

打开

胃和结肠上皮肿瘤的组织病理学分类的深度学习模型

野冈之子^{1, 5}, Fahdi Kanavati^{2, 5}, Kei加藤^{2, 3, 5}, 迈克尔·兰博¹中次郎高二⁴和Masayuki^{1, 2*}

胃和结肠上皮性肿瘤的组织病理学分类是病理学家的常规病理诊断任务之一。基于人工智能（AI）的计算病理学技术将在减轻病理学家不断增加的工作负担方面有很大的好处，特别是在缺乏病理诊断服务的地区。在本研究中，我们对胃和结肠的活检组织病理学全载玻片图像（WSIs）进行了训练，训练了卷积神经网络（cnns）和循环神经网络（Rnns）。训练这些模型将WSI分为腺癌、腺瘤和非肿瘤性。我们在三个独立的测试集上评估了我们的模型，胃腺癌和腺瘤的曲线下面积（auc）分别为0.97和0.99，结肠腺癌和腺瘤的曲线下面积分别为0.96和0.99。研究结果表明，我们的模型具有泛化能力，并且在实际的组织病理学诊断工作系统中具有很高的部署潜力。

根据全球癌症统计数据¹在美国，胃癌和结肠癌是最常见的癌症死亡的主要原因之一，胃癌男性排名第四，女性排名第七，结肠癌男性排名第三，女性排名第二。2018年，胃癌新病例1033701例，胃癌死亡782685例，胃癌新病例1096601例，结肠癌死亡551269例。

在常规的临床病理诊断中，对标本进行组织病理学检查（e.g. 苏木精和伊红（H&E）彩色玻片）通常是在光学显微镜下进行的。全载玻片图像（WSIs）是通过专门的扫描设备获得的玻璃载玻片的数字化对应物，它们被认为可与显微镜下的初步诊断相媲美²。WSIs的出现导致了医学图像分析技术、机器学习的应用，以及最近的深度学习技术，以帮助病理学家检查WSIs和诊断癌症。特别是，深度卷积神经网络（CNNs）已经在大量的计算机视觉中显示出了最先进的结果^{3, 4}以及医学图像分析的应用程序⁵。有前途和成功的计算病理学应用包括肿瘤分类和分割、突变分类和结果预测⁶⁻¹⁷。这些结果突出表明，在部署基于深度学习的工具和工作系统来帮助外科病理学家和支持组织病理学诊断时，特别是在增加初级筛查效率和诊断双读方面，可以获得潜在的巨大利益。

计算病理学的主要挑战之一是WSI的绝对大小。在20倍的放大倍数下获得的一幅图像可以包含几十亿像素，而感兴趣的面积可以小到几千个像素。为了应用深度学习分类器，WSI必须被划分为几千个块，然后将分类器独立地应用于每个块。然后需要聚合所有块的Te输出，以获得final WSI分类。由此产生的另一个挑战是获取带有详细像素级注释的大型wsi数据集的耗时任务，这使得训练具有大量图像的有监督深度学习分类器变得不同。有监督学习的另一种替代方法是弱监督学习，

¹Medmain股份有限公司，日本福冈，810-0042。²梅德曼研究公司，日本福冈，810-0042。³广岛大学医学院，广岛，734-0037，日本。⁴广岛大学医院解剖病理科，日本广岛，734-0037。⁵这些作者的贡献相等：岩冢大、卡纳瓦蒂和加藤圭。*email: tsuneki@medmain.com

其中，只使用全幻灯片的诊断标签，而不是需要像素级的注释。这种技术将需要大量的wsi和计算资源。坎帕内拉等人最近曾尝试过它。¹⁷在44,732个wsi的数据集上获得了令人印象深刻的结果。深度学习以前已经应用于胃和结肠的WSIs，尽管是用于小数据集（11–500）。对于胃癌，Sharma等人。¹⁸使用了11个WSIs的数据集来进行癌的分类。对于结肠癌，深度学习已被用于预测生存结果^{19,20}，核分类²¹和息肉^{10,22}。

在这里，我们提出了深度学习模型来分类胃和结肠的上皮性肿瘤（腺癌和腺瘤），以支持外科病理学家的常规组织病理学诊断。在本研究中，我们收集了两个关于胃和结肠的大数据集，分别包含4128个和4036个WSIs。WSI的收集来自广岛大学医院（日本广岛）和原道医院（日本福冈），并由病理学家手工注释。然后，我们训练了一个基于初始空间-v3架构的CNN²³对于每个器官，使用从WSIs中提取的数百万张瓷砖，将一个瓷砖分为三种标签之一：腺癌、腺瘤和非肿瘤性的。然后，我们汇总了给定WSI中所有块的预测，以使用两种策略获得final分类：简单的最大池方法和递归神经网络（RNN）。

结果

一个用于鉴别WSI组织病理学的深度卷积网络。本研究的目的是开发深度学习模型，可以检测和鉴别胃和结肠活检标本wsi中的上皮性肿瘤（腺癌和腺瘤）。我们的大部分数据集来自广岛大学医院（日本广岛），有4128个胃WSIs和4036个结肠WSIs。我们的数据集的其余部分，包括500个胃WSIs和500个结肠WSIs，来自原道医院（日本福冈）。特广岛大学医院队列在WSI水平上随机分层，获得每个器官的测试集的500个WSIs，其余用于训练和验证（5%）。Te Haradoi医院队列不用于培训，并专门保留作为独立的测试集使用。胃训练集仅包含活检图像，而结肠训练集包含少数手术切除病例（5%）。此外，我们从公开的Te癌症基因组图谱（TCGA）项目的存储库中获得了独立的胃和结肠手术切除病例²⁴要用作测试集。

我们使用初始化-v3网络作为我们的模型的体系结构，并从头开始训练它们。我们使用了它，因为它之前已经被证明在WSI分类任务上工作得很好^{12,25}。Te模型在512×512像素的瓷砖上进行训练，这些瓷砖是以20倍的放大倍数从训练集中随机提取的。每个瓷砖都被分配了三种标签中的一种：腺癌、腺瘤或非肿瘤性的。在训练过程中，以瓷砖旋转和色位变换的形式进一步进行数据增强，以增加鲁棒性，并增加网络的正则化。在推理过程中，训练后的网络以滑动窗口方式使用，输入平铺为512×512像素，f_{xe}步幅小于输入平铺大小。使用滑动窗口是因为所有的瓷砖都需要分类才能获得WSI分类。较小的步幅导致f_{ner}热图和较长的推理时间，而较大的步幅导致较粗的热图和较短的推理时间。我们用了256步。

WSI分类使用两种聚合策略获得：最大池，其中一个WSI从其所有瓷砖中分配最大概率的标签，以及一个RNN模型，该模型被训练使用深度CNN特征作为输入组合所有瓷砖中的信息。图1给出了我们的方法的概述。对于每个聚合策略，我们计算了测试集上的日志损失。日志损失的高值表明模型做出了高可信度的不正确预测。

WSI评价胃和结肠组织病理学图像的高AUC性能。我们对广岛大学医院队列的WSI来源相同)，使用两种WSI分类的聚合方法：最大池（MP-aggr）和RNN（RNN-aggr）。ROC曲线及其对应的曲线下面积（AUC）总结如图2c、d和表1。

Delong检验显示，胃模型（腺癌（ $p = 0.427$ ）和腺瘤（ $p = 0.575$ ））和结肠模型（腺癌（ $p = 0.409$ ）和腺瘤（ $p = 0.945$ ））的auc差异无统计学意义。然而，在对数损失的值在统计上存在显著的差异。表2总结了日志损失及其合并间隔。

对来自不同医疗机构的独立测试集进行评估。其中一个主要的问题是，当模型应用于来自不同来源的测试集时，是否能很好地推广。因此，我们评估了我们的模型在独立的测试集获得的哈拉多伊医院，一个不同的医疗机构。ROC曲线及其对应的曲线下面积（AUC）总结如图所示。2a、b和表3。表4总结了这些模型的日志损失值。

对于胃模型，Delong检验显示腺癌的auc中的差异无统计学意义（ $p = 0.495$ ），而腺瘤的差异无统计学意义（ $p = 0.5e-5$ ）。对于结肠模型，腺癌有统计学意义（ $p = 0.041$ ），而腺瘤无统计学意义（ $p = 0.105$ ）。

模型在包含手术切片的独立tcGA测试集上的性能。为了进一步评估我们的模型的泛化和局限性，我们在从TCGA获得的独立测试集上对它们进行了测试。TCGA只包含手术切片，模型是对活检标本进行训练的。由于手术切除组织比活检要大得多，而且RNN模型是在活检中较小的CNN特征集上训练的，我们只使用MP-aggr方法进行WSI图像分类。

对于胃，我们使用了来自TCGA-STAD项目的475张玻片：360份福尔马林脂肪包埋（FFPE）腺癌诊断样本和115份正常组织的冷冻样本。无正常情况

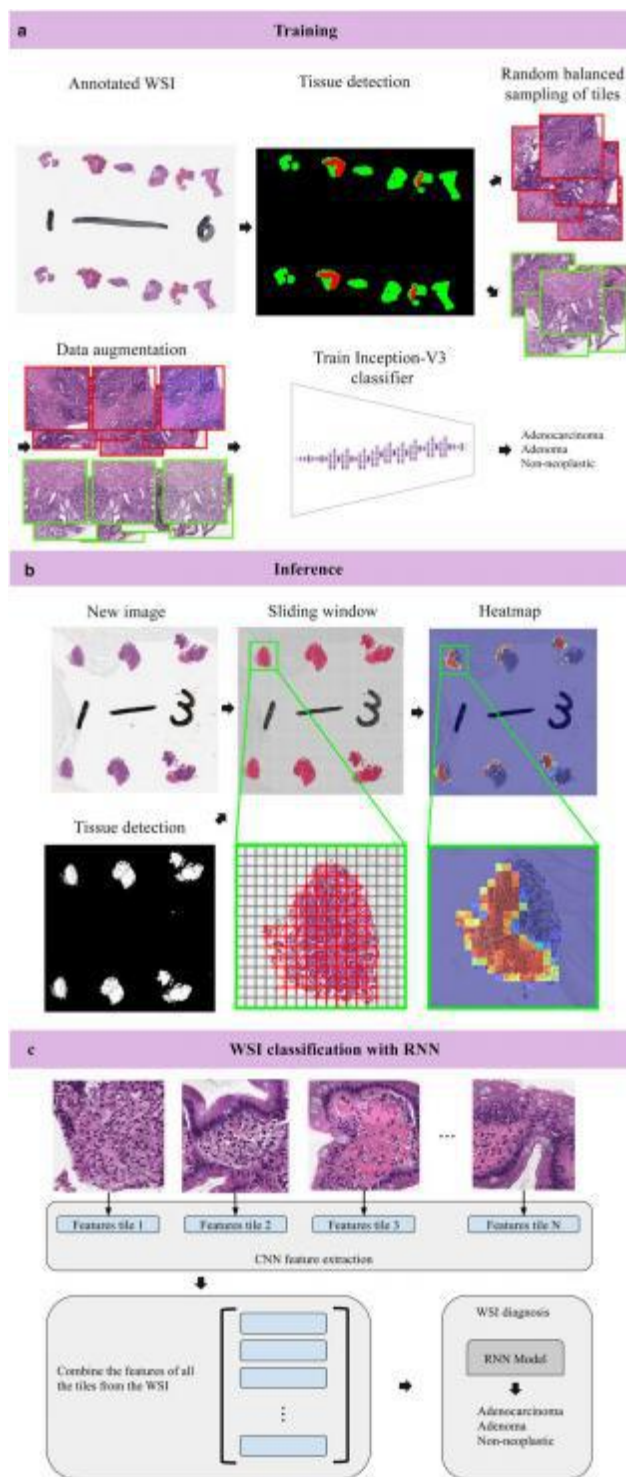


图1. 我们的管道总结。(a) 从训练集中标注的WSIs开始，进行组织检测，然后从每个标签中随机平衡抽样 512×512 块。在训练过程中使用数据增强，使模型对色移和不同方向更加鲁棒。所有的瓷砖都被用来训练一个初始的v3网络，将瓷砖分为3个标签：腺癌、腺瘤和非肿瘤性。(b) 一个新WSI的推理例子。采用阈值法检测Te组织；然后，在检测组织上的滑动窗口提取重叠的 512×512 块，作为网络的输入。这了一个热图图像。使用RNN进行(c) WSI分类，首先需要从训练集中的每个WSI中提取深度CNN特征集来训练RNN模型。然后，通过将训练后的RNN模型应用于WSI的CNN特征集，得到了新的WSI的WSI分类。

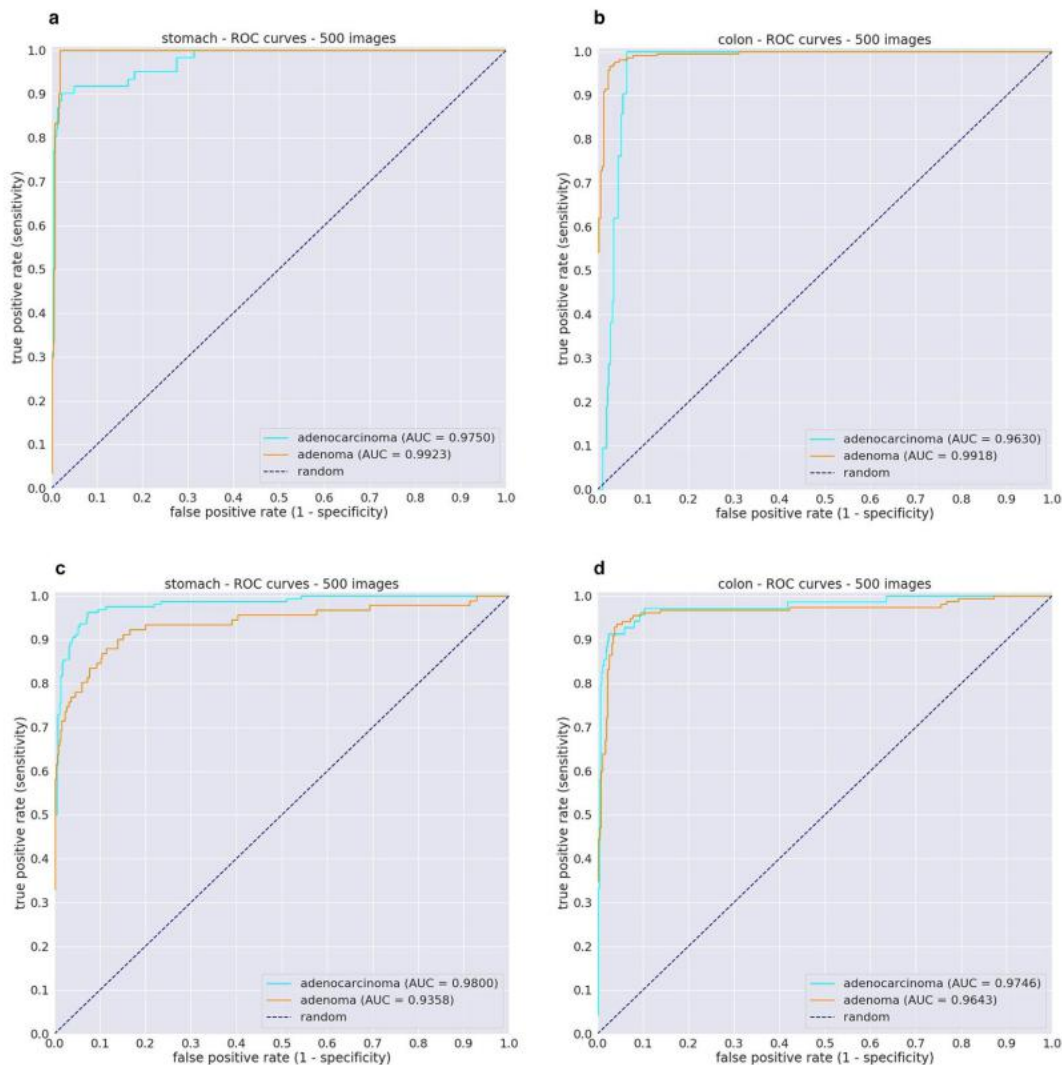


图2。使用RNN-aggr进行WSI分类，评估从不同医疗机构获得的测试案例的模型。（a，b）胃和结肠模型的ROC曲线，分别使用来自Haradai医院队列的500个测试病例。（c，d）胃和结肠的ROC曲线，每个病例使用来自广岛大学医院队列的500个测试病例。

| 标签 | 胃 | | 胃号 | |
|----|-----------------------|-----------------------|-----------------------|-----------------------|
| | RNN-aggr | MP-aggr | RNN-aggr | MP-aggr |
| 腺癌 | 0.980 (0.966 - 0.990) | 0.966 (0.947 - 0.978) | 0.975 (0.948 - 0.993) | 0.972 (0.950 - 0.987) |
| 腺瘤 | 0.936 (0.894 - 0.969) | 0.928 (0.886 - 0.959) | 0.964 (0.940 - 0.984) | 0.961 (0.940 - 0.978) |

表1。在广岛大学医院队列中应用的曲线下面积（AUC）结果，使用两种聚合方法：最大池（MP-aggr）和RNN（RNN-aggr）。括号之间的Te值表示95%的可信区间（CI）s。

组织FFPE样本。腺癌分类的Te AUC为0.924（CI，0.887 - 0.952）（见图. 3a). 对于结肠，我们使用了来自TCGA-COAD项目的547张载玻片： 438个FFPE腺癌样本，以及109个fash冷冻的正常组织样本。腺癌分类的Te AUC为0.982（CI，0.968 - 0.991）（见图. 3b).

胃模型与一组病理学家的比较。我们对一组病理学家和医学生志愿者评估了我们的胃模型，他们没有参与注释训练集，使用45张图像，包括15 WSI腺癌，15 WSI腺瘤和15个非肿瘤病变。在临床工作中使用深度学习模型的最终目的是它们的运行速度；使用训练过的模型，WSI的分类时间从5到

| 标签 | 胃 | | 胃号 | |
|----|-----------------------|----------------------|---------------------------|----------------------|
| | RNN-aggr | MP-aggr | RNN-aggr | MP-aggr |
| 腺癌 | 0.259 (0.175 - 0.368) | 1.264 (1.096 - 1.48) | 0.171 (CI, 0.113 - 0.243) | 0.830 (0.687 - 1.02) |
| 腺瘤 | 0.275 (0.179 - 0.355) | 1.654 (1.348 - 1.82) | 0.241 (CI, 0.155 - 0.344) | 0.895 (0.714 - 1.08) |

表2。在广岛大学医院队列中，使用两种聚合方法：最大池（MP-aggr）和RNN（RNN-aggr）。括号之间的Te值表示95%的可信区间（CI）s。

| 标签 | 胃 | | 胃号 | |
|----|-----------------------|-----------------------|-----------------------|-----------------------|
| | RNN-aggr | MP-aggr | RNN-aggr | MP-aggr |
| 腺癌 | 0.974 (0.952 - 0.994) | 0.980 (0.969 - 0.993) | 0.963 (0.945 - 0.977) | 0.967 (0.943 - 0.984) |
| 腺瘤 | 0.993 (0.986 - 0.998) | 0.987 (0.978 - 0.995) | 0.992 (0.986 - 0.997) | 0.99 (0.979 - 0.996) |

表3。胃和结肠模型的曲线下面积（AUC）结果应用于哈拉多医院的独立测试集，使用两种聚合方法：最大池（MP-aggr）和RNN（RNN-aggr）。括号之间的Te值表示95%的可信区间（CI）s。

| 标签 | 胃 | | 胃号 | |
|----|-----------------------|----------------------|-----------------------|-----------------------|
| | RNN-aggr | MP-aggr | RNN-aggr | MP-aggr |
| 腺癌 | 0.260 (0.173 - 0.338) | 1.436 (1.256 - 1.59) | 0.533 (0.390 - 0.688) | 1.445 (1.238 - 1.62) |
| 腺瘤 | 0.104 (0.053 - 0.153) | 1.339 (1.085 - 1.65) | 0.172 (0.102 - 0.242) | 0.464 (0.327 - 0.557) |

表4。胃和结肠模型的独立测试集上的日志损失，使用两种WSI分类聚合方法：最大池（MP-aggr）和RNN（RNN-aggr）。括号之间的Te值表示95%的可信区间（CI）s。

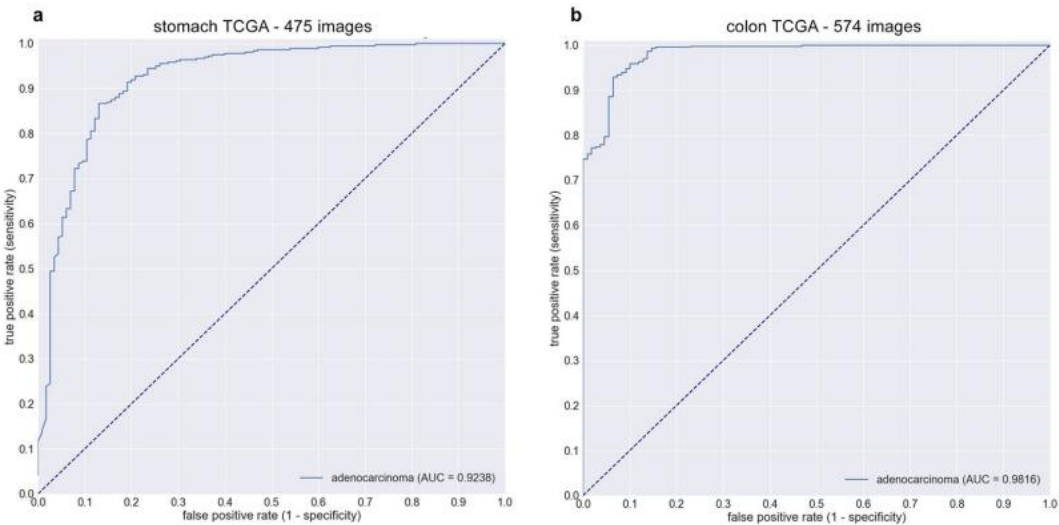
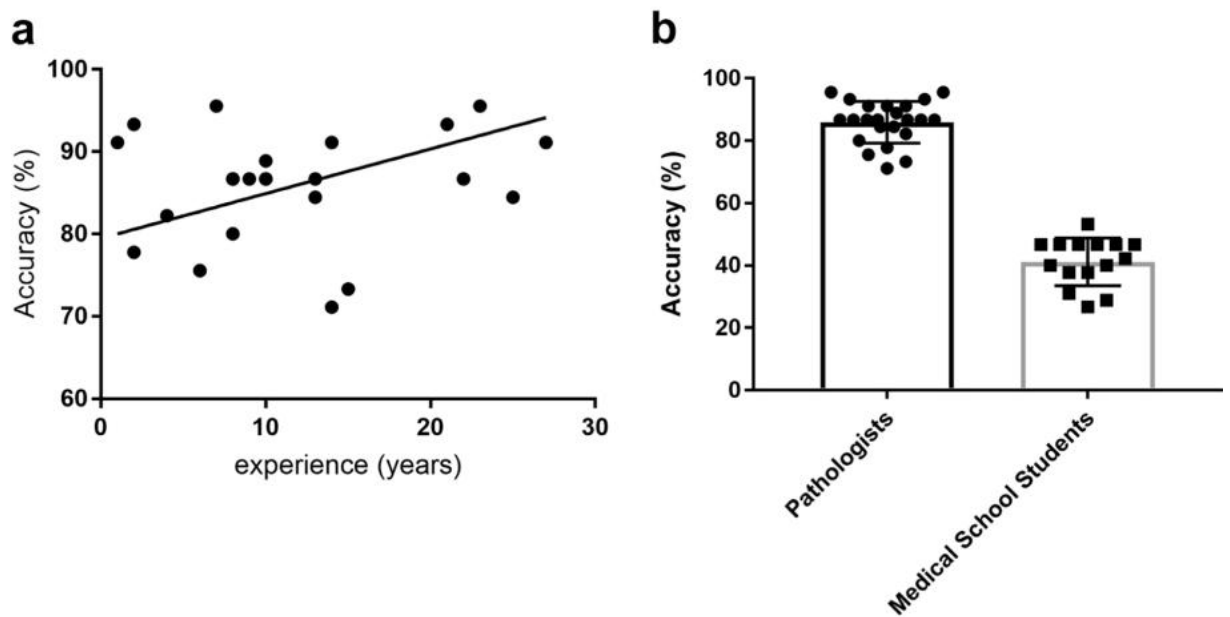


图3。使用MP-aggr对TCGA进行WSI分类的胃(a)和结肠(b)模型的腺癌ROC曲线。即使这些模型在活检切片上进行了训练，它们仍然在手术切片上显示了普遍性。

30秒，具体时间取决于WSI的大小。因此，我们在时间限制下进行了比较。每个病理学家都随机地得到了45张图像，只有30秒的时间来决定诊断。所有病理学家均未使用常规显微镜观察到45例WSIs。病理学家获得的平均准确率（正确诊断的百分比）为85.89% ± 1.401%，（n = 23），而医学院学生的平均准确率为41.19% ± 1.974%，（n = 15）。随机诊断的准确率为33.33%。当使用标准概率阈值为0.5时，Te胃模型的准确率为95.6%。图4比较了病理学家和医科学生的表现。根据多年的经验，病理学家之间未发现具有统计学意义（p = 0.4546）。图5a和图6a比较了……的性能



a: The correlation coefficient: $r = 0.164$, P (two-tailed) = 0.4546
b: t-test: $P < 0.0001$

图4。在45例测试病例下（15例腺癌，15例15例腺瘤和15例非肿瘤）的诊断准确性（%）。(a) Te准确性作为病理学家多年的经验。没有统计学意义 ($p = 0$ 。根据多年的经验在病理学家之间发现。(b) 箱形图比较了病理学家和医学院学生的诊断准确性。病理学家与医科学生之间具有较高的统计学意义 ($p < 0.0001$)。

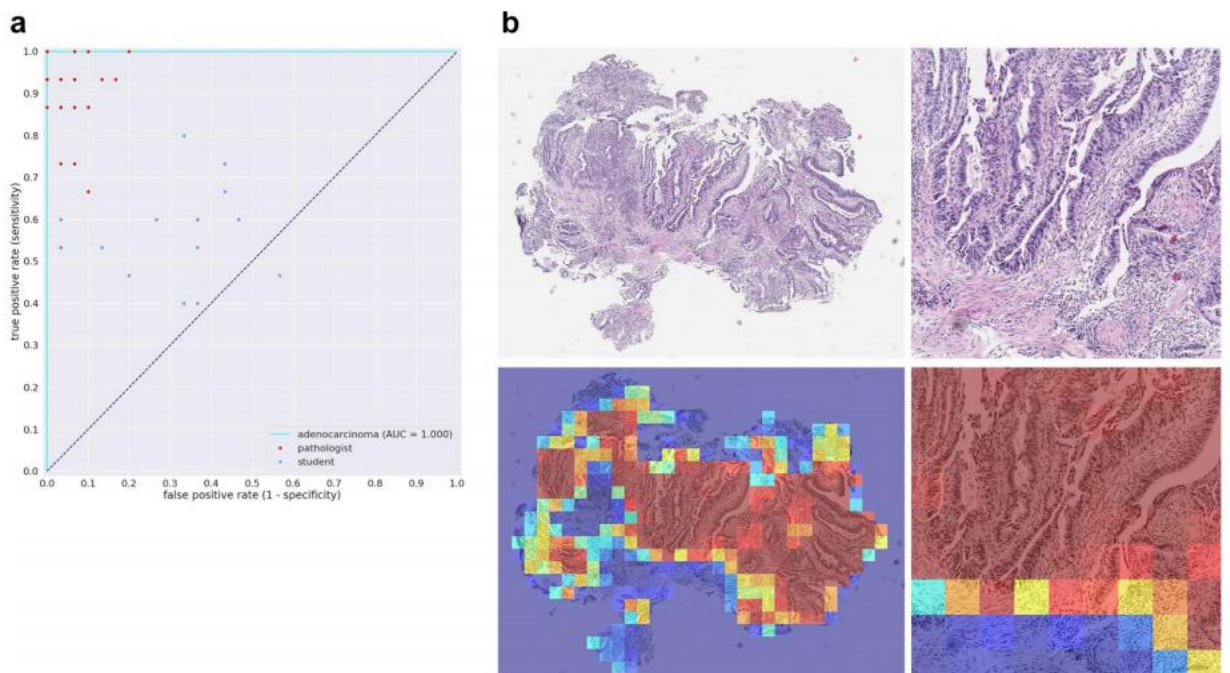


图5。(a) 45例试验病例与一组用于病理学家和医学生比较的胃腺癌ROC曲线。(b) 腺癌胃模型的热图输出示例，使用 128×128 。

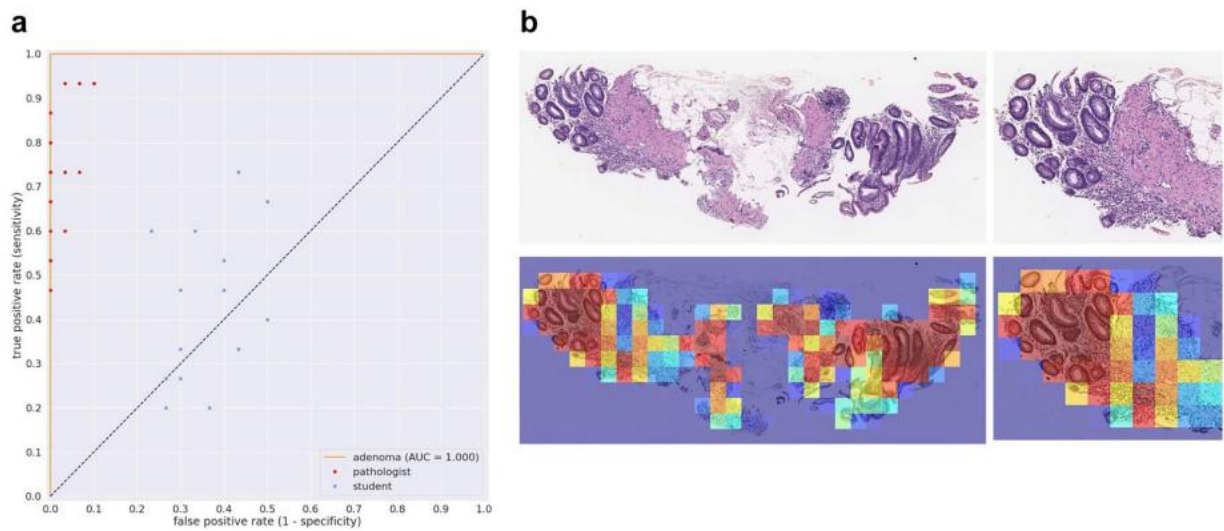


图6。(a)对45例试验病例，用于与一组病理学家和医科学生进行比较的胃腺瘤ROC曲线。(b)腺瘤胃模型的热图输出例子，步幅为128×128。

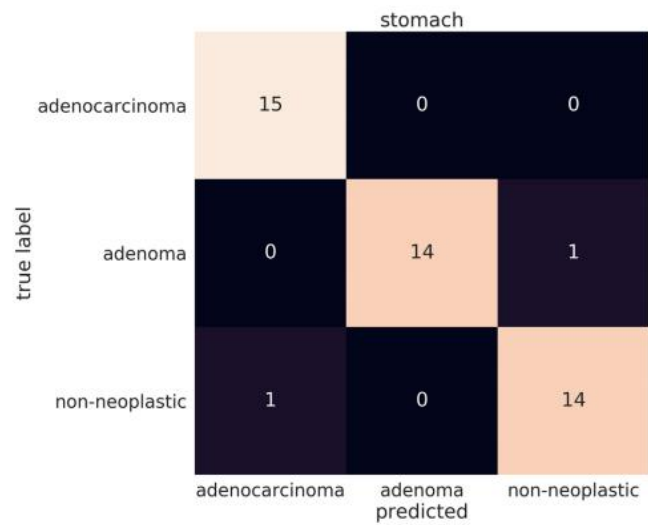


图7。在45个试验病例中，将WSI图像分类为腺癌、腺瘤和非肿瘤的胃模型的混淆矩阵，以评估病理学家在时间限制下的表现。

该模型针对23名病理学家和15名医学院学生。图5b和6b显示了热图输出示例。图7显示了胃模型的混淆矩阵。

讨论

在这项工作中，我们训练了深度学习模型来分类胃和结肠活检WSIs中的上皮性肿瘤。每个模型都在来自单一医疗机构的WSI上进行训练，并分别在来自同一医疗机构和来自不同机构的两个独立测试集上进行评估。在一个给定的WSI上，模型的输出是一组热图，允许可视化预测标签的位置和WSI的主要分类。我们认为，这项研究的结果是重要的，因为，据我们所知，之前没有报道的研究被大数据集验证，特别是对胃和结肠上皮病变。

在不同的测试集上，两种聚合方法之间大多没有统计学差异，当有时，RNN-aggr在统计学上更好。然而，在对数损失方面有统计学意义的差异，RNN-aggr的对g损失要低得多。这意味着使用MP-aggr的WSI概率要高得多，需要较高的阈值切割（约0.99）才能获得分类，而使用RNN-aggr的阈值切割更接近标准的0.5概率。这是

| | 胃 | | | | 胃号 | | | |
|-------|------|------|-----|--------------|------|------|-----|--------------|
| | 广岛队列 | | | 哈拉多伊 步兵大队 | 广岛队列 | | | 哈拉多伊 步兵大队 |
| | 所有 | 列车 | 试验 | 试验 | 所有 | 列车 | 试验 | 试验 |
| 腺癌 | 1386 | 1228 | 158 | 61 | 481 | 412 | 69 | 21 |
| 腺瘤 | 563 | 472 | 91 | 30 | 370 | 215 | 155 | 210 |
| 非肿瘤性的 | 2179 | 1928 | 251 | 409 | 3185 | 2909 | 276 | 269 |
| 合计 | 4128 | 3628 | 500 | 500 | 4036 | 3536 | 500 | 500 |

表5。WSI在数据集中的分布。

这在某种程度上是可以预料到的，因为MP-aggr更容易出现错误，因为一个WSI可以包含多达数千个贴图，而只需要一个假阳性贴图就会导致错误的分类。坎帕内拉等人也注意到了这一点。¹⁷ 这是一个特别的问题，因为这些注释并不是在详细的细胞水平上完成的，一些肿瘤注释的区域包含炎症细胞膨胀的区域。另一方面，RNN-aggr在输出分类之前考虑了幻灯片中的所有瓷砖，并可以抑制假阳性。

所有培训用的幻灯片均全部来自广岛大学医院。这是一个潜在的限制，因为单一机构队列可能不能代表在不同机构的临床环境中可能遇到的样本的完整多样性和异质性。在Haradai医院的独立测试集上的性能没有下降，这是很有希望的。此外，使用FFPE和flash冷冻组织样本对TCGA的独立测试集进行评估显示，在预测手术切除病例中预测腺癌的结果，尽管模型主要是在活检上进行训练的。Te结肠模型在TCGA上的表现优于胃模型，这很可能是由于结肠训练集中包含了一些手术切除病例。我们无法验证腺瘤的预测，因为TCGA没有。

在临床实践中，本研究中对病理学家在时间限制下执行的条件是不现实的，因为给病理学家足够的时间和详尽的资源，他们将达到100%的准确性。然而，这确实显示了将这—个系统集成到临床工作工具中的主要优势：它的运行速度。它可以作为综合工作的一部分，一旦扫描载玻片，就会自动预测诊断，并根据优先顺序对病例进行排序，由病理学家复查。这将允许病理学家检查可能需要最多关注的病例，允许更快的周转。Te工具可以应用于WSI的大型数据库，根据给定的预测诊断组织和检索幻灯片，也可以作为第二个读者，确认诊断或在初步诊断有分歧时提醒病理学家，要求进一步检查。将人工智能集成到计算病理学工作流程中，将对减轻病理学家不断增加的工作量有很大的好处，特别是在获得病理诊断服务短缺的地区。

在下一步中，我们希望进一步对癌的亚型进行分类(e.g., 乳头状腺癌、管状腺癌、粘液腺癌、印戒细胞癌、腺鳞癌等)，在外科病理学家进一步开发实用辅助，并训练模型预测患者预后方面发挥重要作用^{26, 27}。我们希望见证未来病理学家和人工智能合作支持医疗保健基础的发展。

方法

临床病例本回顾性研究收集了4128例人胃上皮病变和4036例结肠上皮病变HE（苏木精&伊红）染色的组织病理学标本，并由外科病理学医生对这些标本进行了组织病理学检查。该实验方案得到了广岛大学伦理委员会的批准。E1316). 500例人胃上皮医院（福冈）和500例结肠上皮病变HE染色手术标本来自人胃上皮病变，经组织病理学检查和批准。所有研究活动均按照广岛大学医院和原道医院的相关准则/规定进行。在两家医院进行手术前，所有患者都获得了使用组织样本和病理诊断报告的知情同意。Te试验病例是随机选择的，因此所获得的比率尽可能地重现了一个真实的临床场景。

数据集和注释。胃在解剖学上分为胃区：贲门、眼底、体、胃窦和幽门。这些解剖区域表现为三种类型的胃粘膜，在边界区域呈过渡性形式：心脏、胃底粘膜和幽门（胃窦）。这三种胃腺由两种成分组成：小窝和分泌部分（腺瘤）。中央凹是胃恶性上皮肿瘤（如腺癌）发病机制的重要组成部分，因为干细胞层位于基部。上皮良性肿瘤。g., 和非肿瘤性病变(e.g., 胃炎和息肉)也发生在胃粘膜。大肠（大肠）分为六个区域：盲肠、升结肠、横结肠、降结肠、乙状结肠和直肠。结肠黏膜由三种成分组成：上皮层、固有层和粘膜肌层。结肠上皮肿瘤（恶性腺瘤；良性腺瘤）和非肿瘤性病变(e.g., 结肠炎和息肉)发生于结肠黏膜，其上皮细胞成分起源于结肠上皮细胞。对于从广岛大学医院获得的数据集，胃

数据集由4128个WSIs组成，冒号数据集由4036个WSIs组成。表5将数据集的分布分为训练集和测试集。Te胃数据集仅由活检图像组成，而结肠数据集包含部分手术标本（5%）。所有用于注释的病例在注释前都被仔细检查，并由病理学家进行验证。

Te WSIs由一组外科病理学家（注释病理学家）手工注释，他们通过绘制与腺癌、腺瘤和非肿瘤之一对应的区域来进行常规组织病理学诊断。腺癌注释包括肿瘤实质和间质（肿瘤内有炎症细胞膨胀或肿胀的组织部分）。任何不包含腺癌或腺瘤的区域都被包括在非肿瘤标签下，其中包括炎性（e. g., 胃炎、结肠炎和增生性增生。g., 基底腺息肉，结直肠息肉）病变及正常。非肿瘤性注释包括上皮细胞和间质。每个WSI的平均注释时间约为10分钟。由注释病理学家执行的注释被修改（如有必要），并由另一组病理学家进行合并和验证，并用作训练数据集。每个注释的WSI至少由两名病理学家看到，并由一名高级病理学家进行WSI检查和验证。在注释标签上有差异的病例被排除在训练之外。一些WSI包含多个注释标签。因此，一个给定的WSI：腺癌、腺瘤、非肿瘤。例如，如果WSI同时包含腺癌和腺瘤的注释，则WSI诊断为腺癌。

图像预处理。我们的变焦倍数为20倍。对于每个WSI，我们对first进行组织检测，以去除大面积的白色背景。Te组织检测是基于一个简单的饱和阈值法。基于获得的组织掩模，我们提取了 512×512 块不同的随机取向，因为组织样本没有特定的取向。我们从给定WSI的每个标签中采样相同数量的贴图（ $n = 400$ ），以避免WSI水平上的标签不平衡（当一个给定标签的图像比另一个标签有更多的图像时）。例如，如果一个WSI同时患有腺癌和非肿瘤，那么我们取样400块腺癌和400块非肿瘤。在采样过程中，如果超过50%的注释标签位于瓷砖内的 128×128 中心窗口之外，则标题将被拒绝；否则，瓷砖被分配给它们重叠的标签。从每个数据集中总共提取了大约400万个贴图。

深度学习模型。我们使用了标准的初始空间-v3网络架构²³输入大小为 512×512 。我们通过使用0.35的深度乘数来减少参数的数量，从而减少了权重参数的简化版本；深度乘数是一个标量因子，用于同时缩放所有卷积层的深度。这允许加快训练和推理，而对预测精度的影响最小。Liu等人观察到使用低至0.1的深度乘数的精简版本，以达到与网络完整版本相似的性能。²⁵乳腺癌WSI。我们没有对网络进行任何进一步的调整。我们从零开始训练这个精简后的网络。

为了使算法对图像变化的鲁棒性，并添加了正则化方法，我们应用了几种数据增强技术。这包括一些技术，如随机旋转随机固定图像或上固定；小比例；次要的位置；色调和饱和度；以及对对比度和亮度扰动。然后将Te图像缩放到 $[-1.0, 1.0]$ 。

虽然我们从给定的WSI中平等地采样每个标签中的贴图，但数据集仍然包含每个标签的WSIs数量不相等，这导致每个标签提取的贴图的总体数量不相等。为了解释这种不平衡，我们使用了一个加权交叉熵损失函数。Te权重被用来减少WSI水平上的标签不平衡，因此网络不利于预测一个标签而不是另一个标签。我们使用Adam优化算法来训练网络²⁸动量为0.9，衰减为0.9，衰减为1.0。我们使用了一个有温暖开始的衰减学习率，最小开始学习率为0.001，最大学习率为0.05。我们使用了128次的批处理大小，并且训练总共运行了625 K次迭代。选择验证误差最低的Te模型作为final模型。

全幻灯片分类的RNN模型。我们之前训练过的初始-v3网络可以通过去除final全连接的分层来作为特征提取器。初始值-v3特征提取器的Te输出，深度乘数为0.35，是一个715特征向量。使用 256×256 步幅像素的滑动窗口从WSI中提取瓷砖。在推理过程中，需要检查来自给定WSI的所有贴图，以便进行WSI分类。

一个RNN模型可以采用任意长度的序列并产生单个输出²⁹。每张幻灯片都从组织区域中提取了任意数量的瓷砖，我们使用所有的瓷砖作为RNN模型的输入。在训练过程中，每一步都随机关闭瓷砖特征的顺序，以减少对瓷砖输入顺序的依赖。我们使用了一个由两个长短期记忆（LSTM）组成的RNN³⁰具有隐藏状态表示大小为128的层。我们使用批量大小为1的随机梯度下降来训练模型。Te模型的训练速率为0.001，50个周期的衰减为 $1e-6$ 。选取在5%验证子集上表现最好的模型作为final模型。

软件和统计分析。利用张量流法实现和训练了Te模型³¹。使用scikit-learn软件包在python中计算auc³²使用matplotlib绘制³³。采用自举法估计了auc的Te95%ci³⁴1000次迭代。对auc的比较采用双尾DeLong检验³⁵对于两条相关的ROC曲线。对对数损失进行比较

双面学生t检验。对病理学家的多年经验与准确性进行了相关性分析。采用学生t检验来检验病理学家和医学院学生的诊断准确性的差异。图形棱镜7.04版。graphpad.com)用于生成图。4.

数据可用性

由于管理隐私保护的机构要求，本研究中使用的大多数数据集都不能公开获得。本研究的支持数据在提交时提供给编辑委员会成员和审稿人，以评估手稿。Te的外部TCGA数据集来自TCGA-STAD和TCGA-COAD项目，并可通过基因组项目公开获得数据共享门户(<https://portal.gdc.癌症gov/>)。

收到日期：2019年10月22日；接受日期：2020年1月14日；

Published online: 30 January 2020

参考文献

1. 布雷, F. 以及其他他人2018年全球癌症统计: 全球癌症统计机构对全球185个国家36种癌症的发病率和死亡率的估计。CA: 癌症J. 临床医生68, 394–424 (2018)。
2. 穆克霍帕德, S. 以及其他他人全载玻片成像与显微镜初步诊断: 1992例病例的多中心盲法随机非劣效性研究(关键研究)。是J. 外科应用程序。42, 39 (2018)。
3. 他, K., 张, x., 任, S. & Sun, J. 深入研究整流器: 在图像集分类上超越了人类水平的性能。《IEEE计算机视觉国际会议论文集》, 1026–1034 (2015)。
4. 埃斯特瓦, 以及其他他人深度神经网络的皮肤癌分类。Nat. 542, 115 (2017)。
5. 利琴斯, G. 以及其他他人对医学图像分析中的深度学习的调查。医学图像J. 42, 60–88 (2017)。
6. Hou, L. 以及其他他人基于补丁的卷积神经网络用于全幻灯片组织图像的分类。在IEEE计算机视觉和模式识别会议论文集上, 2424–2433 (2016)。
7. 马达布希, A. 和李, G. 《数字病理学中的图像分析和机器学习: 挑战与机遇》(2016年)。
8. 利琴斯, G. 以及其他他人深度学习作为一种提高组织病理学诊断的准确性和准确性的工具。科学。棱镜平布6, 26286 (2016)。
9. 克罗斯, Z., Ba, J. L. & Frey, B. J. 利用深度多实例学习对显微镜图像进行分类和分割。生物信息学。32, i52–i59 (2016)。
10. Korb, B. 以及其他他人全幻灯片图像上进行结直肠癌肉分类的深度学习。病理学信息学杂志8 (2017)。
11. 罗, X. 以及其他他人综合的计算病理图像分析可预测肺癌的预后。J. 托拉克。OncoI. 12, 501–509 (2017)。
12. 库德雷, N. 以及其他他人利用深度学习对非小细胞肺癌组织病理学图像进行分类和突变预测。Nat. 医学24, 1559 (2018)。
13. 魏, J. W. 以及其他他人用深度神经网络对切除的肺癌切片上的组织学模式进行病理学家水平的分类。科学。棱镜平布9, 3358 (2019)。
14. 格蒂奇, A. 以及其他他人卷积神经网络可以准确地通过数字切片区分肺腺癌的四种组织学生长模式。科学。棱镜平布9, 1483 (2019)。
15. Bejnordi, B. E. 以及其他他人深度学习算法检测乳腺癌女性淋巴结转移的诊断评估。Jama318, 2199–2210 (2017)。
16. 萨尔茨, J. 以及其他他人利用病理图像的深度学习研究肿瘤膨胀淋巴细胞的空间组织和分子相关性。细胞代表。23, 181–193 (2018)。
17. 坎帕内拉, G. 以及其他他人临床级计算病理学使用弱监督深度学习的整个幻灯片图像。Nat. 医学25, 1301–1309 (2019)。
18. 沙玛, H., 泽贝, N., 克伦伯特, I., 赫尔威奇, O. & Hufnagl, P. 深度卷积神经网络在数字组织病理学中使用全幻灯片图像自动分类胃癌。计算机化的医学。成像图。61, 2–13 (2017)。
19. Bychkov, D. 以及其他他人基于深度学习的组织分析可以预测结肠直肠癌的预后。科学。棱镜平布8, 3395 (2018)。
20. Kather, J. N. 以及其他他人利用深度学习预测结肠直肠癌组织学切片的生存率: 一项回顾性多中心研究。PLoS医学。16, e1002730 (2019)。
21. Sirinukunwattana, K. 以及其他他人局部敏感深度学习用于常规结肠癌组织学图像中细胞核的检测和分类。IEEE跨。医学成像35, 1196 (2016)。
22. Korb, B. 以及其他他人观察: 深度神经网络可视化来解释结直肠癌肉的全幻灯片图像分析结果。在IEEE计算机视觉和模式识别(CVPR)研讨会(2017)。
23. 谢格迪, 范霍克, 美国, 施伦斯, J. 和WojnaZ. 重新思考计算机视觉的初始架构。《IEEE计算机视觉和模式识别会议论文集》, 2818–2826 (2016)。
24. 格罗斯曼, R. L. 以及其他他人走向对癌症基因组数据的共同愿景。N. 引擎。J. 医学375, 1109–1112 (2016)。
25. 刘, Y. 以及其他他人十亿像素的病理图像上检测癌症转移。arXiv预印本arXiv: 1703.02442 (2017)。
26. 胡, B. 以及其他他人胃癌: 分子病理学的分类、组织学和应用。J. 胃肠炎。OncoI. 3, 251 (2012)。
27. 黄, S. S. 以及其他他人全基因组测序揭示了胃癌的基因组景观和遗传异质性。Nat. 通勤。5, 5477 (2014)。
28. 金玛, D. P. & Ba, J. 一种随机优化的方法。arXiv预印本arXiv: 1412 (2014)。
29. 坟墓, A. 监督序列标签。在递归神经网络的监督序列标记中, 5–13 (施普林格, 2012)。
30. Hochreiter, S. & Schmidhuber, J. 长期记忆。神经计算9, 1735–1780 (1997)。
31. 阿巴迪, M. 以及其他他人张量流: 在异构系统上的大规模机器学习(2015)。软件可从小沙发上购买。org。
32. 佩德雷戈萨, F. 以及其他他人Python中的机器学习。J. 马赫数。学习。物品12, 2825–2830 (2011)。
33. 亨特, J. D. 一个2d图形环境。压缩。科学。雕刻9, 90–95, <https://doi.org/10.2007.551109/MCSE> (2007)。
34. Efron, B. 和蒂布希拉尼, R. J. 对引导程序的介绍(CRC出版社, 1994年)。
35. 德隆, E. R., DeLong, D. M. & 克拉克-皮尔森, D. L. 比较在两个或两个以上相关的受试者工作特征曲线下的面积: 一种非参数方法。生物识别学44, 837–845 (1988)。

致谢

我们非常感谢博士们提供的支持。黑月幸、大田真、森圭一和。广岛大学的Tamai; 先生。村上龙, 西小助, 吉田玉木, 岩川永助, 女士。新海诚, 以及梅德曼公司的成像中心和广岛中心。我们感谢来自世界各地参与本研究注释工作的病理学家和广岛大学医学院的志愿学生。利用计算资源进行了Te计算

属于九州大学信息技术研究所大力推广项目的类别。

作者贡献

O. I. ., F. K. .和K. K. .对这项工作的贡献相同; O. I. ., F. K. .和M. T. 设计实验; O. I. ., F. K. ., K. K. ., M. R. ., K. A. 和M. T. 进行了实验, 并对数据进行了分析。I. ., K. K. 和K. A. 收集的组织病理学病例; O. I. ., F. K. ., K. K. 和M. T. 写了手稿; M. T. 监督项目。所有的作者都对稿件进行了审核。

利益冲突

这些作者声明没有相互竞争的利益。

其他信息

有关材料的信件和要求应寄给M. T.

重印和权限信息可在www上找到。自然com/reprints.

出版商的说明《施普林格自然》对已出版的地图和机构纠纷中的管辖权主张保持中立。



本文根据知识共享署名4.0国际进行授权

允许在任何媒介中使用、共享、改编、分发和复制的许可证

使用共享许可证, 并指示是否进行了更改。Te图像或其他第三方材料在此

文章包含在文章的知识共享许可中, 除非在信用额度中另有说明

格式, 只要你给予了适当的信用给原始作者(s)和来源, 提供一个链接到Cre-

经法定规定或超过允许使用, 您将需要获得直接许可

著作权持有人要查看此许可证的副本, 请访问<http://creativecommons.org/licenses/by/4.0/>. ©Te

作者(s) 2020