

文章



<https://doi.org/10.1038/s41467-020-15671-5>

打开

病理图像的计算分析 能够更好地诊断TFE3 Xp11.2 易位性肾细胞癌

郑俊¹、志汉^{2,3}罗希特·梅拉⁴、魏绍²、迈克尔程²前进冯⁵、董妮¹✉, 黄昆^{2,3}✉梁成⁶✉张杰⁷✉

TFE3 Xp11.2易位性肾细胞癌（TFE3-RCC）通常进展更多
与其他RCC亚型相比具有积极的，但TFE3-RCC的诊断具有挑战性
通过传统的对病理图像的视觉检查。在这项研究中，我们收集了苏木精
74例TFE3-RCC的组织病理学全切片图像（最大
74例透明细胞肾细胞癌（ccRCC，最常见的肾CC亚型）
与性别和肿瘤分级相匹配。实现了一种自动计算管道
提取图像特征。比较研究鉴定了52个具有显著差异的图像特征
TFE3-RCC与ccRCC之间的关系。建立了机器学习模型来区分
来自ccRCC的TFE3-RCC。对外部验证集上的分类模型的测试显示
精度高，ROC曲线下面积为0.842~0.894。我们的结果表明
自动推导出的图像特征可以捕捉到细微的形态差异
并为TFE3-RCC提供了一个潜在的指导方针
诊断

¹国家区域医学超声关键技术工程实验室，广东省生物医学测量和超声成像重点实验室，生物医学工程学院，健康科学中心，深圳。²印第安纳大学医学院医学系，印第安纳波利斯，美国。³美国印第安纳波利斯的再生器研究所。⁴密歇根大学病理科，安娜堡，密歇根州，美国。⁵南方医科大学生物医学工程学院，广州，中国。⁶印第安纳大学医学院病理和检验医学系，美国印第安纳波利斯。⁷印第安纳大学医学和分子遗传学系，美国印第安纳波利斯分校。✉电子邮件：nitong@szu.edu.cn; kunhuang@iu.edu; lcheng@iupui.edu; jizhan@iu.edu

R 肾上腺细胞癌（RCC）由多种异质性亚型组成^{1,2}和典型地分为三种主要的组织学亚型：透明细胞肾细胞癌（ccRCC）（~75%）、乳头状肾细胞癌（15–20%）和嫌色肾细胞癌（~5%）^{3,4}。除了组织病理学上定义的RCC亚型外，Xp11.2易位RCC，一种与TFE3基因融合相关的罕见亚型，在2004年的WHO肾肿瘤分类中首次被正式承认。TFE3基因位于染色体Xp11.2上的TFE3基因，有多种融合伙伴^{5–7}。具有t（6；11）易位的肾细胞癌，包含MALAT1–TFEB基因融合，是非常少见。

TFE3 Xp11.2易位肾细胞癌（TFE3–RCC）常在晚期诊断，比非Xp11.2易位RCC更具侵袭性，预后较差。近年来，肾癌的靶向治疗取得了重大进展⁸，特别是VEGF靶向药物（舒尼替尼、索拉非尼等）。和mTOR靶向（替西罗莫司、依维莫司等）阻断血管生成活性的疗法^{9–11}。在过去的几年中，有许多研究调查了靶向治疗TFE3–RCC7患者的疗效^{12–16}。例如，周埃里等人¹⁴在一项小型回顾性研究中显示，VEGF靶向药物在转移性TFE3–RCC患者中显示出一定的疗效。改善这种罕见肾细胞癌亚型的诊断不足将有助于样本管理，改善临床试验的获取，更重要的是，有助于为这组患者开发有效的治疗方法。

然而，基于苏木精和伊红（H&E）染色的病理图像的目视检查，要区分TFE3–RCC与其他亚型是相当具有挑战性的。TFE3–RCC的大体形态与ccRCC5相似^{–7,17}。显微镜下，TFE3–RCC病例的上皮样透明细胞呈分支状，乳头状结构，带有纤维血管核心和/或巢状结构。虽然这些特征提示了TFE3–RCC，但形态谱是相当可变的，可以与其他RCC亚型重叠，如ccRCC或乳头状RCC1,2。例如，癌症基因组图谱（TCGA）项目的ccRCC和乳头状RCC数据集中的一些病例与TFE3或TFEB易位有关^{18,19}。

由于TFE3–RCC难以识别明显和可靠的形态学特征，易位的诊断可以通过双色、分离荧光原位杂交来证实。然而，这需要额外的时间来检测这种诊断，并且对于最初不怀疑为TFE3–RCC的RCC患者并没有常规检查。因此，TFE3–RCC被其他RCC亚型误诊的风险很高，这延误了适当的治疗。我们将机器学习应用于数字化的h&e染色的病理图像，并研究它是否有助于识别TFE3RCC独特的图像特征，并将TFE3–RCC与最常见的RCC亚型ccRCC区分开来。

随着数字幻灯片扫描仪变得越来越可靠和流行，玻片越来越多地数字化成全幻灯片图像。近年来，人们对将机器学习应用于h&e染色的病理图像，用于包括预后预测在内的各种任务越来越感兴趣^{20–22}，癌症分类^{23–26}，以及遗传状态预测，如微卫星不稳定性²⁷和基因突变²⁸。值得注意的是，坎帕内拉等人²³报告了一个临床级的计算病理学框架，并在44,732张全载玻片图像的数据集上进行了评估。结合图像处理技术和机器学习模型，Yu等人²⁶区分正常和肿瘤切片的曲线下面积（AUC）为0.85，肺腺癌和鳞状细胞癌切片的曲线下面积为0.75。这些研究证明了计算病理学在临床决策支持方面的有效性。

在这项研究中，我们收集了74例来自多个来源的TFE3–RCC患者（基于我们所知的最大的TFE3–RCC报道研究）和74例性别和肿瘤分级匹配的ccRCC患者的h&e染色全切片图像。本研究的目的是（i）识别TFE3–RCC和ccRCC之间存在显著差异的不同的定量图像特征；（ii）基于这些特征建立和评价客观和全自动的分类模型，以区分TFE3–RCC和ccRCC。

结果

患者特征和病理图像分析工作流程。我们收集了两个完整的幻灯片图像数据集：数据集1和数据集2。数据集1来自印第安纳大学，包括50例TFE3–RCC患者和50例性别和肿瘤分级匹配的ccRCC患者。数据集1被随机分为训练集（80%）和内部验证集（20%），使用5次交叉验证。数据集2来自密歇根大学和TCGA大学。它被用作外部验证集。其中包括24例TFE3–RCC患者和24例ccRCC患者，其性别和肿瘤分级也相匹配。表1总结了这两个数据集的患者的人口统计学和临床特征。

分析的工作效率如图所示。1. 148例切除活检病例的h&e染色切片在×40倍的徕卡Aperio扫描仪在40倍下进行数字化（图。1a）。一个病理图像分析管道从全幻灯片图像中提取定量图像特征²¹，表征细胞核的大小、染色、形状和密度（图。1b）。研究图像特征与疾病亚型的关系（即TFE3–RCC vs ccRCC；图。1c），首先使用Mann–WhitneyU检验比较两种亚型之间各图像特征的分布。然后，结合图像特征，建立4个机器学习模型（逻辑回归、线性核SVM、高斯核SVM、随机森林），将患者分为TFE3–RCC组和ccRCC组。

特征提取管道包括核分割、核级特征提取和图像级特征提取三个步骤。2）。首先，采用分层多层阈值方法对全幻灯片图像中的细胞核进行分割²⁹（图2a）。接下来，对每个分段的核，计算10个核水平的特征（图。2b）。10个核级特征的代表性图像斑块如表2所示。最后，由于每张全幻灯片图像包含数百万个细胞核，通过结合10–bin直方图和5个分布统计数据（平均值、标准数据、偏度、峰度和熵），将每种类型的核级特征分解为15个图像级特征。2c）。直方图的bin中心是通过聚类从训练集采样的每一种核级特征来确定的聚类中心；因此，直方图特征在患者之间具有可比性。15个成象立面特征的命名规则如图所示。2c，使用核级特征（e. g.，比率我们总共计算了150个图像级特征

表1两个全幻灯片图像数据集的人口统计学和肿瘤特征。

特点	数据集1: TFE3–RCC/ccRCC	数据集2: TFE3– RCC/ccRCC
不患者性别男性	50/50	24/24
女性2级2级3级	22/22	9/9
福尔曼等级: 4	28/28	15/15
	10/10	6/6
	29/29	15/15
	11/11	3/3

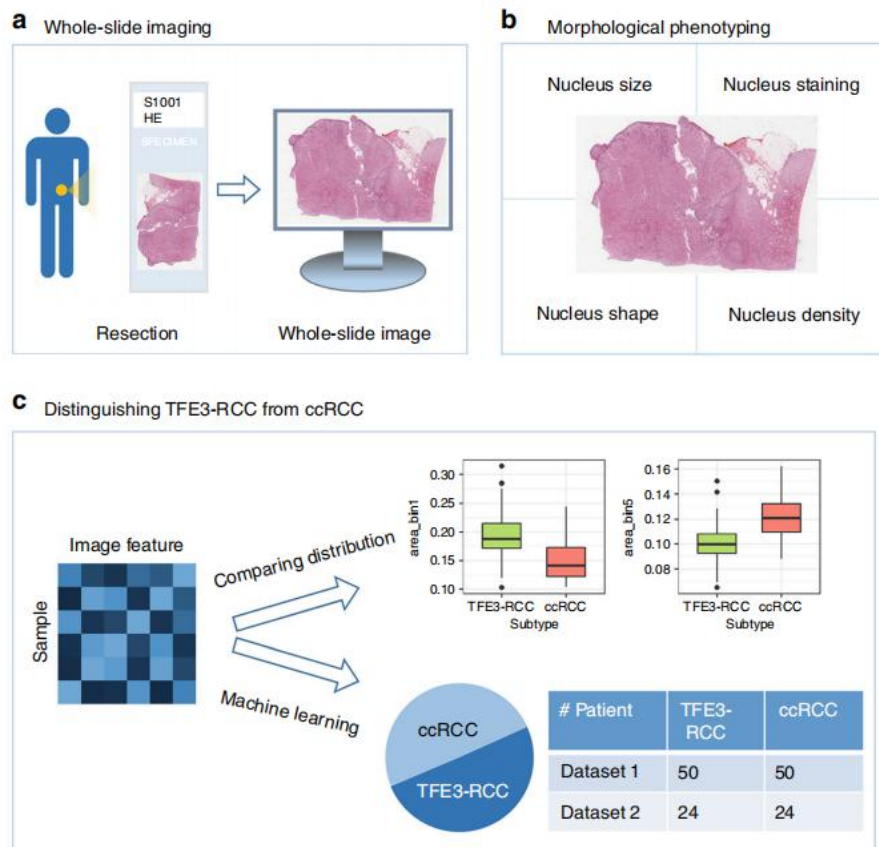


图1工作方案。用扫描仪对h&e染色的组织切片进行数字化，以获得全载玻片图像。b提取大量的定量图像特征，表征细胞核的大小、染色、形状和密度。c采用Mann-WhitneyU检验比较TFE3RCC和ccRCC的图像特征，并基于图像特征建立机器学习模型，对两种癌症亚型进行自动分类。在c的框图上，中心标记表示中位数，框的底部和顶部边缘表示第25和第75百分位数（ q_1 和 q_3 各自地上端晶须从 q 开始延伸3到 $q_3 + 1.5 \times (q_3 - q_1)$ ，而较低的晶须从 q 开始延伸出来1到 $q_1 - 1.5 \times (q_3 - q_1)$ ），而须末端以外的数据是单独绘制的孤立点。

每个幻灯片图像。更多的细节可以在圆形到拉长中找到。如图所示。3、ratio_bin1为“方法”部分。未被充分代表的相比之下，ratio_bin3，ratio_bin4，ratio_bin5，

和ratio_std的代表过多。总之，这些观察结果

这表明ccRCC倾向于有更多的非常圆的细胞核。

定量图像特征显示TFE3-RCC与ccRCC之间存在显著差异。我们对每个特征应用Mann-WhitneyU检验，经过多次检验校正，发现TFE3-RCC和ccRCC之间存在52个显著差异的特征（5%的错误发现率；图。3）。与TFE3-RCC亚型相比，显著特征代表过多或不足；i.e.，如果TFE3-RCC组中该特征的中位数高于ccRCC组，则该特征被定义为过多的代表。

对于图中与细胞核大小相关的特征。3，我们发现area_bin1、area_bin9和area_bin10在TFE3-RCC中占的比例过多，而area_bin4、area_bin5和area_bin6的比例不足。从area_bin1到area_bin10的图像特征代表了大小的核的比例。因此，这些显著特征表明，TFE3-RCC的核大小比ccRCC的核大小更不均一，更接近两个极端，这也得到了过度代表的特征area_std（核大小的标准偏差）的支持。

名字以大调、小调和比例在无花果。3是由拟合的椭圆得到的核这些特征与细胞核的形状有关。特别是，从ratio_bin1到ratio_bin10的特征直接描述了形状变化的百分比

以红色和绿色通道计算的11个细胞核染色相关特征显示TFE3-RCC和ccRCC之间存在显著差异。在这些特征中，rMean_bin8、rMean_bin9、rMean_mean、rMean_skewness和gMean_mean在TFE3-RCC病例中的比例过高。rMean_bin8和rMean_bin9表示在红色通道中平均像素值很大的原子核的比例。rMean_mean和gMean_mean分别表示红色通道和绿色通道中所有原子核的平均像素值的平均值。rMean_skewness的比例过高，说明TFE3-RCC中红色通道中细胞核的平均像素值的数据分布比ccRCC中更不对称。

在15个显著的核密度相关特征中，我们发现5个特征被过度代表：distMin_bin1、distMin_bin2、distMean_bin1、distMean_bin2和distMax_bin1。这五种特征的过度表现表明，与ccRCC相比，TFE3-RCC倾向于呈现更多彼此非常接近的核。换句话说，TFE3-RCC中的细胞更容易聚集在一起。

基于图像特征的分类模型可以有效地区分TFE3-RCC和ccRCC。我们首先在数据集1上使用五倍交叉验证来训练和评估我们的分类器

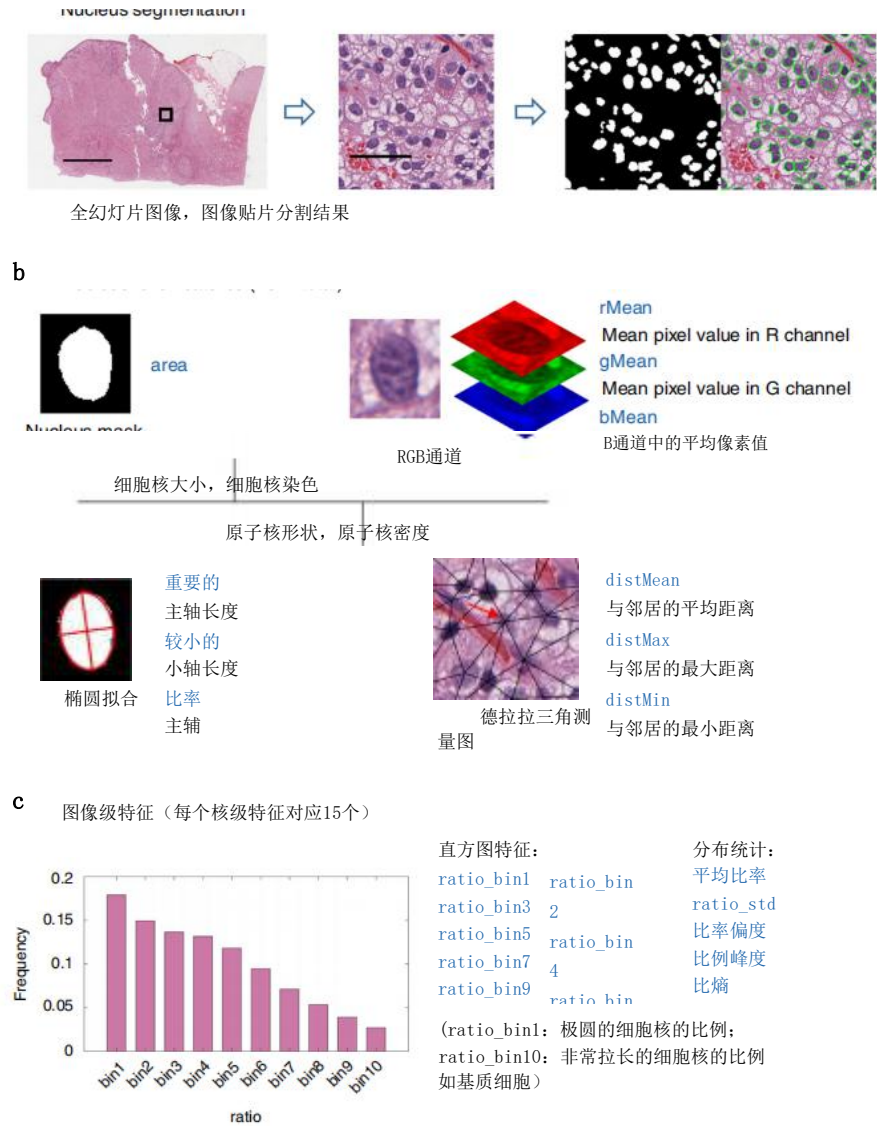


图2特征提取管道。a全幻灯片图像中的细胞核被自动分割。b对于每个分段核，提取10个核水平特征，包括核大小、染色强度、形状和密度。c对于同一全幻灯片图像中的每一种核级特征，使用10-bin直方图和five分布统计数据将其分解为15个图像级特征。比例尺：5mm（全幻灯片图像）和50μm（图像贴片）。

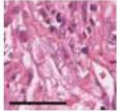
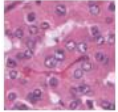
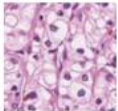
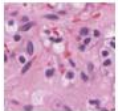
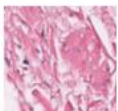
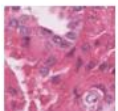
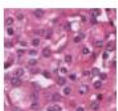
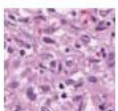
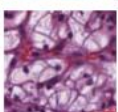
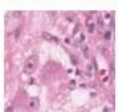
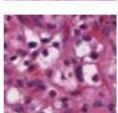
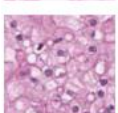
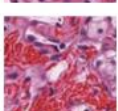
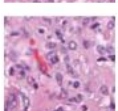
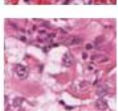
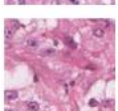
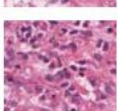
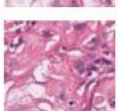
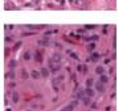
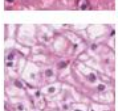
从印第安纳大学获得（详见表1）。在这五轮测试中，数据集1被随机划分为两组：80%的训练和20%的内部验证。我们的结果表明，使用最小冗余最大相关性（mRMR）算法选择的30个特征，我们的最佳分类器，高斯核SVM，平均AUC为0.886。四种分类器（逻辑回归、SVM线性核SVM、SVM高斯核和随机森林）的性能没有显著差异（方差分析检验P值= 0.77）。四种分类器的五次交叉验证结果的柱状图如图所示。4a.

使用外部数据集（数据集2；表1）进一步验证了我们的定量图像特征在诊断分类中的效用。具体来说，我们使用数据集1训练相同的四个分类器，然后使用数据集2验证其性能。所有分类器的AUC与上述内部交叉验证集的AUC相似。4b). 我们还观察到，除了随机森林分类器外，其他三个分类器的AUC都略高

比使用数据集1进行5倍交叉验证的平均AUC相比。这可能是因为数据集1中的所有患者都被用来训练在数据集2上测试的分类模型。相比之下，在数据集1上的5倍交叉验证中，数据集1中只有80%的患者被用作训练集。mRMR选择的顶级定量特征（以特征重要性评分衡量）包括ratio_bin3、rMean_mean、minor_std、area_bin5、rMean_skewness、distMin_bin5、rMean_std和ratio_std。

讨论

据我们所知，这是第一个提供一个计算模型，利用从h&e染色的全载玻片图像中提取的定量组织病理学特征来区分TFE3-RCC和ccRCC的研究。在这项研究中，我们实现了一个自动工作流，从图像中计算出150个客观特征。从整个幻灯片中提取图像特征，不仅覆盖了一个很大的肿瘤区域，而且

表2是10个核水平特征的说明。		
功能名称	解释	具有大值的补丁
核面积大小（单位：像素）		 322  545
长主轴长度（单位：像素）		 26  30
小小轴长度（单位：像素）		 14  20
比率	主要与次要比率	 1.4  1.7
R通道中的平均像素值		 99  169
G通道中的平均像素值		 55  108
B通道中的平均像素值		 102  153
距离测量到邻居的平均距离（单位：像素）		 60  95
与邻居的最大距离（单位：像素）		 86  123
距离到邻居的最小距离（单位：像素）		 31  41
每个图像补丁旁边的数字是补丁中所有核的特征值的平均值。例如，使用所有的补丁。		50µm

覆盖了广泛的细胞核形态，包括细胞核的大小、染色、形状和来自异质性肿瘤组织的密度。我们建立并评估了机器学习模型，将患者分为TFE3-RCC或ccRCC。该工作效率的有效性是通过从不同来源收集的独立数据集来确认的。

大多数癌症是异质性的，并包含几个亚型^{1,2}。这些亚型通常具有不同的分子谱特征，从而驱动肿瘤的发展和进展方式不同⁹⁻¹¹。在诊断癌症时，通常会收集组织病理学切片。我们的假设是肿瘤的形态学特征

表型可以通过人工智能算法进行定量检测，该算法可以反映潜在的遗传畸变，包括易位。TFE3-RCC由Xp11.2上的特异性易位决定。据我们所知，我们报道了74例最大的TFE3-RCC队列，并使用计算病理图像分析对TFE3-RCC和ccRCC的显微镜外观进行了广泛的分析。我们的研究表明，应用基于定量组织病理学特征的机器学习模型来区分TFE3-RCC和ccRCC，具有令人印象深刻的准确性（AUC在0.842–0.894之间）

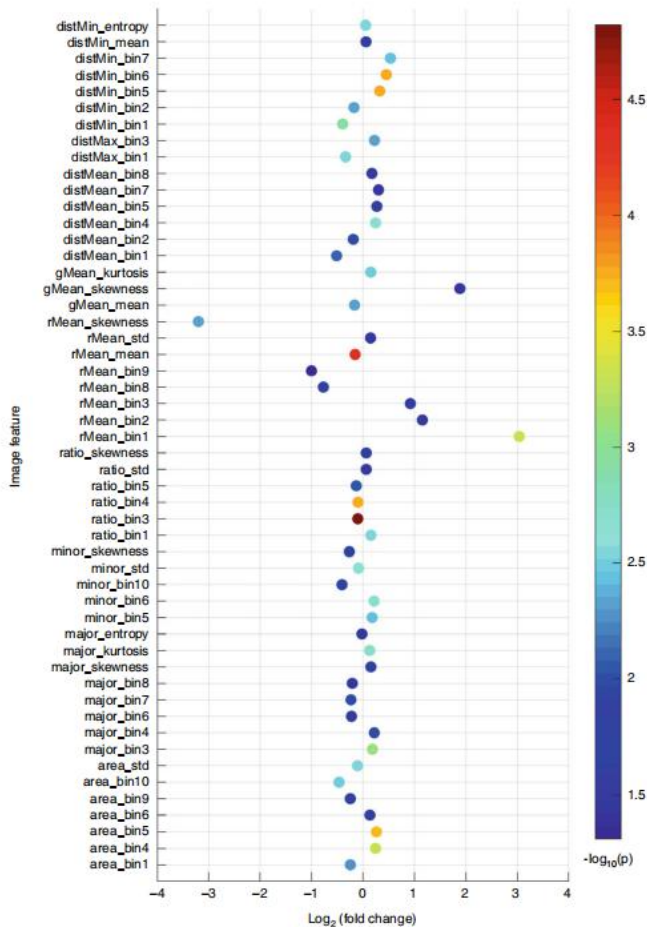


图3TFE3-RCC与ccRCC的图像特征比较。对于每个特征，折叠变化被定义为ccRCC和TFE3-RCC之间的中值特征值的比值。采用双侧Mann-WhitneyU检验确定了TFE3-RCC和ccRCC之间存在显著差异的52个图像特征。在5%的水平上，使用错误发现率程序进行了多重比较校正。

验证集该工具的优势将减轻TFE3-RCC的诊断不足，并促进针对这组患者的样本管理或临床试验访问。

我们鉴定了52个在两种亚型之间有显著不同的图像特征。例如，与ccRCC相比，TFE3-RCC的非常小和非常大的核比例更高（见图中的area_bin1、area_bin9和area_bin10。3），这与TFE3-RCC更具侵袭性与更高的肿瘤分级相关的事实相一致³⁰因为高级别肿瘤的细胞增殖速度更快。一位高级病理学家（LC）被咨询了显著不同的特征。虽然对于某些特征，很难通过人的眼睛分辨它们的差异，但其他特征可以被视觉感知。例如，我们发现ccRCC有更高比例的非常圆的核（见图中的ratio_bin1。3）比TFE3-RCC。病理学家证实，ccRCC确实比TFE3RCC的细胞核更圆。另一个例子是，我们的特征的过度表示（即，distMean_bin1和distMean_bin2；图。3）表明TFE3-RCC中的细胞团块比ccRCC更多，这也被观察到（补充图。1）。

由于TFE3易位导致TFE3蛋白的过表达，TFE3蛋白的免疫组化（IHC）被认为是这一遗传事件的替代物。我们将我们的方法与其他使用IHC的研究的性能进行了比较。夏兰等人。³¹发现在一个两个

实验室研究表明，TFE3 IHC对TFE3重排肿瘤的总敏感性和特异性分别为85%和57%，B实验室分别为70%和95%，导致约登指数分别为0.42和0.65（约登指数=敏感性+特异性-1）。他们的数据集包含27个TFE3重排肿瘤和98个对照。我们的高斯核SVM分类器的灵敏度为91.7%，特异性为79.2%，约登指数为0.708。4b）。值得注意的是，我们的基于病理图像的分类器只依赖于常规的H&E染色，而不是对特定分子的染色。

以往对TFE3-RCC临床病理特征的研究往往是样本量小³²。我们的基于病理图像的分类器可以帮助病理学家诊断新的TFE3-RCC病例，也可以帮助进行大规模的回溯性研究，检索被误诊的旧TFE3RCC病例。当使用适当的阈值时，分类器可以自动从组织病理学切片档案中识别TFE3-RCC病例，具有非常高的灵敏度和相对较低的假阳性率（图。4b）。例如，我们的带有高斯核的SVM分类器可以达到91.7%的灵敏度，同时保持20.8%的假阳性率。鉴于大多数RCC是ccRCC，其临床应用将使病理学家排除许多真阴性（ccRCC）进行进一步评估，或提名可疑病例进行进一步评估。

我们还测试了不同机构之间（即不同的扫描仪器或载玻片制备）对H&E载玻片染色的差异是否会影响我们的方法的泛化性能。我们的外部验证集（数据集2）中的幻灯片来自几个机构（密歇根大学和TCGA；TCGA案例本身也来自不同的机构），它们与数据集1中的幻灯片有不同的颜色外观。我们在没有颜色归一化步骤的情况下应用了相同的分析工作效率，并观察到在外部验证集上的泛化性能大幅下降（补充图。2）。这表明，在处理来自不同来源的全幻灯片图像时，颜色归一化是一个关键的步骤。

此外，我们还在数据集1上测试了一个卷积神经网络ResNet18。整张幻灯片的图像被调整到224-224像素，以便输入ResNet18。ResNet18在80%的病例上进行了训练，并在其余病例上进行了5倍交叉验证。实施了两种训练策略，即从头开始训练网络和迁移学习。对于基于预先训练的ResNet18网络的迁移学习，只更新了最后两层（全连接层和softmax层）的权值，并冻结了早期层的权值。五倍交叉验证产生的平均AUC为0.518，迁移学习产生的平均AUC为0.696。迁移学习的表现更好，这可能是由于在使用迁移学习时需要学习的参数要少得多。与auc在0.8-0.9之间的分类模型相比，ResNet18的性能较差。众所周知，深度神经网络学习到的特征难以解释。然而，我们的分类管道是基于细胞图像特征，这些特征在细胞和组织形态上具有明确的意义，因此在临床诊断中更容易解释和可取。

这项研究有几个局限性。肿瘤内异质性是肾细胞癌中一个有充分证据证明的现象⁹⁻¹¹。由于我们无法从同一病例中收集多个福尔马林固定的、石蜡包埋的组织块，因此我们不能准确地评估肿瘤内的异质性（ITH）。尽管如此，在我们的研究中，全载玻片图像都来自于手术切除标本。手术切除标本覆盖了更大的区域

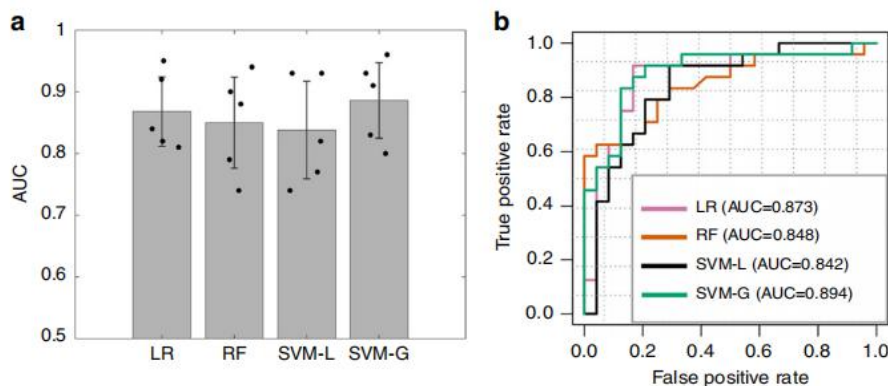


图4. 四种机器学习模型的性能。在数据集1上使用fove折叠交叉验证的分类性能（每个模型n=5个实验）。对于每个患者，80%的患者作为训练集，其余的患者作为内部验证集。b在外部验证集（数据集2）中对TFE3-RCC和ccRCC进行分类的接收机工作特征曲线。模型使用数据集1进行训练，并使用数据集2进行评估。AUC的95%证据区间：LR（0.763–0.984）、RF（0.736–0.960）、SVM-L（0.725–0.959）和SVM-G（0.797–0.991）。LR、逻辑回归；RF、随机森林；SVM-L，带线性核的SVM；SVM-G，SVM带有高斯核。数据用a中的平均±SD表示。

肿瘤与穿刺活检的比较。此外，我们的算法通过使用形态学特征值（超过十个箱子的直方图）的分布作为成像特征来考虑ITH。尽管内部和外部验证集的一致相似的性能证明了我们的成像特征和分类模型的稳定性和可重复性，但如果从同一肿瘤的多个部位进行评估，证明这些特征是稳定的将更严格。另一个重要的局限性是，我们的研究使用了匹配的ccRCC与TFE3-RCC进行比较。TFE3-RCC1有多种形态学表现^{2,5}。它们还模拟了乳头状肾细胞癌、透明细胞乳头状肾细胞癌、未分类的肾细胞癌、嫌色肾细胞癌、嗜酸细胞瘤和其他罕见的肾肿瘤。未来的研究应包括其他肾肿瘤类型和匹配病例中的组织学变异进行比较。

综上所述，我们通过证明了基于定量特征的组织病理学图像分类器可以在外部验证集上成功区分TFE3-RCC和ccRCC，准确率较高（AUC为0.894），这证实了我们的假设，即肿瘤组织学表型可以改变潜在的基因易位。我们的方法可以促进基于常规收集的h&e染色组织病理学切片的TFE3-RCC诊断，从而有助于这种罕见和侵袭性癌症亚型的准确样本管理和治疗发展。

方法

样本收集。两个h&e染色的全幻灯片图像的数据集（148张图像

总共）被收集。TFE3-RCC患者与ccRCC患者的比值为1:1，两种亚型之间的性别和肿瘤分级信息相匹配。数据集1包括50名TFE3-RCC患者和50名ccRCC患者，均来自印第安纳大学。数据集2作为外部验证集收集，包含14名来自密歇根大学的TFE3-RCC患者，10名来自TCGA的TFE3-RCC患者³³，24例ccRCC患者。所有肿瘤标本均采用手术切除的方式采集。组织切片以×40倍放大倍数扫描。分析中未纳入TFEB重排易位RCC。我们没有试图根据TFE3与不同伴侣基因的重排来对TFE3-rcc进行亚分类。个人健康信息在我们的数据集中被去识别，因此这是一项机构审查委员会批准的豁免研究。

荧光原位杂交技术。对所有肿瘤均进行了间期荧光原位杂交检测，描述如下^{34–36}。所有TFE3-RCC病例均经FISH分析确诊。具体来说，4 μm厚的组织切片由含有肿瘤的缓冲福尔马林固定、石蜡包埋的组织块制备。载玻片用二甲苯（每次15 min）洗涤两次，然后用无水乙醇（每次10 min）洗涤两次，然后在引擎盖中风干。然后用10 mm柠檬酸（pH 6.0）（Zymed，旧金山，CA，USA）在95° C下处理10 min，冲洗

蒸馏水洗涤3 min，然后用2×SSC洗涤5 min。通过应用0.4 ml胃蛋白酶（5mg/ml，0.01N盐酸和0.9%氯化钠）（Sigma，圣路易斯，Mo，美国）在37° C下进行40 min，进行组织消化。用蒸馏水冲洗3 min，2×SSC洗涤5 min，风干。TFE3的分离探针使用BAC克隆RP11-528A24（116 kbp，位于TFE3的着丝粒，用5-荧光素dUTP标记）和RP11-RP116B14（182 kbp，位于TFE3的端粒，用5-ROX dUTP标记）（帝国基因组学，布法罗，纽约州，美国）。用DenHyb2按1:25的比例稀释TFE3的BAC克隆。在还原光条件下，将稀释后的探针（5 μl）应用于每张载玻片。然后用一个22×22毫米的覆盖物覆盖载玻片，并用橡胶水泥密封。将载玻片在83° C的湿箱中孵育12 min，在37° C杂交过夜，实现变性。取下盖子，用0.1×SSC洗涤2次。在45° C（每个20 min）条件下，每1.5 M尿素洗涤1×SSC，然后用2×SSC洗涤20 min，用2×SSC每0.1×SSC。1%NP-40，10 min，45° C。再用室温2×SSC洗涤载玻片5 min。载玻片风干，用10 μl 4',6-二胺-2-苯基乙烯（天然）染色，盖上盖子，指甲油密封。

切片用蔡司Axioplan 2显微镜（蔡司，哥廷根，德国）进行检查。图像用CMOS相机获取，并用元系统软件（MetaSystem，Belmont，MA，USA）进行分析。获得5个具有0.4 mm间隔的连续聚焦堆栈，然后集成到单个图像中，以减少与厚度相关的伪影。对于每个病例，至少用×1000倍放大的荧光显微镜检查了100个肿瘤细胞核。只评估了不重叠的肿瘤细胞核。TFE3融合导致了一种分裂信号模式。当绿色和红色信号被两个或两个以上的信号直径分开时，信号被认为是分裂的。在此基础上，以及其他商业上可行的分离FISH检测和TFE3分离FISH检测，当≥10%的肿瘤核显示分裂信号模式时，报告了阳性结果（补充图。3）。

从全幻灯片图像中提取定量特征。每个维度

整张幻灯片的图像从4万到13万像素。这些图像被细分为2000×2000大小的瓷砖，以方便处理。考虑到机构间的颜色变化，在特征提取之前，我们使用结构保留颜色归一化算法将数据集2中图像的颜色外观转换为数据集1中图像的颜色外观³⁷。为了从患者身上提取的核水平特征汇总为患者水平特征，采用了组织图和分布统计。为了构建直方图特征，我们使用了一个视觉词袋模型^{38–40}。词袋模型是一种最初用于自然语言处理和检索的特征表示方法。在这个模型中，一个文本被表示为一个单词频率的直方图（即，直方图中的每个bin表示某些单词在文本中出现的频率）。这种方法已被考虑图像特征的计算机视觉广泛采用。在本研究中，对于每种类型的核水平特征，我们创建了一个核水平特征的直方图。在这个直方图中，这些单词（即箱子的中点）是通过聚类得到的聚类质心

来自训练集的核级特征。具体来说，对于每种类型的核级特征，从训练集中收集大量的核级特征，并输入K-means算法学习10个代表性单词（即聚类质心）。使用交叉验证的方法来选择集群的数量（补充图。4）。然后，利用欧氏距离将从整个幻灯片图像中提取的核级特征分配到它们最近的箱子中，从而得到一个直方图

每个患者和每种核水平特征的单词计数。对得到的直方图进行11归一化，以消除具有不同核数的全载玻片图像的影响。在分布统计方面，计算了每种单元级特征的5个参数：*i. e.*，平均值、标准差、偏度、峰度和熵。熵基于归一化直方图计算熵。

TFE3-RCC与ccRCC的图像特征分布的比较。

为了确定TFE3-RCC和ccRCC之间特定图像特征的形态差异，我们使用双侧Mann-WhitneyU检验比较了两种亚型之间每个图像特征的分布。为了校正多重比较，我们根据本杰米尼和霍赫伯格的调整，通过错误发现率程序调整了P值⁴¹。调整后的P值<0.05被认为有统计学意义。

-TFE3和ccRCC的机器学习方法。由于高

图像特征的维数和相对较小的样本量，数据可能存在过拟合；因此，在建立分类模型之前，我们进行了特征选择，以避免过拟合问题。采用mRMR算法降低了特征维数⁴²使用R包mRMR。在各种任务中，mRMR已被证明是一种鲁棒的特征选择算法⁴³⁻⁴⁵。将mRMR算法应用于与样本的类标签相关的所有图像特征(*i. e.*，TFE3-RCC或ccRCC)来选择一个信息丰富和非冗余的特征集。

采用逻辑回归、带有线性或高斯核的SVM和随机森林进行监督机器学习。R版本3.5用于训练和测试分类模型，glmnet包用于逻辑回归，随机森林包用于随机森林，e1071包用于SVM。在数据集1中，使用了5倍交叉验证。为了使用外部验证集进一步验证我们的方法，我们使用数据集1训练分类模型，并使用数据集2进行评估。AUC用R包pROC计算和置信区间。

数据可用性

从H&E染色的全幻灯片图像中提取的定量图像特征可从GitHub网站获得(<https://github.com/chengjun583/tRCC-ccRCC-classification>)。其余数据可在文章、补充信息文件中获得，或在合理的要求下可从作者处获得。

代码可用性

这项工作的源代码可以从GitHub下载到(<https://github.com/成军583/tRCC-ccRCC-分类>)。

收到日期：2019年9月5日；接受日期：2020年3月23日；

Published online: 14 April 2020

参考文献

- 程, L., 等。泌尿外科病理学, 第4版。(爱思唯尔, 2019)。
- 麦克伦南, T. 和程, L. 50年泌尿病理学: 肾细胞肿瘤知识的加速扩展。哼唱帕索尔。95, 24 - 45 (2020)。
- MochH. 以及其他2016年世界卫生组织泌尿系统和男性生殖器官肿瘤分类-A部分: 肾脏、阴茎和睾丸肿瘤。欧洲人Urol。70, 93 - 105 (2016)。
- 里基茨, C. J. 以及其他肾细胞癌的癌症基因组图谱的综合分子特征。细胞代表。23, 313 - 326. e315 (2018)。
- Argani, P. MiT家族易位性肾细胞癌。塞明。诊断。帕索尔。32, 103 - 113 (2015)。
- Komai, Y. 以及其他经细胞遗传学和免疫组化诊断的成人Xp11易位性肾细胞癌。Clin. 癌症Res. 15, 1170 - 1176 (2009)。
- 马格斯, M. J. 以及其他MiT家族易位相关的肾细胞癌: 一个强调形态学, 免疫表型, 和分子模拟的当代更新。拱门。帕索尔。实验室医学139, 1224 - 1233 (2015)。
- Posadas, E. M. 以及其他肾细胞癌的靶向治疗。Nat. 发动机的旋转尼弗罗尔。13, 496 - 511 (2017)。
- 程, L. 以及其他了解肾细胞瘤的分子遗传学: 对诊断、预后和治疗的意义。专家牧师。抗癌者。10, 843 - 864 (2010)。
- 乔伊里, T. K. & Motzer, R. J. 转移性肾细胞癌的全身性治疗。N. 引擎。J. 医学376, 354 - 366 (2017)。
- Sanfrancesco, J. M. 和程, L. 肾细胞癌基因组景观的复杂性: 对靶向治疗和精确免疫肿瘤学的意义。令状。发动机的旋转Oncol. 血红素。119, 23 - 28 (2017)。
- 阿姆斯特朗, J. 以及其他依维莫司与舒尼替尼治疗转移性非透明细胞肾细胞癌 (ASPEN) 患者: 一项多中心、开放标签、随机的2期临床试验。《柳叶刀》。17, 378 - 388 (2016)。
- 贝尔蒙特, J. &达彻, J. 非透明细胞肾细胞癌的靶向治疗与治疗。安Oncol。24, 1730 - 1740 (2013)。
- 乔伊里, T. K. 以及其他血管内皮生长因子靶向治疗成人转移性Xp11.2易位性肾细胞癌。癌症116, 5219-5225 (2010)。
- 达马安蒂, N. P. 以及其他TFE3/IRS1/PI3K/mTOR轴靶向治疗易位性肾细胞癌。Clin. 癌症Res. 24, 5977 - 5989 (2018)。
- Tannir, N. M. 以及其他依维莫司与舒尼替尼对转移性非透明细胞肾细胞癌 (ESPN) 的前瞻性评价: 一项随机多中心2期临床试验。欧洲人Urol。69, 866 - 874 (2016)。
- S. 斯卡拉L. 以及其他检测6例TFEB扩增的肾细胞癌和25例MITF易位的肾细胞癌: 对85例采用临床TFE3和TFEB FISH检测的85例进行系统形态学分析。模块。帕索尔。31, 179 - 197 (2018)。
- 癌症基因组图谱研究网络。透明细胞肾细胞癌的综合分子特征。自然499, 43 - 49 (2013)。
- 癌症基因组图谱研究网络。以及其他肾乳头状细胞癌的综合分子特征。N. 引擎。J. 医学374, 135 - 145 (2016)。
- 郑, J. 以及其他肾肿瘤微环境中与患者生存相关的拓扑结构特征的识别。生物信息学34, 1024-1030 (2018)。
- 郑, J. 以及其他人对组织病理学图像和基因组数据的综合分析可预测透明细胞肾细胞癌的预后。癌症Res. 77, e91 - e100 (2017)。
- 纳特拉詹, R. 以及其他微环境的异质性与乳腺癌的进展相似: 一种组织学-基因组整合分析。PLoS医学。13, e1001961 (2016)。
- 坎帕内拉, G. 以及其他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)。
- 苏达山, P. J. 以及其他组织病理学乳腺癌图像分类的多实例学习。专家系统。Appl 117, 103 - 111 (2019)。
- 徐, Y. 以及其他弱监督的组织病理学, 肿瘤图像的分割和分类。医学图像肛门。18, 591 - 604 (2014)。
- 余, K. H. 以及其他利用全自动显微镜病理图像特征预测非小细胞肺癌的预后。Nat. 通勤。7, 12474 (2016)。
- Kather, J. N. 以及其他深度学习可以直接从胃肠道肿瘤的组织学中预测微卫星的不稳定性。Nat. 医学25, 1054 - 1056 (2019)。
- 库德雷, N. 以及其他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 - 1567 (2018)。
- 艾哈迈迪Phowlady, H. 以及其他分层多层阈值分割的组织图像的核分割。医学成像2016: 数字病理学, (2016)。
- 徐, L. 以及其他Xp11.2. 年轻人中的易位性肾细胞癌。BMC Urol. 15, 57 (2015)。
- Sharain, R. F. 以及其他TFE3的免疫组化在诊断TFE3-重排肿瘤时缺乏特异性和敏感性: 一项比较的, 2个实验室研究。哼唱帕索尔。87, 65 - 74 (2019)。
- 克莱斯, M. 以及其他易位性肾细胞癌的发病率、临床病理特征和融合转录图谱。组织病理学70, 1089-1097 (2017)。
- 爸爸, M. 以及其他TFE3 Xp11.2易位肾细胞癌小鼠模型揭示了新的治疗靶点, 并确定了GPNMB作为人类疾病的诊断标志物。摩尔癌症Res. 17, 1613 - 1626 (2019)。
- 卡利奥, 以及其他伴有人TFE3易位和琥珀酸脱氢酶B突变的肾细胞癌。模块。帕索尔。30, 407 - 415 (2017)。
- 程, L. 以及其他荧光原位杂交在外科病理学中的原理和应用。J. 帕索尔。Clin. Res 3, 73 - 99 (2017)。
- 饶, 问。以及其他人与单独的TFE3或组织蛋白酶K免疫组化染色相比, TFE3分离的FISH对Xp11.2易位相关的肾细胞癌具有更高的敏感性: 扩大了形态学谱。是J. 外科医生帕索尔。37, 804 - 815 (2013)。
- 瓦哈丹。以及其他组织学图像的结构彩色归一化和稀疏染色分离。IEEE跨。医学1962年至1971年 (2016年)。
- 本泰布, A. 以及其他从组织病理学切片中得出的卵巢癌亚型的结构化潜伏模型。医学图像肛门。39, 194 - 205 (2017)。
- 郑, J. 以及其他通过肿瘤区域的增强和分割提高脑肿瘤分类性能。《公共科学图书馆》ONE 10, e0140381 (2015)。
- 郑, J. 以及其他利用自适应空间池化和费雪向量表示法检测脑肿瘤。《公共科学图书馆》ONE 11, e0157112 (2016)。
- Benjamini, Y. 霍奇伯格, Y. 控制错误发现率: 一种实用而强大的多重测试方法。J. R. 斯达社会重量的单位B统计。方法。57, 289 - 300 (1995)。

42. 彭, H. 以及其他基于互信息的特征选择: 最大依赖性、最大相关性和最小冗余性的标准。IEEE跨. 模式肛门. 马赫数. 知识27, 1226 – 1238 (2005).
43. 拉多维奇, M. 以及其他时间基因表达数据的最小冗余最大相关性特征选择方法。BMC Bioinforma. 18, 9 (2017).
44. 里奥斯·委拉斯开兹, E. 以及其他人体细胞突变驱动肺癌的不同成像表型。癌症Res. 77, 3922 – 3930 (2017).
45. 徐, Y. 以及其他Mal-Lys: 通过mRMR特征选择预测整合序列特征的蛋白质中赖氨酸丙二酰化位点。科学. 棱纹平布6, 38318 (2016).

致谢

这项工作得到了美国癌症协会对印第安纳大学(J. Z.), 国家自然科学基金资助项目。国家重点研发计划项目(第275号。2019YFC0118300)、深圳孔雀计划(KQTD2016053112051497和KQJSCX20180328095606003)、印第安纳大学预防健康计划、SZU健康科学中心青年教师支持计划(No. 71201-000001), SZU自然科学基金(编号为。2019131), 广东省医学科研基金项目(No. B2018031)。

作者贡献

J. C., L. C., K. H. 和J. Z. 构思和设计了这项研究。J. C. 和Z. H. 在W. S., R. M., M. C., Q. F., 和D. N. 这篇论文是由J. C., J. Z. 和K. H. 与所有合著者的贡献。

竞争利益

作者声明没有任何相互竞争的利益。

其他信息

本文的补充信息可在<https://doi.org/10.1038/s41467-02015671-5>上获得。

有关材料的信件和要求应寄给D. N., K. H., L. C. 或J. Z.

自然通信感谢Samra Turajlic和另一个, 匿名的, 审稿人(s) 为他们对这项工作的同行评审的贡献。可提供同行评审员报告。

重印和许可信息可在<http://www.nature.com/reprints>上获得。

出版商的说明施普林格自然保持中立的管辖权主张在已出版的地图和机构附属机构。



本文是在知识共享协议下获得授权的

CC BY 4.0国际许可证, 允许使用、共享、

以任何媒介或格式进行改编、分发和复制, 只要您给予原作者和来源, 提供知识共享许可的链接, 并指出是否进行了更改。本文中的图片或其他第三方材料都包含在文章的知识共享许可中, 除非在材料的信用额度中另有说明。如果材料没有包含在文章的知识共享许可中, 并且您的预期使用不被法律法规允许或超过了允许的使用, 您将需要直接获得版权所有者的许可。要查看此许可证的副本, 请访问<http://creativecommons.org/licenses/by/4.0/>。

©作者(s) 2020