德系数（JC）、Adamic-Adar指数（AA）、共同邻居（CN）和优先连接（PA）等指标，以及它们的加权变体。这些指标是根据合著网络的结构得出的。研究表明，这些基于结构的因素在新合著模式的形成中发挥着重要作用（Liben-Nowell 和 Kleinberg，2007年；Pavlov 和 Ichise，2007年；Yu 等，2014年）。值得注意的是，除加权和未加权的优先连接指标外，所有基于结构的因素都是共同邻居指标的变体。

a) **b)**

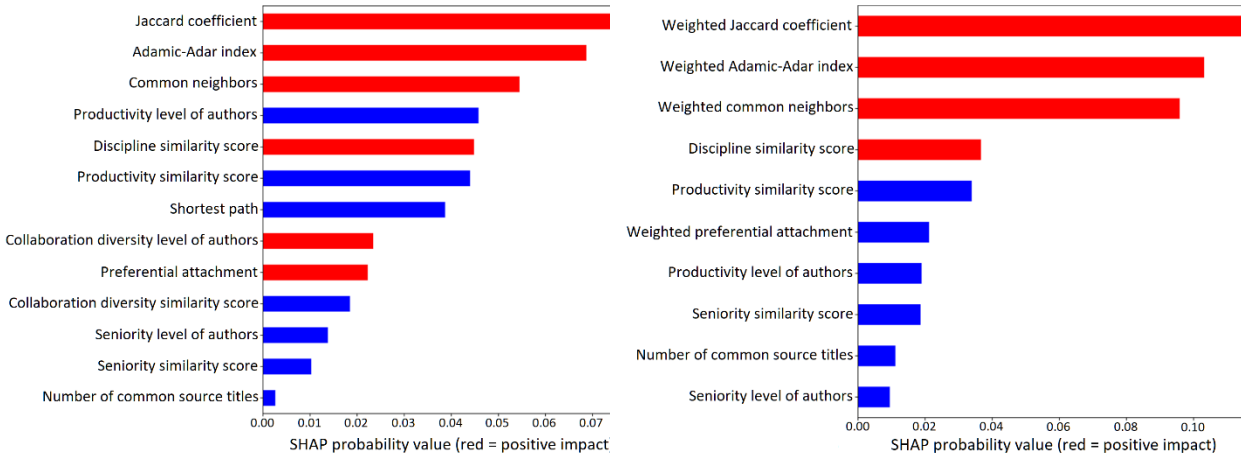


图 4。 SHAP分析，预测新合作关系中最重要的特征：a）网络 2，b）网络 5。

图4-a和图4-b展示了这些基于共同邻居的特征对新共同署名链接的发展具有积极影响。共同邻居指标用于评估两位作者之间共同合著者的数量，这表明共同合著者越多的作者，未来越有可能开展合作。杰卡德系数是另一种基于共同邻居的指标，其计算方式是共同合著者数量与潜在合著者总数之比；杰卡德系数越高，未来合作的可能性越大。Adamic-Adar指数同样基于共同邻居，并对自身连接较少的共同合著者赋予更高权重。它对新共同署名链接的积极影响表明，对于共同合著者数量相同的两组作者对而言，共同合著者连接较少的那组作者更有可能建立共同署名链接。

基于属性的特征，包括学科相似度得分、作者生产力水平和生产力相似度得分，也是新合著模式的关键驱动因素。这些特征源自作者的个人属性，其中学科相似度得分用于评估两位作者在研究领域和学科方面的相似性，并且如图所示，会正向影响新共同署名发生的可能性。这表明，与来自不同学科的作者相比，研究领域相似的作者更倾向于合作。此外，生产力相似度得分和作者生产力水平也是影响新合著模式的重要基于属性的因素。这些指标利用出版物数量来衡量作者的生产力。有趣的是，在大多数网络中，这些生产力指标往往会对新共同署名链接的形成产生负面影响。

5.2 持续共同署名

5.2.1 性能评估

表 2 展示了各种机器学习模型在预测持续合作方面的性能。各分类器的平均召回率如下：逻辑回归为 0.75，决策树为 0.67，随机森林为 0.73，XGBoost 为 0.74。与表1中的分类器相比，这些分类器具有更高的精确率，表明预测的持续共同署名链接中有相当一部分确实发生了，从而凸显了它们在预测持续合作方面的有效性。此外，较高的 AUC 值反映出这些分类器能够准确区分持续共同署名链接的存在与否，彰显了它们在预测任务中的可靠性。

表 2. 预测持续合作关系的模型性能。

网络	模型	召回率	AUC	AP
1	LR	0.79	0.82	0.97
	DT	0.67	0.87	0.97
	RF	0.78	0.90	0.98
	XGB	0.76	0.88	0.97
2	LR	0.81	0.69	0.95
	DT	0.68	0.79	0.95
	RF	0.73	0.84	0.98
	XGB	0.71	0.82	0.97
3	LR	0.74	0.76	0.95
	DT	0.71	0.86	0.96
	RF	0.72	0.87	0.97
	XGB	0.74	0.84	0.96
4	LR	0.74	0.77	0.92
	DT	0.60	0.73	0.90
	RF	0.71	0.80	0.94
	XGB	0.68	0.78	0.93
5	LR	0.73	0.74	0.92
	DT	0.64	0.78	0.92
	RF	0.73	0.84	0.96
	XGB	0.71	0.85	0.97
6	LR	0.61	0.70	0.91
	DT	0.64	0.78	0.92
	RF	0.67	0.79	0.95
	XGB	0.67	0.75	0.93
7	LR	0.82	0.77	0.93
	DT	0.71	0.78	0.91
	RF	0.77	0.85	0.96
	XGB	0.81	0.82	0.95
8	LR	0.78	0.79	0.92
	DT	0.70	0.76	0.91
	RF	0.69	0.81	0.94
	XGB	0.73	0.78	0.94
9	LR	0.79	0.82	0.93
	DT	0.70	0.74	0.9
	RF	0.77	0.85	0.95
	XGB	0.80	0.84	0.95
10	LR	0.72	0.74	0.92
	DT	0.67	0.78	0.93
	RF	0.73	0.81	0.95
	XGB	0.76	0.85	0.96
11	LR	0.67	0.79	0.93
	DT	0.68	0.79	0.93
	RF	0.71	0.86	0.97
	XGB	0.71	0.85	0.96
12	LR	0.74	0.79	0.94
	DT	0.70	0.81	0.94
	RF	0.74	0.85	0.96
	XGB	0.82	0.86	0.96

图5展示了各机器学习模型之间以及与基线相比的平均精度（AP）表现。所有分类器的表现均优于基线，体现了它们在预测持续共同署名链接方面的强大能力。各分类器的表现相近，AP 值介于 0.89 至 0.97 之间。各分类器的平均精确率如下：逻辑回归为

0.91、决策树为0.93、随机森林为0.94，XGBoost为0.93。与新建共同署名模式的预测结果相似，随机森林模型在大多数合著网络中的整体表现通常最佳。不过，在网络5和网络10中，极端梯度提升的表现优于随机森林模型。

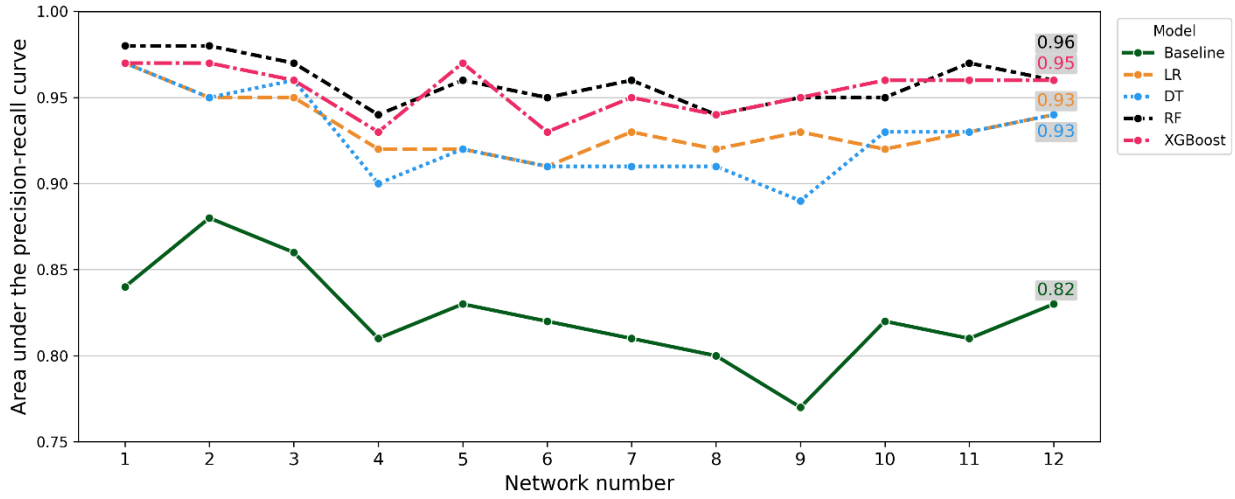


图5. 基于精确率-召回率曲线下面积对用于预测持续合作的模型进行性能评估。每条线末端的数值表示平均性能。

5.2.2. 驱动因素

图6-a和图6-b突出显示了在共同署名网络7和11中预测持续合作时最重要的特征，这两个网络上的模型表现最佳。与新合著模式分析（图4）类似，基于结构和基于属性的指标都成为重要的预测特征。关键预测特征包括作者生产力、学科相似度得分、资历水平、优先连接及其加权变体、合作权重、合作多样性相似度得分、加权/未加权杰卡德系数，以及加权/未加权Adamic-Adar指数。

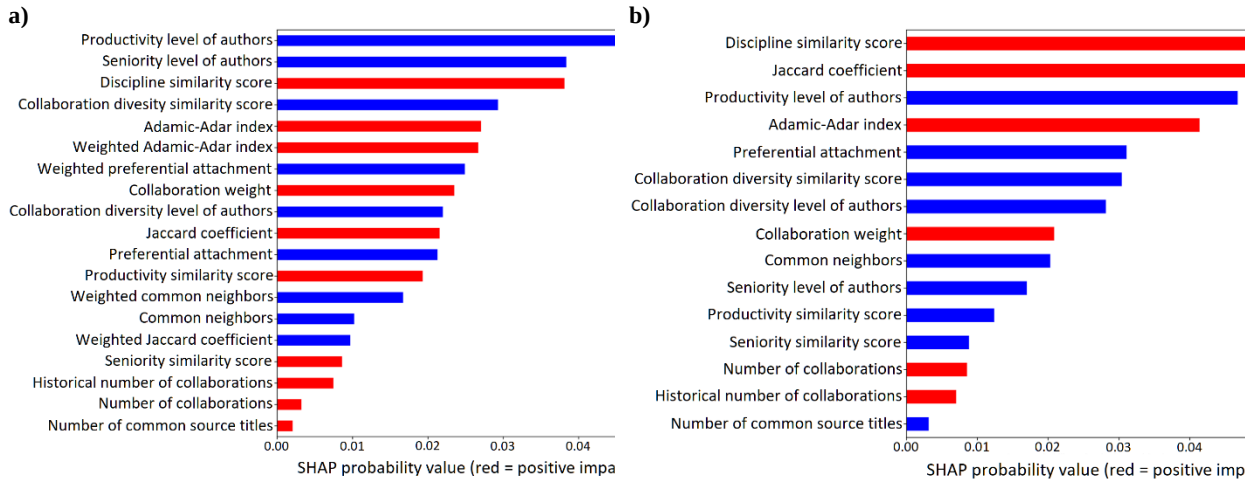


图5. SHAP分析：预测持续合作关系时最重要的特征：a）网络7，b）网络11。

分析发现，生产力水平（定义为作者对发表的出版物总数）会降低维持持续共同署名链接的可能性。相反，学科相似度得分会对持续共同署名模式产生积极影响，这表明研究学科相似的作者对未来更有可能再次合作。另一项基于属性的特征——作者资历，也会对持续共同署名产生负面影响，这表明职业生涯跨度较长的作者对不太可能续约合作关系。同样，优先连接也会对共同署名链接的延续产生负面影响。这种负面影响以及作者生产力水平的影响，可能是因为高产的作者对在研究界更为显眼，因此更容易接触到新的合作者，从而降低了与之前的合作伙伴再次合作的可能性（Yu 等，2014）。

合作权重同时考虑作者对合作发表的出版物数量，以及参与这些出版物的作者人数，对持续的共同署名模式有积极影响。如图6所示，这表明，合作发表的出版物更多、且每篇出版物涉及的作者较少的作者对，更有可能再次合作。此外，作为一种基于结构的特征，合作多样性相似度得分也是持续的共同署名模式的重要预测指标。合作多样性相似度的负面影响可能意味着，作者倾向于与合作多样性水平相近的人再次合作。与推动新合著模式的因素一致，杰卡德系数和Adamic-Adar指数均对持续的共同署名模式有积极影响，这表明共同合著者较多的作者对更有可能再次合作。

5.3. 中断的共同署名

5.3.1. 性能评估

表3展示了各种机器学习模型预测中断合作的性能。平均召回率用于评估模型准确预测作者对之间合作关系中断的能力；逻辑回归的平均召回率为0.63，决策树为0.72，随机森林为0.75，XGBoost为0.73，其中随机森林的平均召回率最高。就平均AUC而言，各模型的表现如下：逻辑回归的AUC为0.74，决策树为0.77，随机森林为0.82，XGBoost为0.81。这些AUC得分反映了模型区分共同署名链接存在与否的能力，其中随机森林的判别能力最高。AUC值越高，表示模型识别真正例和负例的性能越好，这凸显了随机森林和XGBoost在预测中断的共同署名模式方面的有效性。

表3. 预测中断合作的模型性能。

网络	模型	召回率	AUC	AP
1	LR	0.71	0.67	0.40
	DT	0.93	0.89	0.48
	RF	0.79	0.86	0.64
	XGB	0.79	0.86	0.65
2	LR	0.58	0.77	0.44
	DT	0.58	0.74	0.27
	RF	0.68	0.83	0.46
	XGB	0.63	0.77	0.43
3	LR	0.68	0.80	0.48
	DT	0.75	0.82	0.48

	RF	0.82	0.89	0.73
	XGB	0.82	0.89	0.61
4	LR	0.54	0.68	0.40
	DT	0.63	0.66	0.30
	RF	0.80	0.70	0.37
	XGB	0.51	0.64	0.37
5	LR	0.38	0.60	0.38
	DT	0.69	0.75	0.33
	RF	0.67	0.79	0.50
	XGB	0.64	0.76	0.47
6	LR	0.66	0.68	0.41
	DT	0.80	0.72	0.30
	RF	0.70	0.75	0.43
	XGB	0.68	0.77	0.46
7	LR	0.54	0.73	0.51
	DT	0.75	0.78	0.41
	RF	0.77	0.82	0.56
	XGB	0.79	0.83	0.59
8	LR	0.78	0.86	0.68
	DT	0.76	0.79	0.42
	RF	0.84	0.87	0.64
	XGB	0.88	0.86	0.60
9	LR	0.55	0.76	0.62
	DT	0.67	0.76	0.45
	RF	0.65	0.84	0.69
	XGB	0.69	0.79	0.61
10	LR	0.62	0.77	0.49
	DT	0.67	0.80	0.46
	RF	0.73	0.84	0.61
	XGB	0.81	0.85	0.57
11	LR	0.80	0.82	0.54
	DT	0.78	0.82	0.48
	RF	0.81	0.86	0.61
	XGB	0.85	0.88	0.64
12	LR	0.72	0.79	0.46
	DT	0.58	0.75	0.36
	RF	0.70	0.81	0.55
	XGB	0.66	0.79	0.50

图7展示了各分类器相较于基线的平均精确率表现，结果表明，所有分类器在预测中断的合著关系链接方面均超过了基线。其中，随机森林和XGBoost的表现最佳。

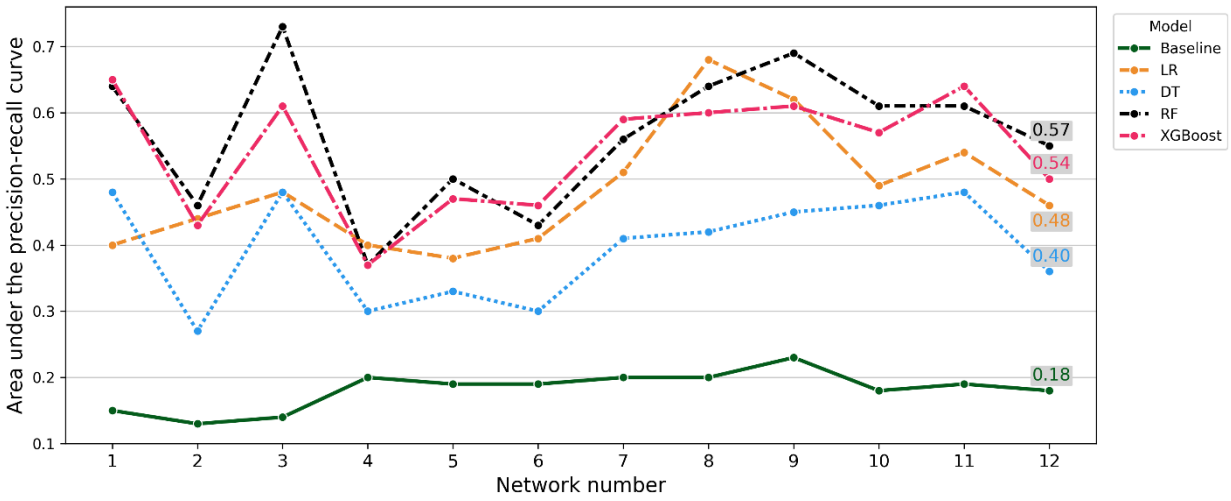


图7. 基于精确率-召回率曲线下面积评估模型预测中断合作的性能。每条线末端的数值表示平均性能。

5.3.2. 驱动因素

图8突显了影响中断合作的关键特征。与持续的共同署名模式中观察到的趋势不同，作者生产力水平会正向影响合作结束的可能性。此外，以作者对的学术生涯年限总和计算的资历水平也是一个重要因素，同样会增加共同署名中断的可能性。这表明，与职业生涯较早阶段的作者对相比，集体职业生涯跨度较长的作者对更有可能结束合作。另一方面，合作权重会降低共同署名结束的可能性。这意味着，合作发表的出版物数量较多、且参与贡献者较少的作者对，不太可能终止合作。这种倾向可能源于这些作者对此前建立的牢固合作纽带，使他们更愿意继续合作。

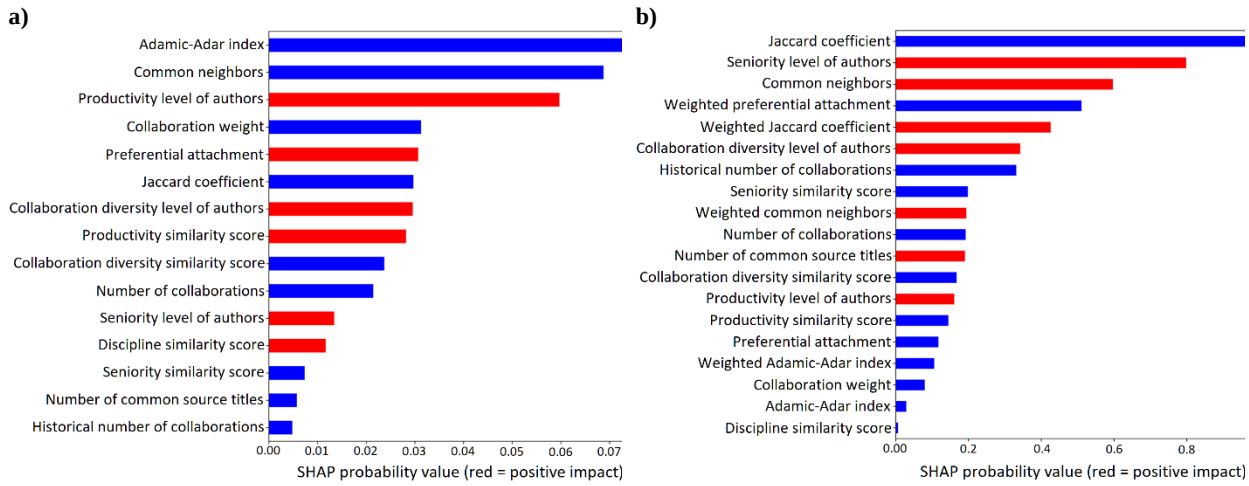


图8. SHAP分析：预测终止的合作关系时最重要的特征：a) 网络3，b) 网络8。

基于结构的指标，包括未加权和加权杰卡德系数、Adamic-Adar指数、共同邻居、优先连接以及加权优先连接，都会影响终止合著模式。然而，这些指标的影响并非始终为正或为负，这表明共同合著者数量较多，并不一定意味着作者对未来停止合作的可能性较低。这可能说明，这些指标与共同署名中断之间的关系较为复杂，并取决于具体情境。

6. 结论

本研究旨在预测各种共同署名链接，并揭示影响研究人员不同共同署名模式的关键因素，同时以“人工智能驱动的癌症研究”为案例，说明我们方法的有效性。通过从合著网络和个体作者中提取基于结构的指标和基于属性的指标，我们将这些特征作为机器学习分类器的输入，用于预测新建、持续和终止的共同署名模式。分类器展现出良好的预测性能，凸显了整合这两类指标对于准确进行共同署名链接预测的价值。

我们的分析表明，基于属性的指标（如作者生产力水平和资历）在塑造这些模式方面发挥着重要作用。虽然高生产力会对持续共同署名产生负面影响，但其与共同署名中断呈正相关，这表明高产作者可能会寻求新的合作关系。同样，资历会正向影响终止的共同署名，说明职业生涯较长的作者可能更倾向于使其合作关系多样化。此外，学科相似度得分成为一个关键因素：它会正向影响新建和持续的共同署名模式，同时负向影响终止模式。这表明，在相似领域工作的作者更有可能发起或重复合作，也较不可能终止合作关系。

基于结构的指标（如杰卡德系数、Adamic-Adar指数和优先连接）也被发现是关键影响因素，不过其影响会因不同的共同署名模式而异。这种复杂性凸显了合著动态的细微特征：拥有共同合著者并不一定能预测持续合作或未来合作。

我们的研究结果强调，在分析作者合作网络时，理解基于属性和基于结构的因素都至关重要。分类器稳健的性能和可解释性，有助于世界各地的研究人员识别潜在合作者，并协助科研机构组建凝聚力强、效率高的研究团队。本研究不仅加深了对合著动态的理解，也支持了共同署名源于社交网络这一观点，为建立稳固的学术合作提供了宝贵见解。

7. 局限性与未来工作

本研究存在若干局限性，主要源于所采用的数据来源。尽管正如Falagas等人（Falagas等人 2008年）所指出的，Scopus和PubMed被公认为医学领域的综合性资源，但纳入其他来源或许能够提供更全面的视角。此外，排除非英语出版物可能限制了数据集的多样性。

需要注意的是，共同署名并不总是意味着合作。作者可能开展了合作，但最终并未形成共同署名；而且，并非所有共同署名论文都源于真正的合作（Ebadi和Schiffauerova 2015年）。在本研究中，基于结构的指标源自“同质”合著网络，该网络仅包含一种节点类型（作者）和一种边类型（共同署名链接）。这些指标被用作机器学习分类器和SHAP方法的输入，以识别合作因素。未来研究可以构建“异质”合著网络，纳入多种节点类型（例如作者、研究机构、论文、领域、出版场所）和边类型（例如论文中的作者角色、隶属关系、出版场所、领域、引文）。正如Hu等人（Hu等人 2020年）所建议的，这些复杂网络能够提取更加细致的特征，从而支持先进的共同署名链接预测技术，并深入揭示推动共同署名的因素。此类发展将显著加深我们对学术合作动态的理解，并改进预测模型。

最后，未来研究可以通过明确将作者隶属关系划分为学术机构、私营机构、政府机构和非营利机构等类别，进一步完善当前框架，从而考察特定部门合作模式，并揭示人工智能驱动的癌症研究中私营—公共合作伙伴关系与纯学术合作之间潜在的结构差异。

参考文献

Amafah、Joy、Tarilate Temedie-Asogwa、Julie Atta 和 Mohammad Ahmmad Mahmoud Al Zoubi。2023年。“治疗摘要对以患者为中心的沟通及癌症幸存者护理质量的影响。”International Journal of Medical and All Body Health Research 04(01):70– 80。doi:10.54660/IJMBHR.2023.4.1.70-80。

Barabási, A. L. 和 R. Albert。1999年。“Emergence of Scaling in Random Networks。”Science 286(5439):509–12。doi:10.1126/science.286.5439.509。

Beck, Susanne, Bergenholtz, Carsten, Bogers, Marcel, Brasseur, Tiare-Maria, Conradsen, Marie Louise, Di Marco, Diletta, Distel, Andreas P., Dobusch, Leonhard, Dörler, Daniel, Effert, Agnes, Fecher, Benedikt, Filiou, Despoina, Frederiksen, Lars, Gillier, Thomas, Grimpe, Christoph, Gruber, Marc, Haeussler, Carolin, Heigl, Florian, Hoisl, Karin, Hyslop, Katie, Kokshagina, Olga, LaFlamme, Marcel, Lawson, Cornelia, Lifshitz-Assaf, Hila, Lukas, Wolfgang, Nordberg, Markus, Norn, Maria Theresa, Poetz, Marion, Ponti, Marisa, Pruschak, Gernot, Pujol Priego, Laia, Radziwon, Agnieszka, Rafner, Janet, Romanova, Gergana, Ruser, Alexander, Sauermann, Henry, Shah, Sonali K., Sherson, Jacob F., Suess-Reyes, Julia, Tucci, Christopher L., Tuertscher, Philipp, Vedel, Jane Bjørn, Velden, Theresa, Verganti, Roberto, Wareham, Jonathan, Wiggins, Andrea, and Sunny Mosangzi and Xu. 2022年。“开放科学创新研究领域：协作概念化方法。” Industry and Innovation 29(2):136–85。doi:10.1080/13662716.2020.1792274。

Blei, David M. 2 003. “潜在狄利克雷分配。”《机器学习研究杂志》993–1022。

Cheng, Zhe, Maoyu Li, Raja Dey 和 Yongheng Chen。2021年。《用于癌症治疗的纳米材料：当前进展与展望》。《血液学与肿瘤学杂志》14(1):85。doi:10.1186/s13045- 021-01096-0。

Chuan, P. M., L. H. Son, M. Ali, T. D. Khang, L. T. Huong 和 N. Dey。2018年。《基于混合内容相似度度量的合著网络链接预测》。《Applied Intelligence》48(8):2470–86。doi:10.1007/s10489-017-1086-x。David Crosby、Sangeeta Bhatia、Kevin M. Brindle、Lisa M. Coussens、Caroline Dive、Mark Emberton、Sadik Esener、Rebecca C. Fitzgerald、Sanjiv S. Gambhir、Peter Kuhn、Timothy R. Rebbeck 和 Shankar Balasubramanian。2022年。《癌症早期检测》。《Science》375(6586):eaay9040。doi:10.1126/science.aay9040。Benjamin Davies、Jason Gush、Shaun C. Hendy 和 Adam B. Jaffe。2022年。《研究资助与合作》。《Research Policy》51(2):104421。doi:10.1016/j.respol.2021.104421。Davis, J. 和 M. Goadrich。2006年。《精确率—召回率与受试者工作特征曲线之间的关系》。Ashkan Ebadi 和 Andrea Schiffauerova。2015年。《如何成为科研合作网络中的重要参与者？》。《Journal of Informetrics》9(4):809–25。doi:10.1016/j.joi.2015.08.002。Ashkan Ebadi 和 Andrea Schiffauerova。2016年。《如何提升科学产出？研究资助及其他影响因素的统计分析》。《Scientometrics》106(3):1093–1116。doi:10.1007/s11192-015-1825-x。Ashkan Ebadi、Patrick J. Tighe、Lei Zhang 和 Parisa Rashidi。2017年。《DisTeam：手术团队选择决策支持工具》。《Artificial Intelligence in Medicine》76:16–26。doi:10.1016/j.artmed.2017.02.002。Dennis Essers、Francesco Grigoli 和 Evgenia Pugacheva。2022年。《网络效应与研究合作：来自国际货币基金组织工作论文合著的证据》。《Scientometrics》127(12):7169–92。doi:10.1007/s11192-022-04335-4。Matthew E. Falagas、Eleni I. Pitsouni、George A. Malietzis 和 Georgios Pappas。2008年。《PubMed、Scopus、Web of Science 和谷歌学术的比较：优势与不足》。《FASEB j.》22(2):338–42。doi:10.1096/fj.07-9492LSF。Jacques Ferlay、Murielle Colombet、Isabelle Soerjomataram、Donald M. Parkin、Marion Piñeros、Ariana Znaor 和 Freddie Bray。2021年。《2020年癌症统计概览》。《International Journal of Cancer》。doi:10.1002/ijc.33588。加拿大政府。2021年。《2021年加拿大癌症统计》。https://www.canada.ca/en/public-health/services/reports-publications/health-promotion-chronic-disease-prevention-canada-research-policy-practice/vol-41-no-11-2021/canadian-cancer-statistics-2021.html。M. Al Hasan、V. Chaoji、S. Salem、M. Zaki 和 N. York。2006年。《使用监督学习进行链接预测》。Hajar Ali Hasin 和 Diman Hassan。2022年。《合著网络中的链接预测》。《Science Journal of University of Zakho》10(4):235–57。doi:10.25271/sjuoz.2022.10.4.1040。Hu, Z.、Y. Dong、K. Wang 和 Y. Sun。2020年。《异构图Transformer》。载于《2020年万维网大会——万维网大会论文集》，2704–10页。WWW。P. Jaccard。1901年。《阿尔卑斯山脉和汝拉山脉一处区域植物区系分布的比较研究》。

Knoop, T., D. Wujcik, and K. Wujcik. 2017. “癌症护理中跨专业协作的新兴模型。” Seminars in Oncology Nursing 33(4):459–63. doi:10.1016/j.soncn.2017.08.009. Lalli, Roberto, Riaz Howey, and Dirk Wintergrün. 2020. “合作网络的动态与广义相对论史，1925—1970年。” Scientometrics 122(2):1129–70. doi:10.1007/s11192-019-03327-1. Liben-Nowell, David, and Jon Kleinberg. 2003. “社交网络的链接预测问题。” Pp. 556–59 in《第十二届信息与知识管理国际会议论文集》，*2003*年信息与知识管理国际会议。美国纽约州纽约市：美国计算机协会。Liben-Nowell, David, and Jon Kleinberg. 2007. “社交网络的链接预测问题。” Journal of the American Society for Information Science and Technology 58(7):1019–31. Lundberg, S. M., and S. I. Lee. 2017. “解释模型预测的统一方法。” Advances in Neural Information Processing Systems 2:4766–75. Martínez, V., F. Berzal, and J. C. Cubero. 2016. “复杂网络链接预测综述。” ACM Computing Surveys 49(4). doi:10.1145/3012704. Moed, Henk F., Wolfgang Glänzel, a nd Ulrich Schmoch, eds. 2005.《定量科学与技术研究手册：在科技系统研究中使用出版物和专利统计数据》。多德雷赫特：施普林格荷兰。Murata, T., and S. Moriyasu. 2007. “基于加权邻近度量的社交网络链接预测。” Newman, M. E. J. 2001. “增长网络中的聚类与优先连接。” Physical Review E 64(2):025102–025102. doi:10.1103/PhysRevE.64.025102. Okoli, George N., Otto L. T. Lam, Viraj K. Reddy, Leslie Copstein, Nicole Askin, Anubha Prashad, Jennifer Stiff, Satya Rashi Khare, Robyn Leonard, Wasifa Zarin, Andrea C. Tricco, and Ahmed M. Abou- Setta. 2021. “改善有症状人群癌症早期诊断的干预措施：范围综述。” doi:10.1136/bmjopen-2021-055488. Parimi, R., and D. Caragea. 2011. “使用主题建模方法预测社交网络中的友谊链接。” 6635 LNAI(PART 2:75–86. Pavlov, M., and R. Ichise. 2007. “通过链接预测在合著网络中寻找专家。” Pp. 42– 55 in CEUR Workshop Proceedings. Vol. 290. Röder, Michael, Andreas Both, and Alexander Hinneburg. 2015. “探索主题一致性度量的空间。” Pp. 399–408 in《第八届ACM国际网络搜索与数据挖掘会议论文集——*2015*年*ACM*国际网络搜索与数据挖掘会议》。美国计算机协会出版社。Saito, T., and M. Rehmsmeier. 2015. “在不平衡数据集上评估二分类器时，精确率-召回率图比受试者工作特征曲线图更具信息量。” PLoS ONE 10(3):1–21. doi:10.1371/journal.pone.0118432.

Thakur, D., J. Wang, and S. Cozzens. 2011. “国际合著衡量的是什么？2011年亚特兰大科学与创新政策会议：建设科学创新与成果能力。” ACSIP。Ullgren, Helena. 2021. “谁来关心？癌症护理的碎片化还是整合。” https://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-188860. 世界卫生组织。2022年。“癌症。” https://www.who.int/news-room/fact-sheets/detail/cancer. Yu, Q., C. Long, Y. Lv, H. Shao, and P. He. 2014. “预测医学合著网络中的合著者关系。” PLoS ONE 9(7):101214. doi:10.1371/journal.pone.0101214.t001.