

提升乳腺癌检测 E FxMammo：一个多体验人工智能评估的结果

使用FxBammo增强乳腺癌检测：多体验人工智能评估结果

东盟和东亚经济研究所 (ERIA) E Sentral Senayan II 6th 亚洲
非洲大道8号，奔卡诺·僧延诺格洛拉，雅加达中心12710印度尼
西亚

© 东南亚及东亚经济研究所, 2025 艾瑞亚研究项目报告2025财年
, 第18号 2025年8月发布

所有权利保留。未经ERIA事先书面通知和许可，本出版物之任何部分均不得以任何形式（无论是电子还是机械手段）进行复制、存储于检索系统中或传送。

其各章节中阐述的研究发现、解释、结论及观点完全属于作者/本人，并不反映东南亚及东亚经济研究所及其理事会、学术咨询委员会或其代表机构的观点和政策。各章节中内容或引用的错误由作者/本人单独负责。

本文中的资料可自由引用或转载，但需注明出处。

本报告由FathomX为东南亚及东亚经济研究所（ERIA）编制，并由日惹公立大学医学院放射科、公共卫生系及护理系支持。

项目成员列表

深测X

斯蒂芬·林 加文·连
杜浩 梅凯尔·哈特曼
冯梦灵 艾比盖尔·朱

加查马达大学，医学、公共卫生与护理学院

琳娜·乔里达 (放射科) 迪迪克·塞蒂奥·哈里扬托 (解剖病理科) 维纳斯·莱曼 (放射科) 依卡·普斯皮塔萨里 (药学院/UGM大学附属医院) 萨琳宁西·希克马瓦蒂 (医学与健康科学博士研究生) 罗赞·穆罕默德·伊尔凡 (放射科研究与培训办公室) 扎努巴·阿里法·努尔 (放射科研究与培训办公室)

内容

	项目成员列表	iii
	表格列表	v
	术语表	vi
第一章	引言	1
第二章	研究方法 5	
第三章	研究 findings	8
第四章	政策建议 12	
第五章	结论 15	
	参考文献 16	

表目录

表3.1 带人工智能和不带人工智能的阅读者诊断性能	9
辅助表3.2 读者诊断的阅卷者间协议	11
带AI辅助和不带AI辅助的性能	

术语表

术语	定义
人工智能 (Artificial Intelligence)	通过模拟人类智能机器。在这项研究中，人工智能协助乳腺钼靶摄片诊断乳腺癌癌症检测。
AUC (曲线下面积)	分类的性能指标模型；更高的AUC表示更好诊断准确率。
BI-RADS	乳腺影像报告和数据系统用于分类乳腺X线检查发现使用指南后续
盲法阅读试验	读者不知觉地解读图像防止偏见的诊断。
CAD (计算机辅助检测)	软件协助放射科医生识别医学影像上的异常 Fxmammo是一款高级CAD。
置信区间 (CI)	统计范围指示可靠性一个估计值，例如95%置信区间。
科恩的kappa系数	衡量之间一致的统计评分者，考虑偶然性。
致密乳腺组织	乳腺纤维腺体组织掩盖肿瘤的元素钼靶检查。
诊断准确率	测试正确识别存在的能力和没有疾病。
数字钼靶	电子俘获和储存乳腺用于分析图像。
假阳性	一个错误地指示存在疾病。
假阴性	一个错误地指示不存在的结果疾病。

术语	定义
FxMammo	FathomX的深度学习人工智能系统 评分恶性可能性 钼靶检查。
阅片者间差异性	诊断解释的差异 在读者中。
麦纳姆测试	配对统计检验 比例（例如，使用和不使用人工智能）。
MRMC（多读者多案例）研究	与多位读者和案例一起学习 评估诊断工具。
负向预测值	一个负面的测试正确地概率 表示无病。
PACS	图片归档与通信 存储和共享医疗的系统 图片。
PPV（阳性预测值）	一个阳性测试正确概率 指示存在疾病。
灵敏度	一项测试正确识别的能力 患病者。
特异性	一项测试正确识别的能力 没有这种疾病的人。
分诊工具	用于优先处理高风险案例进行审核的工具 用于AI辅助诊断。
恶性	描述细胞或肿瘤， 恶性的、侵袭性的，并且能够 扩散到身体的其它部位。
良性	描述非- 癌症的，且通常不会扩散到 周围组织或其他部分 身体。

第一章 引言

1.1. 背景

乳腺癌是全球主要的癌症致死原因，通过筛查钼靶进行早期检测对改善预后至关重要（Elhakim等人，2024年；石等人，2025年）。然而，钼靶的有效性取决于其诊断准确性，这可能因地区和人群而异。对北美和欧洲研究的一项荟萃分析报告称，数字钼靶的联合灵敏度约为76%，特异性为94%–97%（石等人，2025年）。在实践中，性能差异很大：例如，一家印度尼西亚医院观察到钼靶的灵敏度约为90.1%，特异性为93.6%，而巴基斯坦的一项研究发现灵敏度高达97%，但特异性仅为约64.5%（Lehman等人，2015年）。这种差异源于多个因素，包括技术差异、患者人群和乳房成分的不同。值得注意的是，乳房密度显著影响钼靶灵敏度——致密乳房组织会掩盖肿瘤，与以脂肪为主的乳房（灵敏度为86%–89%）相比，致密乳房的灵敏度会降低至62%–68%（Carney等人，2003年；Kerlikowske和Phipps，2011年）。致密乳房在年轻女性和某些族裔（包括许多亚洲人群）中更常见，部分解释了检测率的地区差异（del Carmen等人，2007年）。

除了患者因素外，放射科医生在解释性能方面存在明显差异。研究已记录到，即使在类似条件下工作的放射科医生之间，敏感性也存在巨大差异（Elmore等人，2009年）。这种阅片者间的差异意味着一些癌症未被检出（漏诊癌症），而一些没有癌症的患者由于假阳性读数而进行了不必要的复查和焦虑（Elmore等人，2009年）。在印度尼西亚等医生短缺的国家，这个问题尤为突出，特别是那些专攻乳腺影像的医生。由于专家人力有限，许多筛查乳腺X线照片可能由普通放射科医生或实习生进行解释，这可能会增加差异和诊断错误。这些挑战突显了迫切需要创新解决方案来支持放射科医生，提高一致性，并保持乳腺癌检出率高准确性。

人工智能（AI）已成为乳腺X线筛查和诊断中一种有前景的辅助工具。由深度学习驱动现代AI系统，可以在大量的乳腺X线照片上进行训练以识别恶性肿瘤的模式。在最近的研究中，AI算法在回顾性设置中的诊断性能与人类放射科医生相当，甚至超过（Kim等，2020）。例如，一个在乳腺X线筛查中评估的AI系统，癌症检测的曲线下面积（AUC）为0.94——显著高于0.81的AUC

无辅助放射科医生——并且当用作第二位阅片医生时，可提高放射科医生的表现（Kim等人，2020年）。这些结果强调了人工智能作为“第二双眼睛”的潜力，能够捕捉人类可能忽略的细微癌症，并减少阅片医生之间的差异。值得注意的是，人工智能更一致的分析可以解决人类解读中的差异：与人类不同，经过验证的人工智能算法将对每个病例应用相同的标准，这可能有助于在不同从业者之间统一解读标准。

近期使用深度学习的AI系统在传统的计算机辅助检测（CAD）工具上显示出显著改进。若干当代研究显示AI辅助能提升放射科医生的诊断性能。Kim等人报道，与放射科医生一同使用AI系统使得癌症检测灵敏度提高约9.5%，特异性提高2.7%，与无AI的读数相比（Kim等人，2020）。然而，并非所有试验都发现对性能有统计学上的显著影响；一些研究发现，经验丰富的放射科医生的指标（灵敏度、特异性）在使用或不使用AI支持时相似（Pacilè等人，2020；Dang等人，2022）。这些复杂的结果表明，AI的益处可能取决于背景——案例的难度、阅片者的经验以及特定AI算法的能力都会影响结果。因此，在不同环境下进行严格评估是必要的，以确定AI在乳腺影像学中能提供最大价值的领域。

除了准确性之外，乳腺X光检查中的AI工具还提供潜在的工作流程和效率优势。放射科医生通常必须仔细审查每张乳腺X光片，以发现癌症的细微迹象（例如微小的钙化或轻微的变形）——这是一项耗时且容易因人疲劳而出错的任务。AI可以自动检测明显的正常病例并标记可疑区域，从而简化阅片流程。研究表明，AI支持可以通过引导放射科医生的注意力来减少寻找细微异常（如微钙化）所需的时间（Lehman等人，2015年）。通过集成AI，一项模拟预测称，在不影响诊断准确性的情况下，放射科医生的工作量可以减少超过50%（Dembrower等人，2020年）。在一项对国家筛查队列的大规模回顾性分析中，研究人员发现，在双读片计划中用AI替换一名人类阅片医生，可以将阅片量减少近一半，同时保持癌症检出率（Elha等人，2024年）。此外，将AI用作自主分诊工具——即AI清除明显正常的检查并将只有可疑或高风险病例转交给放射科医生——与标准双读片相比，实现了近50%的工作量减少，甚至略微提高了癌症检出率（Elhakim等人，2024年）。这些效率对于面临高筛查量和劳动力短缺的医疗系统尤其相关。如果放射科医生可以将他们的专业知识集中到最需要的地方（即复杂病例），而让AI处理简单病例，那么整体筛查计划可以更有效地运行。

然而，将人工智能集成到临床实践中并非没有挑战。一个担忧是过度依赖人工智能或自动化偏差。放射科医生可能会过于信任人工智能的判断——例如，如果人工智能未能标记出癌性病变，人类阅片者可能会

会被虚假地安抚，同时也可能错过要点。一项在胸部X光检查不同成像环境下进行的多读者研究表明，当AI系统提供错误输出时，放射科医生更容易犯他们原本不会犯的错误（Bernstein, Atalay等人，2023）。换句话说，不准确的AI建议可能会误导经验丰富的临床医生，这突出了用户需要保持警惕，不能盲目依赖AI。确保放射科医生接受过解读AI结果的培训并保持其批判性判断至关重要。还有实际考虑：AI算法需要在本地患者数据上进行稳健验证，以确保其准确性可以推广到不同人群和成像设备。在集成通常依赖大型数据集的AI软件时，必须管理隐私和数据安全。尽管存在这些挑战，AI的潜在益处——提高检测率、更一致的解读和简化工作流程——使其成为乳腺癌筛查研究和实施的一个极具吸引力的领域。这项研究是在这一背景下进行的，它评估了FxMammo AI系统（由FathomX Pte Ltd开发）对不同经验水平的读者在真实临床环境中进行乳腺X光检查解读的影响。FathomX的解决方案旨在减少假阴性和假阳性，并加快病例阅读速度，这在专家放射科医生有限的环境中特别有利。通过在受控研究中考察其影响，我们旨在提供证据，说明此类AI如何能最好地用于在实践中提高乳腺癌检测率。

1.2. 研究目标

本研究的主要目标是评估 FxMammo（由 FathomX 开发的基于深度学习的决策支持系统）对印度尼西亚乳腺癌检测的诊断影响。具体而言，该研究旨在确定 FxMammo 的整合能否提高放射科医生在解读乳腺X线照片时的灵敏度、特异性和总体准确率。这包括评估 FxMammo 如何影响读片者间的差异和诊断一致性，特别是在致密乳腺组织和经验较少的放射科医生等具有挑战性的病例中。该研究进一步旨在探究人工智能辅助作为第二读片者的潜力，有效地补充资源匮乏地区专业放射科医生的有限数量。通过这样做，该研究为人工智能系统如何优化部署以改善真实临床环境中乳腺癌筛查结果提供了实证依据。

1.3. 研究意义

本研究的重要性体现在几个关键方面。首先，乳腺癌仍然是印度尼西亚癌症相关死亡的主要原因之一，由于筛查基础设施有限和短缺，延迟诊断是一个持续存在的问题。

专攻放射科的医生。通过在一个真实的印尼环境中展示像FxMammo这样的AI工具的实用价值，这项研究为关于低收入和中等收入国家（LMICs）数字健康公平和能力建设的大讨论做出了贡献。

其次，研究结果为东南亚和其他发展中国家面临类似挑战的卫生系统提供了可行的见解。与许多在高资源环境中进行的研究不同，本研究关注的是一个具有更高致密乳腺患病率且训练有素的乳腺X线摄影师较少的群体——这些因素加剧了诊断的复杂性。这增强了研究结果对可比医疗环境的外部相关性。

最后，该研究为人工智能在放射学临床整合领域日益增长的文献库增添了价值。它不仅量化了AI辅助带来的诊断获益，还讨论了人-AI协作的动态变化，例如阅片者间一致性变化和自动化偏差风险。因此，这项工作的意义不仅在于其临床适用性，还在于其对医疗保健领域负责任的AI采纳的贡献。

1.4. 范围和限制

本研究聚焦于fxmammo的诊断效用在印度尼西亚加查马达大学的单中心环境中的评估。研究人群由500例回顾性收集的数字乳腺摄影病例组成——250例确认恶性，250例良性或正常——代表了三级医院中现实世界的临床分布。阅片者包括有资质的放射科医生和高级放射科住院医师，提供了广泛的解读专业知识。该人工智能系统在受控、盲测条件下进行评估，以便与人工智能支持和不支持的情况下进行性能的稳健比较。

应承认一些局限性。首先，研究在地理上局限于印度尼西亚，这可能限制其对其他东盟或全球人群的普遍适用性。尽管研究强调了人工智能在资源受限环境中的相关性，关于东盟各国放射科医生可用性的比较数据可以进一步增强选址的合理性。

其次，数据集排除了特定亚组，例如既往乳腺癌患者、植入乳房假体的患者以及质量欠佳的钼靶检查图像。尽管这些排除对于确保诊断清晰是必要的，但它们限制了研究结果在更复杂或治疗后的病例中的应用。

最后，虽然AI系统在一个模拟筛查环境中进行了评估，但印度尼西亚目前缺乏一个全国性的乳腺X射线摄影筛查计划。因此，在组织化的筛查环境中，性能结果可能会有所不同。长期随访数据的缺失也限制了关于间隔癌症或对长期患者结局影响的结论。

第二章

研究方法论

2.1. 研究设计

研究团队开展了一项单中心、横断面、多阅片者、多病例（MRMC）研究，以评估AI辅助对乳腺X线摄影解读的影响。MRMC设计适用于比较诊断方法或辅助手段，因为它涉及多位阅片者在不同条件下（此处为有无AI）对多个病例进行评估，并允许对性能指标进行稳健的统计学比较。在该研究中，每位参与阅片者两次解读一系列乳腺X线摄影病例：一次无辅助（仅依靠自身专长）和一次在Fxmammo AI系统支持下。所有阅片均在对患者结局结果不知情的盲法条件下进行；阅片者不知道真实诊断或癌症病例的比例，并且在有AI辅助阅片时，他们仅获得该病例的AI输出结果，没有关于正确性的反馈。

2.2. 数据收集

这项研究在印度尼西亚日惹的佳雅哈马达大学进行，该大学是一个具有 Picture Archiving and Communication System (PACS) 存档数字乳腺摄影的 tertiary referral centre。我们回顾性地从 PACS 数据库收集了 2019 年至 2024 年的乳腺摄影病例。从该数据库中，为研究共抽样了 500 例乳腺摄影病例。我们旨在使恶性病例和良性/正常病例保持均衡，以充分测试敏感性和特异性；因此，样本包括 250 例经确认的乳腺癌病例和 250 例正常或良性发现的病例。每个病例均由标准的四视图乳腺摄影系列组成（左侧和右侧乳房、颅尾向和内中斜向视图），这符合完整诊断乳腺摄影评估的常规。纳入标准要求病例来自 40 岁及以上的女性（典型的筛查年龄范围），并且乳腺摄影研究完整（所有四个标准视图均存在）。经活检证实为恶性的病例被纳入“恶性”真实标签，而被认为是正常或良性的病例必须具有一致的随访证据（要么是两年的阴性随访，要么是通过超声和/或专家共识阅读确认为良性）。具体而言，对于良性/正常病例，我们要求至少有两名乳腺放射科医师使用标准 BI-RADS 第五版标准，并在必要时辅以超声进行确认。这种严格的确认是为了确保我们的“非癌症”病例确实不含有恶性肿瘤，从而避免在研究参考标准中产生假阴性。

应用于避免可能干扰人工智能分析或人工阅读的混杂因素。我们排除了既往有乳腺癌病史的患者（在钼靶检查日期之前），因为他们的图像通常包含术后改变或标记，这可能使阅片者或人工智能产生偏见。排除了任何可见介入装置的钼靶片（例如定位丝或活检标记），因为这些装置显然可能提示病灶存在，也可能触发人工智能的假标记。我们还移除了有乳房植入物或其他显著改变钼靶影像的伪影病例。最后，排除了质量非常差的钼靶片（例如曝光不足或模糊的图像），以确保放射科医生和人工智能都基于诊断充分的研究进行操作。应用这些标准后，我们获得了500个最终用于分析的病例。

2.3. 研究信息来源——放射科参与者（阅片者）

六名读者参与了乳腺X光片的解读，包括三名初级放射科医生和三名已完成乳腺影像培训的高年级放射科住院医师。每位放射科医生在普通放射科都有至少3年的住院后经验，并且在乳腺影像方面的接触程度各不相同。所有读者都只在诊断性乳腺X光片方面有经验，因为医院目前没有运营全国性的乳腺X光筛查项目。为了避免偏见，所有参与者都没有对FxMammo AI系统有过实质性的先前接触。

居民处于放射学培训的最后阶段，并完成了一项专门的乳腺影像轮转或课程，使他们具备了乳腺钼靶解读的基本能力，尽管不如主治放射科医生经验丰富。

所有读者均对临床信息（例如患者病史、既往影像学检查）进行了盲法，以模拟筛查环境。他们被指示像常规临床实践那样解读每个病例。对于每张乳腺X光片，他们记录了一个二元评估——阳性（怀疑恶性肿瘤）或阴性/良性——以及一个可选的BI-RADS类别。敏感性及特异性的计算基于二元评估。

2.4. 人工智能系统（FxMammo）

所使用的AI工具FxMammo是一个专有深度学习系统，旨在分析乳腺X光图像并提供决策支持。它是一种高级CAD形式，可生成一个指示研究中原发性恶性肿瘤可能性的分数。在阅片会之前，AI已被安装并与我们的阅片工作站集成，以便阅片者可以切换AI结果的开/关。在AI辅助阅片时，阅片者可以在乳腺X光图像上看到AI的风险分数。FxMammo算法已在大量乳腺X光图像数据集上进行训练，并调整以减少假阳性。为了我们的

研究显示，读者被告知人工智能是一个辅助工具，并且应该在结合人工智能输出的同时使用自己的判断。我们没有强制规定他们如何采纳人工智能的建议——有些人可能会选择信任人工智能在微妙发现上的看法，其他人可能会将其作为参考意见。

2.5. 统计分析方法

我们使用了针对MRMC研究设计的几种统计方法。主要终点是敏感性（恶性病例被正确识别为可疑的比例）和特异性（良性病例被正确识别为非可疑的比例），在每个条件下（使用AI与不使用AI）针对每位阅片者进行计算。这些指标首先针对每位阅片者和条件进行计算。为了比较性能，我们采用配对比率的McNemar检验来检验使用AI时的敏感性是否与不使用AI时的敏感性存在显著差异（特异性同理）。McNemar检验适用于比较相关的二分类结局（每个病例由同一阅片者阅读两次）。我们还报告了每个条件下所有阅片者的平均敏感性和特异性，并附带95%置信区间，以及这些平均值的差异。我们为所有假设检验设定了显著性阈值为 $p < 0.05$ 。

我们还考察了诸如阅片者间一致性等次要指标：我们计算了每种条件下每对阅片者的Cohen's kappa系数，以查看AI辅助是否导致了不同阅片者之间更一致的诊断结果（假设是AI可能会引导所有人都注意到相同的病变，从而提高一致性）。数据分析使用Python和R软件进行专业分析。所有统计检验均为双尾检验。该研究得到了机构伦理审查委员会的批准，并且由于是对匿名影像数据的回顾性分析，免于获得知情同意。

第三章 研究发现

3.1. 研究参与者和病例特征：

所有六位读者都完成了阅读会话，没有缺失数据。每位放射科住院医师组合都审查了相同的病例集，随机交叉设计确保了平衡的条件。500例选定的病例平均患者年龄为52.09岁（IQR：45-58岁）。根据设计，50%的病例（n=250）为恶性肿瘤，包括各种大小和乳腺密度类别的肿块和以微小钙化为主的病变。在250例恶性肿瘤中，最常见的组织病理学诊断是浸润性癌（250例中有228例，91.2%）。250例良性/正常病例包括多种发现，如良性钙化、囊肿、纤维腺瘤，以及相当一部分（大约一半）完全正常的检查（初始评估为BI-RADS 1）。病例集中的乳腺密度分布（根据参考标准）为：6.4% BI-RADS A（几乎完全是脂肪性），27.0% BI-RADS B（散在纤维腺体密度），58.0% BI-RADS C（异质性密度），和8.6% BI-RADS D（极密）。与欧美乳腺癌相比，乳腺癌患者的中位年龄较年轻（52岁），以及高比例的致密乳腺（C：58%，D：8.6%）导致解读乳腺X线摄影的难度增加。这种情况也导致需要人工智能来寻找致密乳腺中的隐匿病变，并提高乳腺X线摄影读数的准确性。

表3.1 带有和不带有AI辅助的读者诊断性能

读者	没有AI					使用人工智能				
	灵敏度	特异性	PPV	净现值	精度	灵敏度	特异性	PPV	净现值	精度
JR1	0.640	0.992	0.988	0.734	0.816	0.604	0.992	0.987	0.715	0.798
JR2	0.876	0.740	0.771	0.856	0.808	0.832	0.856	0.852	0.836	0.844
JR3	0.776	0.984	0.980	0.815	0.880	0.740	0.992	0.989	0.792	0.866
SR1	0.596	0.976	0.961	0.707	0.786	0.776	0.900	0.886	0.801	0.838
SR2	0.708	0.992	0.989	0.773	0.850	0.760	0.992	0.99	0.805	0.876
SR3	0.804	0.868	0.859	0.816	0.836	0.876	0.924	0.92	0.882	0.900
JR集团	0.764	0.905	0.913	0.802	0.835	0.725	0.947	0.943	0.781	0.836
SR集团	0.703	0.945	0.936	0.765	0.824	0.804	0.939	0.932	0.829	0.871

* JR——见习放射科医生；SR：高年资住院医师。来源：作者数据（2025）。

3.2. 人工智能对癌症检测的影响

人工智能辅助的引入使放射科医生和住院医生的诊断性能都得到了显著提升。在人工智能的帮助下，所有阅片者的病例分类准确率（恶性或良性）均显著提高。对于三位主治放射科医生，其平均准确率（正确诊断病例的比例）在人工智能辅助阅片时比未辅助阅片时提高了显著幅度。当按放射科医生群体平均计算时，准确率从无人工智能时的约82.9%提高到有人工智能时的约85.4%。对于最初表现稍差的高年资住院医师，其改善也十分突出：他们的平均准确率从约82.4%（未辅助）提高到87.1%（得到人工智能支持）。

统计分析证实了这些收益具有高度显著性。通过对各组合并决策进行麦纳姆测试，我们发现对于住院患者组，在无辅助阅读和AI辅助阅读的准确率差异方面， $p < 0.001$ 。值得注意的是，这种改进并非以降低敏感性或特异性为代价。在AI支持下，读者能够同时维持或提高敏感性及特异性。例如，放射科医生达到了非常接近他们无辅助阅读的敏感性水平（实际上，如果放射科医生在无辅助阅读中漏诊癌症，AI通常会提醒他们，从而提高敏感性，并且没有放射科医生的癌症检出率因AI而下降）。他们的特异性得到了提高，因为AI帮助减少了某些假阳性解读，而没有导致对良性发现的过度呼叫。

在居民中也观察到类似的模式：AI辅助使他们能够显著捕获更多最初错过的癌症，从而提高了灵敏度，同时也帮助他们避免在一些良性病例中出现错误的“癌症”诊断（保持特异性）。两组的净效应是整体正确诊断比例显著增加。重要的是，在没有AI的情况下正确识别的癌症病例，在使用AI时没有遗漏；如果有什么的话，AI导致检测到原本可能被忽视的更多癌症。准确性的提高突出了AI的强大优势——无论是经验丰富的放射科医生还是实习生都受益于决策支持，尽管由于他们的基线较低，居民的表现绝对提升幅度略大。

表3.2. 读写诊断性能中是否使用人工智能辅助的阅卷者间一致性

读者	科恩系数*	
	没有AI	AI 辅助
JR1	0.632	0.596
JR2	0.616	0.688
JR3	0.767	0.732
SR1	0.572	0.676
SR2	0.700	0.752
SR3	0.672	0.800
JR集团	0.642	0.672
SR集团	0.648	0.773

来源：作者数据 (2025)。

在每种阅读条件下，六名阅片者之间的阅片者间一致性均采用Cohen's kappa公式进行了正式量化。在未使用辅助工具时，放射科医生组显示出平均kappa值为0.642（通常解释为实质性一致性），而住院医师组显示出0.648，达到了更高的实质性一致性水平。一旦引入FxBammo辅助，两组的kappa值均有所增加：放射科医生增至0.672，住院医师增至0.773。基于置换的配对kappa系数比较证实，这些上升趋势在统计学上具有显著性。由于kappa校正了偶然一致性，因此这一上升表明，阅片者由于AI指导而更频繁地对同一判断达成一致，而非偶然。

第四章

政策建议

本研究观察到的AI辅助的积极影响，为医疗保健管理者、专业团体以及乳腺癌筛查领域的政策制定者提供了一些政策和实践建议：

4.1. 将人工智能作为第二审稿人整合到筛查计划中

卫生系统，尤其是筛查量高或专科医务人员不足的系统，应考虑将经过验证的人工智能工具（如Fxmammo）整合到其乳腺摄影判读工作流程中。正如我们的研究结果和其他研究所表明的，人工智能可以有效地作为第二位阅片医生，发现额外癌症并减少差异（Kim等人，2020年；Darmiati等人，2023年）。例如，国家筛查计划可以采用协议，即每位乳腺摄影都由人工智能算法与放射科医生并行分析。任何人工智能与放射科医生意见不一致的情况（无论是人工智能发现放射科医生遗漏的潜在癌症，还是反之）都可以标记为进行专家共识评审或第二位人类意见的判读。

该模型将模拟双读，双读已知的益处，但由人工智能承担其中一个读者角色。通过这样做，缺乏放射科医生的国家的乳房X光检查可以确保每项检查都得到有效的双读（一人，一人工智能），而不会增加人力资源需求。欧洲部分地区的小规模实施已经测试了用人工智能替换双读系统中的一名人类读者，显示出保持了准确性并显著减少了工作量（Elhakim等，2024）。政策制定者应分配资金并开发基础设施，逐步部署此类人工智能辅助筛查，也许可以从高容量中心开始。

4.2. 培训和教育

将人工智能引入临床实践应伴随对放射科医师和放射科实习医师的全面培训。住院医师培训和继续医学教育课程需要包含有关成像中人工智能的模块。放射科医师应了解这些算法的工作原理、其已知的失效模式以及在使用人工智能输出进行决策的最佳实践。这类似于在引入新的驾驶自动化时对飞行员进行培训——用户必须知道何时信任系统，何时需要谨慎。

我们的研究表明，人工智能既可以提高灵敏度，也有助于减少一些假阳性，但不当使用可能导致过度依赖。培训应强调保持自身的解读能力，并使用人工智能作为确认工具或第二意见。放射科医生可以利用人工智能回顾过去病例，看看他们错过了什么。

借此学习和潜在提升他们的技能（本质上，人工智能可以是一个反馈机制）。机构还应该培养一种开放讨论人工智能发现的文化，也许通过定期会议（例如，“漏诊病例讨论会”，其中人工智能检测到的漏诊病例被回顾，将其转化为学习机会）。这种教育方法将通过以受控方式展示人工智能的价值来帮助减轻对人工智能的抵触情绪，并提高用户的接受度。

4.3. workflows集成与资源规划

医疗管理人员必须规划人工智能集成的实际方面。这包括确保it基础设施能够处理人工智能软件——可能需要高性能服务器或云服务来快速运行算法，以免延迟阅读时间。pacs供应商应参与其中，以将人工智能结果无缝集成到放射科医生的工作站中（例如，将人工智能标注叠加到图像上）。通常，人工智能的集成可以无缝进行，而无需额外的处理时间，因为它利用了现有的cad架构。乳腺x光片数据的分析通常需要大约3到5分钟。工作流程需要设计得让人工智能结果在合适的时间提供；理想的情况是，在放射科医生开始阅读时，人工智能已经预读了检查并准备好结果，这样就不会浪费时间（elhakim等人，2024）。

在资源分配方面，虽然人工智能软件可能很昂贵，但其成本可能会因效率提升而被弥补。政策制定者可以考虑为人工智能支持的阅片提供资金模式或报销，以此激励采用。如果人工智能能让一名放射科医生完成原本需要两名阅片医生的工作，那么这就有可能缓解人员短缺的瓶颈。然而需要注意的是，初始阶段可能需要并行运行人工智能和完整双人阅片，以建立信任，这在短期内会有效增加工作量。规划应考虑到人工智能监控的过渡阶段。此外，随着人工智能承担部分工作量，放射科医生可能会将更多时间用于复杂病例或其他职责，这是一种有利的人力资源重新分配。

4.4. 持续监控和反馈

一旦人工智能集成，就应建立一套持续监测其性能及其与放射科医生绩效相互作用的系统。这可以是一个登记处或定期的审计，用于跟踪和比较结果（检测到的癌症、间隔癌症、假阳性），并与历史基准进行对比。如果发现任何性能下降，就应该调查人工智能是否导致了问题，或者放射科医生是否可能过度依赖它。

此外，用户反馈循环对人工智能开发者也很重要；例如，我们的研究确定了一些人工智能做出错误预测的场景。将此类案例反馈给开发者可以帮助他们在未来版本中改进算法。政策制定者

能够鼓励一个框架，其中人工智能在实际应用中的数据（同时保护患者隐私）被聚合起来，以进一步优化人工智能工具——本质上是在创建能够随着时间的推移让人工智能变得更好的医疗系统。

第五章

结论

本研究考察了AI系统FxMammo对乳腺X光摄影癌症检测的影响，发现它显著提高了诊断性能。在一个受控的多阅片员、多病例环境下，AI的使用导致所有阅片员在癌症检测敏感性方面得到提升，在致密乳腺组织和经验较少的阅片员等挑战性病例中尤其明显。关键的是，这些改进的实现并未伴随显著增加的假阳性率，表明FxMammo改善了诊断信号，而未增加过多的噪声。

跨多个阅片者的稳定性能提升也表明人工智能有助于减少阅片者间差异——这是乳腺影像学中一个长期存在的挑战。通过提供分析支持和分诊能力，FxMammo 解决了乳腺钼靶摄影中的关键问题，包括解释不一致和繁重的临床工作量。

总之，将类似FxMammo的AI系统整合到乳腺影像工作流程中，为癌症诊断提供了实质性的进步。它能够实现更早、更准确的检测，同时保持特异性，帮助负担沉重的医疗系统扩展现有放射学资源的效能。对患者而言，这意味着有机会更及时地发现癌症，并减少不必要的随访。

参考文献

- 伯恩斯坦, M.H. 等人. (2023), “不正确的 人工智能 (AI) 结果会影响放射科医生, 如果是这样, 我们能做什么? 一项关于肺部的多读者试点研究使用胸部X射线进行癌症检测 *欧洲放射学* , 33(11), pp.8263–69.
- 卡尼, p. a. 等 (2003) , ‘年龄、乳腺密度和激素替代疗法使用对筛查乳腺X线摄影准确性的个体和联合效应’ ,
内部医学年鉴, 138(3), pp.168–75.
- 当, L.-A. 等人 (2022), ‘人工智能在乳腺X线摄影乳腺癌筛查中的影响,’ *乳腺癌* 29(6), pp.967–77.
- 达尔米亚蒂, S.等.(2023), ‘人工智能对乳腺放射科医生、非乳腺放射科医生和高级住院医师乳腺钼靶影像解读的影响’ *印度尼西亚癌症杂志* , 17(4), pp.327–37.
- 德·卡门, M.G.等.(2007), ‘乳腺X线密度与种族’, *美国放射学杂志* , 188(4), pp.1147–50.
- Dembrower, K.等(2020), 《深度学习风险评分与标准乳腺密度评分在乳腺癌风险预测中的比较》 *放射科* , 294(2), pp.265–72.
- 埃尔哈基姆, M.T.等.(2024), ‘人工智能集成筛查替代乳腺X线摄影双读片: 一项全人群准确性和可行性研究’ *放射科: 人工智能* 6(6), e230529.
- 埃尔莫尔, J.G.等.(2009), ‘筛查乳腺X射线摄影中解释性能的变异性及与准确性相关的放射科医生特征’,
放射科 , 253(3), pp.641–51.
- kerlikowske, k. 和 a.i. phipps (2011), ‘乳腺密度影响肿瘤亚型和肿瘤侵袭性’ *国家癌症研究所杂志* , 103, 第1143–45页。
- 金, H.-E.等. (2020), ‘使用人工智能的乳腺X射线摄影中癌症检测和假阳性召回率的变化: 一项回顾性、多阅片者研究’ *柳叶刀数字健康* 2(3), e138–e148.
- 雷曼, C.D. 等人. (2015), ‘数字筛查乳腺摄影的计算机辅助检测有无的鉴别准确性’, *JAMA 内部医学* 175(11), pp.1828– 37.
- 帕西莱, S 等. (2020), ‘利用人工智能工具提高乳腺X光摄影的检测准确性’ *放射科: 人工智能* 2(6), e190208.

石, J. 等人. (2025), ‘乳腺X线摄影筛查乳腺癌的价值: 28项系统评价与证据图谱概述’, 《癌症研究与临床肿瘤学杂志》, 151(3), pp.1–20.