病例图像分析

摘要：数字病理图像分析通常以通用图像识别技术为基础。然而，由于数字病理图像和任务具有一些独特的特征，因此通常需要特殊的处理技术。在这篇综述中，我们描述了使用机器学习算法进行数字病理图像分析的应用，及其特定于数字病理图像分析的问题和可能的解决方案。

**机器学习方法**

HI的细分是本综述涵盖的头几年的一个流行类别。大多数工作基于图像处理技术，例如滤波，阈值和轮廓检测技术。由于分割的注释是一项非常耗时的任务，因此通常会找到无监督方法和监督方法。大多数早期工作都使用分割来向专家突出显示HI中的信息。特征提取旨在发现 HI 中的判别特征，并将其聚合到特征向量中以训练 ML 算法。大多数浅层分类器和集成方法使用这种向量表示来学习线性或非线性决策边界。我们将浅层方法分为两个子类别：整体分类器和集成方法。集成方法结合了几种不同的基础模型，以减少预测中的偏差和方差，并提高预测的准确性。属于这两个子类别的作品需要上一步的特征提取。

一、HI的分割

1、无监督方法

k 均值算法是一种用于聚类的无监督 ML 方法，已用于像素区域的分割。k均值[聚类](https://baike.baidu.com/item/%E8%81%9A%E7%B1%BB/593695?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/K%E5%B9%B3%E5%9D%87%E7%AE%97%E6%B3%95/_blank)算法发明于1956年，是一个聚类算法，把n的对象根据他们的属性分为k个分割，k < n。它与处理[混合正态分布](https://baike.baidu.com/item/%E6%B7%B7%E5%90%88%E6%AD%A3%E6%80%81%E5%88%86%E5%B8%83/1392906?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/K%E5%B9%B3%E5%9D%87%E7%AE%97%E6%B3%95/_blank)的最大期望算法很相似，因为他们都试图找到数据中自然[聚类](https://baike.baidu.com/item/%E8%81%9A%E7%B1%BB/593695?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/K%E5%B9%B3%E5%9D%87%E7%AE%97%E6%B3%95/_blank)的中心。它假设对象属性来自于[空间向量](https://baike.baidu.com/item/%E7%A9%BA%E9%97%B4%E5%90%91%E9%87%8F/132918?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/K%E5%B9%B3%E5%9D%87%E7%AE%97%E6%B3%95/_blank)，并且目标是使各个群组内部的[均方误差](https://baike.baidu.com/item/%E5%9D%87%E6%96%B9%E8%AF%AF%E5%B7%AE/9024810?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/K%E5%B9%B3%E5%9D%87%E7%AE%97%E6%B3%95/_blank)总和最小。假设有k个群组Si, i=1,2,...,k。μi是群组Si内所有元素xj的重心，或叫中心点。

其他聚类算法也已用于分割 HI。例如迭代自组织数据分析技术（ISODATA）等。值得一提的是，基于DL方法的无监督方法也被提出用于分割HI。

2、监督方法

二、4. HI 的特征提取

监督浅层方法在执行分类之前依赖于从原始数据中提取的特征。特征提取方法处理图像并提供合理数量的特征来汇总图像的信息。这些方法旨在降低图像的维度并突出显示与问题相关的信息，例如是否存在特定结构、单个元素的数量、纹理和结构的形状。理想情况下，要素应独立于平移、缩放和旋转。

从HI中提取特征的主要挑战是从此类图像中存在的结构中提取形态特征，并寻找允许捕获与医学诊断相关的信息的更高级别的表示。形态特征与识别细胞变化有关，例如由于某些问题或有丝分裂阶段引起的细胞核变形，或组织变化，例如细胞密度或异常数量。形态特征与病理学家分析 HI 的方式有关，寻找对 HI 进行分类的具体理由。另一方面，高级特征是HI中所有结构的抽象，而不仅仅是细胞结构。出于这个原因，研究人员通常会利用频域中的纹理描述符或表示。HI 使用了几个特征，例如形状、大小、纹理、分形，甚至是这些功能的组合。

鉴于HI中丰富的几何结构和复杂的纹理，大多数作品都结合了不同类型的特征。形态特征对于表征几何结构至关重要，但由于它们需要复杂的预处理，例如，找到原子核的轮廓来计数它们，因此获得它们更复杂。

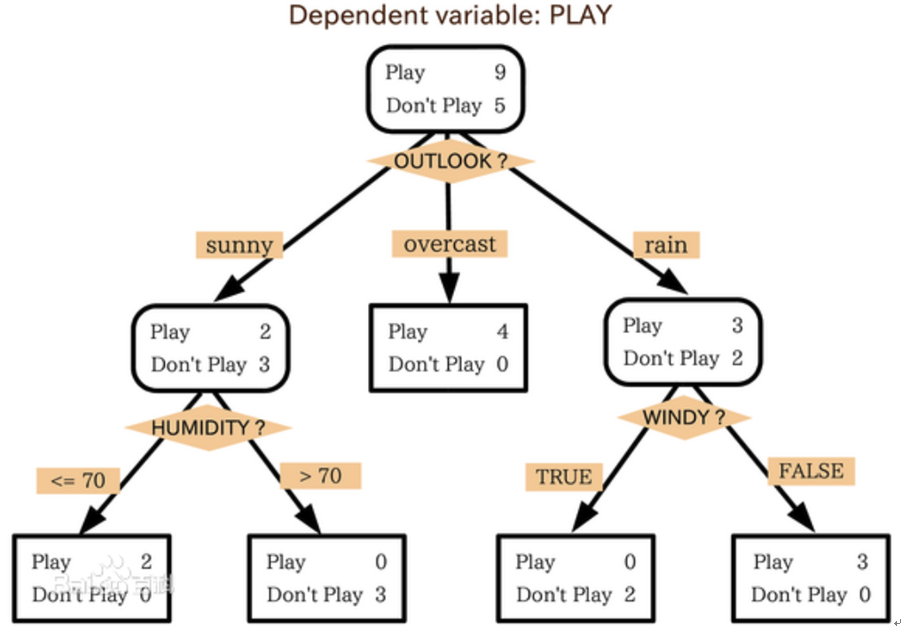
三、HI分类的浅层方法

1、整体分类器

HI分析中采用了不同的监督学习方法，例如支持向量机（SVM）、决策树（DT）、朴素贝叶斯（NB）、k 最近邻（k-NN）、多层感知器 （MLP）等。

支持向量机方法是建立在统计[学习](https://baike.so.com/doc/5378098-5614271.html" \t "https://baike.so.com/doc/_blank)理论的VC维理论和结构[风险](https://baike.so.com/doc/1002008-1059378.html" \t "https://baike.so.com/doc/_blank)最小[原理](https://baike.so.com/doc/5418626-5656791.html" \t "https://baike.so.com/doc/_blank)基础上的，根据有限的样本信息在模型的复杂性（即对特定训练样本的学习精度）和学习能力（即无错误地识别任意样本的能力）之间寻求最佳折中，以期获得最好的推广能力。

[机器学习](https://baike.so.com/doc/3290018-3465759.html" \t "https://baike.so.com/doc/_blank)中，决策树是一个预测模型;他代表的是对象属性与对象值之间的一种映射关系。树中每个节点表示某个对象，而每个分叉路径则代表的某个可能的属性值，而每个叶结点则对应从根节点到该叶节点所经历的路径所表示的对象的值。决策树仅有单一输出，若欲有复数输出，可以建立独立的决策树以处理不同输出。[数据挖掘](https://baike.so.com/doc/3186393-3357840.html" \t "https://baike.so.com/doc/_blank)中决策树是一种经常要用到的技术，可以用于分析数据，同样也可以用来作预测。



K最近邻(K-Nearest Neighbor,KNN)分类算法是数据挖掘分类技术中最简单的方法之一，是著名的模式识别统计学方法，在机器学习分类算法中占有相当大的地位。它是一个理论上比较成熟的方法。既是最简单的机器学习算法之一，也是基于实例的学习方法中最基本的，又是最好的文本分类算法之一。

2、集成方法

集成方法结合了多个碱基分类器的预测，试图提高单个分类器的泛化和鲁棒性。一些研究人员提出了组合分类器以提高HI方法的性能。

四、基于深度学习的方法

由于最近在解决大型数据集上的复杂机器学习问题方面取得的成就，深度学习方法正在引起科学界的关注。卷积神经网络（CNN）可以在单个优化过程中学习表示和决策边界。然而，CNN通常需要大量的数据来进行充分的训练，以避免过度拟合问题。尽管如此，大多数HI数据集只有少数患者和数百张图像，这可能会限制DL的使用。数据增强和迁移学习是规避HI数据集中数据缺乏的两种可能方法。例如，拥有超过14万张图像的ImageNet是用于训练CNN进行对象识别的最常见数据集之一。数据增强通过使用仿射变换或形态学运算从现有 HI 生成新的 HI。修补 HI 是另一种常见的数据增强方式。修补会产生选择具有相同结构但属于不同类的 HI 片段的效果。另一方面，迁移学习方法重用了以前在大型数据集中训练的CNN，这些数据集通常属于与目标问题不同的领域。

数字病理图像分析中机器学习的典型步骤。在对整个幻灯片图像进行预处理后，可以应用各种类型的机器学习算法，包括监督学习、无监督学习、半监督学习和多实例学习。组织病理学图像来自癌症基因组图谱。 [数字病理](https://link.zhihu.com/?target=https%3A//www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/digital-pathology" \t "_blank)图像分析中常用的机器学习技术分为有监督学习和无监督学习。监督学习的目标是推断一个函数，该函数可以使用训练数据将输入图像很好地映射到其适当的标签例如癌症。标签与 WSI 或 WSI 中的对象相关联。监督学习的算法包括[支持向量机](https://link.zhihu.com/?target=https%3A//www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/support-vector-machine" \t "_blank)、随机森林和卷积神经网络。另一方面，无监督学习的目标是从未标记的图像中推断出一个可以描述隐藏结构的函数。任务包括聚类、异常检测和降维。无监督学习的算法包括 k-means、自动编码器和主成分分析。

## 2023-04-07 00:04:40.495000

**机器学习在数字病理学中的应用**

**2.1 . 计算机辅助诊断**

在应用机器学习算法之前，应该进行一些预处理。例如，当在 WSI 中检测到癌症区域时，从大型 WSI 中采样大约 256 × 256 的局部小块。然后在每个局部补丁中进行癌症和非癌症之间的特征提取和分类。特征提取的目标是为机器学习任务提取有用的信息。各种局部特征，如灰度共生矩阵 (GLCM) 和局部二值模式 (LBP) 已被用于组织病理学图像分析，但深度学习算法如卷积神经网络从特征提取开始分析。在深度学习中同时优化特征和分类器，并且在深度学习中学习的特征在组织病理学图像分析中通常优于其他传统特征。

计算机辅助诊断数字病理图像分析中研究最活跃的任务是[计算机辅助诊断](https://link.zhihu.com/?target=https%3A//www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/computer-assisted-diagnosis" \t "_blank)(CAD)，这是病理学家的基本任务。诊断过程包含将一个 WSI 或多个 WSI 映射到疾病类别之一的任务，这意味着它本质上是一项监督学习任务。据报道，由于机器学习系统所犯的错误与人类病理学家所犯的错误不同，因此使用 CAD 系统可以提高分类精度。CAD 还可以通过调查 WSI 内的所有像素来减少解释的可变性并防止忽略。其他诊断相关的任务包括：（ROI）感兴趣的区域的检测或分割如在WSI肿瘤区域，记分[免疫染色](https://link.zhihu.com/?target=https%3A//www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/immunostaining" \t "_blank)，癌症分期，有丝分裂检测、腺体分割以及血管侵袭的检测和量化。

**2.2 . 基于内容的图像检索**

基于内容的图像检索 (CBIR) 检索与查询图像相似的图像。在[数字病理学中](https://link.zhihu.com/?target=https%3A//www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/digital-pathology" \t "_blank)，CBIR 系统在许多情况下都很有用，特别是在诊断、教育和研究中。例如，学生和初级病理学家可以将 CBIR 系统用于教育目的，以检索相关病例或组织的组织病理学图像。此外，此类系统对专业病理学家也有帮助，尤其是在诊断罕见病例时。

由于 CBIR 不一定需要标签信息，因此可以使用无监督学习。当标签信息可用时，监督学习方法可以学习比无监督学习方法更好的相似性度量，因为组织病理学图像之间的相似性可能因定义而异。然而，准备足够数量的标记数据可能是一个严重的问题，这将在后面描述。

在 CBIR 中，不仅需要准确性，还需要从大量图像中高速搜索相似图像。因此，各种图像特征降维技术（如主成分分析）和快速近似最近邻搜索（如 kd-tree 和哈希 [ [31](https://link.zhihu.com/?target=https%3A//www.sciencedirect.com/science/article/pii/S2001037017300867%3Fvia%253Dihub%23bb0160" \t "_blank) ]）被用于高速搜索。

**2.3 . 发现新的临床病理关系**

历史上，许多关于肿瘤、传染病等疾病的重要发现都是由病理学家和研究人员通过仔细、密切观察病理标本而取得的。例如，Hpylori是由一位病理学家在检查胃炎患者的[胃黏膜](https://link.zhihu.com/?target=https%3A//www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/stomach-mucosa" \t "_blank)时发现的。还尝试将癌症的形态特征与其临床行为相关联。例如，肿瘤分级对于规划治疗和确定患者对某些类型癌症的预后很重要，例如软组织肉瘤、原发性脑肿瘤、乳腺癌和前列腺癌。

同时，由于近年来医疗信息数字化的进步和基因组分析技术的进步，大量的数字信息如基因组信息、数字病理图像、MRI和CT图像已经变得可用。通过分析这些数据之间的关系，可以发现新的临床病理关系，例如癌症的[形态特征](https://link.zhihu.com/?target=https%3A//www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/morphological-trait" \t "_blank)与[体细胞突变](https://link.zhihu.com/?target=https%3A//www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/somatic-mutation" \t "_blank)之间的关系。然而，由于数据量巨大，病理学家和研究人员通过查看标本来手动分析所有关系是不现实的。这就是机器学习技术的用武之地。例如，Beck 等人。从乳腺癌的病理图像中提取纹理信息并用 L1 - 正则化逻辑回归分析，并表明基质的组织学与乳腺癌的预后相关 。其他研究包括癌症组织病理学图像预后预测、体细胞突变预测，以及基于图像QTL发现与自身免疫性甲状腺炎相关的新基因变异。

**特定于组织病理学图像分析的问题**

3.1 . 计算机辅助诊断

我们将描述病理图像分析的独特特征和治疗它们的计算方法。当使用深度学习对狗或房子等图像进行分类时，通常使用 256 × 256 像素等小尺寸图像作为输入。大尺寸的图像通常需要调整到足够小的尺寸才能进行足够的区分，因为输入图像尺寸的增加导致要估计的参数、所需的计算能力和内存的增加。相比之下，WSI 包含许多单元格，图像可能包含多达数百亿个像素，通常很难按原样进行分析。然而，将整个图像调整为较小的尺寸，例如 256 × 256 会导致细胞水平的信息丢失，从而导致识别精度显着下降。因此，整个 WSI 通常被划分为大约 256 × 256 像素的部分区域（“补丁”），并且每个补丁都是独立分析的，例如检测 ROI。由于计算能力和内存的进步，补丁大小正在增加（例如 960 × 960），这有望有助于提高准确性。整合每个补丁的结果的方法仍有改进的余地。例如，由于整个 WSI 可能包含数十万个补丁，因此即使对单个补丁进行了准确分类，也极有可能出现误报。一种可能的解决方案是对每个决策进行区域平均，这样只有当 ROI 扩展到多个补丁时，这些区域才被归类为 ROI。然而，这种方法可能会出现假阴性，导致遗漏小的 ROI，例如孤立的肿瘤细胞。由于计算能力和内存的进步，补丁大小正在增加（例如 960 × 960），这有望有助于提高准确性。整合每个补丁的结果的方法仍有改进的余地。例如，由于整个 WSI 可能包含数十万个补丁，因此即使对单个补丁进行了准确分类，也极有可能出现误报。一种可能的解决方案是对每个决策进行区域平均，这样只有当 ROI 扩展到多个补丁时，这些区域才被归类为 ROI。然而，这种方法可能会出现假阴性，导致遗漏小的 ROI，例如孤立的肿瘤细胞。由于计算能力和内存的进步，补丁大小正在增加（例如 960 × 960），这有望有助于提高准确性。整合每个补丁的结果的方法仍有改进的余地。例如，由于整个 WSI 可能包含数十万个补丁，因此即使对单个补丁进行了准确分类，也极有可能出现误报。一种可能的解决方案是对每个决策进行区域平均，这样只有当 ROI 扩展到多个补丁时，这些区域才被归类为 ROI。然而，这种方法可能会出现假阴性，导致遗漏小的 ROI，例如孤立的肿瘤细胞。整合每个补丁的结果的方法仍有改进的余地。例如，由于整个 WSI 可能包含数十万个补丁，因此即使对单个补丁进行了准确分类，也极有可能出现误报。一种可能的解决方案是对每个决策进行区域平均，这样只有当 ROI 扩展到多个补丁时，这些区域才被归类为 ROI。然而，这种方法可能会出现假阴性，导致遗漏小的 ROI，例如孤立的肿瘤细胞。整合每个补丁的结果的方法仍有改进的余地。例如，由于整个 WSI 可能包含数十万个补丁，因此即使对单个补丁进行了准确分类，也极有可能出现误报。一种可能的解决方案是对每个决策进行区域平均，这样只有当 ROI 扩展到多个补丁时，这些区域才被归类为 ROI。然而，这种方法可能会出现假阴性，导致遗漏小的 ROI，例如孤立的肿瘤细胞。一种可能的解决方案是对每个决策进行区域平均，这样只有当 ROI 扩展到多个补丁时，这些区域才被归类为 ROI。然而，这种方法可能会出现假阴性，导致遗漏小的 ROI，例如孤立的肿瘤细胞。一种可能的解决方案是对每个决策进行区域平均，这样只有当 ROI 扩展到多个补丁时，这些区域才被归类为 ROI。然而，这种方法可能会出现假阴性，导致遗漏小的 ROI，例如孤立的肿瘤细胞。在一些应用中，例如 IHC 评分、标本或患者淋巴结转移的分期，以及通过一张载玻片内多个区域的 Glisson 评分诊断的前列腺癌分期，需要更复杂的算法来整合补丁级别或对象级别的决策。[14](https://link.zhihu.com/?target=https%3A//www.sciencedirect.com/science/article/pii/S2001037017300867%3Fvia%253Dihub%23bb0070" \t "_blank)，[17](https://link.zhihu.com/?target=https%3A//www.sciencedirect.com/science/article/pii/S2001037017300867%3Fvia%253Dihub%23bb0085" \t "_blank)，[18](https://link.zhihu.com/?target=https%3A//www.sciencedirect.com/science/article/pii/S2001037017300867%3Fvia%253Dihub%23bb0090" \t "_blank)，[39](https://link.zhihu.com/?target=https%3A//www.sciencedirect.com/science/article/pii/S2001037017300867%3Fvia%253Dihub%23bb0200" \t "_blank)，[40](https://link.zhihu.com/?target=https%3A//www.sciencedirect.com/science/article/pii/S2001037017300867%3Fvia%253Dihub%23bb0205" \t "_blank)，[75](https://link.zhihu.com/?target=https%3A//www.sciencedirect.com/science/article/pii/S2001037017300867%3Fvia%253Dihub%23bb0380" \t "_blank) ]。例如，对于转移性乳腺癌的 pN 分期，这是 Camelyon 17 中的任务之一，包括我们在内的多个参与团队应用了像素或补丁级概率的随机森林分类器，该分类器使用各种特征（例如估计的肿瘤大小）通过深度学习估计。

**3.2.1标记图像不足**

使用机器学习进行病理图像分析的最大问题可能是只有少量的训练数据可用。深度学习在一般图像识别任务中取得成功的一个关键是训练数据极其丰富。尽管在[数字病理学的](https://link.zhihu.com/?target=https%3A//www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/digital-pathology" \t "_blank)大多数任务中都需要补丁级别或像素级别（例如癌变区域的内部/外部边界）的标签信息如诊断，WSI 的大多数标签最多是在案例级别（例如诊断）。一般图像分析中的标签信息可以很容易地从互联网上检索，也可以使用众包标签，因为任何人都可以识别对象并执行标记工作。然而，只有病理学家才能准确地对病理图像进行标注，而在巨大的WSI中进行区域层面的标注需要大量的劳动力。

可以在机器学习中重用公共准备分析数据作为训练数据，例如自然图像中的ImageNet 和皮肤宏观诊断中的国际皮肤成像协作 。在数字病理学领域，有一些公共数据集包含手工注释的组织病理学图像，如[表 2](https://link.zhihu.com/?target=https%3A//www.sciencedirect.com/science/article/pii/S2001037017300867%3Fvia%253Dihub%23t0010" \t "_blank)、[表 3 所示](https://link.zhihu.com/?target=https%3A//www.sciencedirect.com/science/article/pii/S2001037017300867%3Fvia%253Dihub%23t0015" \t "_blank). 如果分析的目的、载玻片条件（例如染色）和图像条件（例如放大级别和图像分辨率）相似，它们可能会很有用。然而，因为这些数据集中的每一个都专注于特定的疾病或细胞类型，所以这些数据集没有涵盖许多任务。也有含有高分辨率峰会几次大规模的组织图象数据库：癌症基因组图谱（TCGA）包含来自不同癌症类型10000峰会，和基因型与组织中的表达（GTEx包含来自各种组织的 20,000 多个WSI。这些数据库可以作为各种任务的潜在训练数据。此外，TCGA 和 GTEx 还提供基因组图谱，可用于研究基因型和形态之间的关系。问题在于这些存储库中的 WSI 包含案例级别的标签，为了能够将它们用于训练数据，需要一些预处理或专门的机器学习算法来处理案例级别的标签。

Table 2. Downloadable WSI database.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset or author's name | # slides or patches | Stain | Disease | Additional data |
| TCGA | 18,462 | H&E | Cancer | Genome/transcriptome/epigenome |
| GTEx | 25,380 | H&E | Normal | Transcriptome |
| TMAD | 3726 | H&E/IHC |  | IHC score |
| TUPAC16 | 821 from TCGA | H&E | Breast cancer | Proliferation score for 500 WSIs, position for mitosis for 73 WSIs, ROI for 148 cases |
| Camelyon17 | 1000 | H&E | Breast cancer (lymph node metastasis) | Mask for cancer region (in 500 WSIs with 5 WSIs per patient) |
| Köbel et al. | 80 | H&E | Ovarian carcinoma |  |
| KIMIA Path24 | 24 | H&E/IHC and others | various tissue |  |

Table 3. Hand annotated histopathological images publicly available.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dataset or paper | Image size (px) | # images | Stain | Disease | Additional data | Potential usage |
| KIMIA960 | 308 × 168 | 960 | H&E/IHC | various tissue |  | Disease classification |
| Bio-segmentation | 896 × 768, 768 × 512 | 58 | H&E | Breast cancer |  | Disease classification |
| Bioimaging challenge 2015 | 2040 × 1536 | 269 | H&E | Breast cancer |  | Disease classification |
| GlaS | 574–775 × 430–522 | 165 | H&E | Colorectal cancer | Mask for gland area | Gland segmentation |
| BreakHis | 700 × 460 | 7909 | H&E | Breast cancer |  | Disease classification |
| Jakob Nikolas et al. | 1000 × 1000 | 100 | IHC | Colorectal cancer | Blood vessel count | Blood vessel detection |
| MITOS-ATYPIA-14 | 1539 × 1376, 1663 × 1485 | 4240 | H&E | Breast cancer | Coordinates of mitosis with a confidence degree/six criteria to evaluate nuclear atypia | Mitosis detection, nuclear atypia classification |
| Kumar et al. | 1000 × 1000 | 30 | H&E | Various cancer | Coordinates of annotated nuclear boundaries | Nuclear segmentation |
| MITOS 2012 | 2084 × 2084, 2252 × 2250 | 100 | H&E | Breast cancer | Coordinates of mitosis | Mitosis detection |
| Janowczyk et al. | 1388 × 1040 | 374 | H&E | Lymphoma | None | Disease classification |
| Janowczyk et al. | 2000 × 2000 | 311 | H&E | Breast cancer | Coordinates of mitosis | Mitosis detection |
| Janowczyk et al. | 100 × 100 | 100 | H&E | Breast cancer | Coordinates of lymphocyte | Lymphocyte detection |
| Janowczyk et al. | 1000 × 1000 | 42 | H&E | Breast cancer | Mask for epithelium | Epithelium segmentation |
| Janowczyk et al. | 2000 × 2000 | 143 | H&E | Breast cancer | Mask for nuclei | Nuclear segmentation |
| Janowczyk et al. | 775 × 522 | 85 | H&E | Colorectal cancer | Mask for gland area | Gland segmentation |
| Janowczyk et al. | 50 × 50 | 277,524 | H&E | Breast cancer | None | Tumor detection |
| Gertych et al. | 1200 × 1200 | 210 | H&E | Prostate cancer | Mask for gland area | Gland segmentation |
| Ma et al. | 1040 × 1392 | 81 | IHC | Breast cancer |  | TIL analysis |
| Linder et al. | 93–2372 × 94–2373 | 1377 | IHC | Colorectal cancer | Mask for epithelium and stroma | Segmentation of epithelium and stroma |
| Xu et al. | Various size | 717 | H&E | Colon cancer |  |  |
| Xu et al. | 1280 × 800 | 300 | H&E | Colon cancer | Mask for colon cancer | Segmentation |

增加训练数据的一种方法是减少病理学家在 WSI 中指定 ROI 的工作时间。易于使用的 GUI 工具可帮助病理学家在更短的时间内有效地标记更多样本 。例如，Cytomine 不仅允许病理学家用椭圆、矩形、多边形或手绘图包围 WSI 中的 ROI，而且还应用基于内容的图像检索算法来加速注释。另一个减少工作时间的有趣想法是在诊断期间自动定位 ROI，它通过跟踪病理学家的行为将诊断的通常工作时间用作标记。这种方法跟踪病理学家的眼球运动 、鼠标光标位置并更改视口 。然而，从这些跟踪数据中准确定位 ROI 并不总是那么容易，因为病理学家并不总是花时间查看 ROI，而且通过这些方法获得的边界信息往往不太清楚。

另一种利用机器学习方法的方法是主动学习当标签数据（即病理图像）的获取成本较大时，这通常是有效的。主动学习是监督学习中使用的一种方法，它自动选择最有价值的未标记样本（即当正确标记并用作训练数据时有望提高分类器识别性能的样本）并将其显示给病理学家标记. 由于这种方法可能会在标记图像数量较少的情况下提高识别性能，因此获得相同识别性能的总标记时间将缩短。许多标准，例如不确定性抽样、委员会查询方差减少和假设空间缩减 已被应用于选择有价值的未标记样本。

**3.2.2 . 包含不足的标签**

即使不知道 WSI 中 ROI 的确切位置，也有可能从分配给 WSI 或 WSI 级别标签的病理诊断中获得有关 WSI 中 ROI 存在/不存在的信息。即使 WSI 没有进一步的信息，与补丁级标签相比，这些所谓的弱标签也很容易获得，在这方面，WSI 被视为机器学习设置中由许多补丁（实例）制成的“袋子”。在诊断癌症时，如果至少一个贴片包含癌组织，则 WSI 被标记为癌症，如果没有一个贴片包含癌组织，则 WSI 被标记为正常。这种设置是多实例学习或弱监督学习 。在典型的多实例学习问题中，正袋至少包含一个正实例，负袋不包含任何正实例。多实例学习的目的是根据只包含袋子标签的训练数据来预测袋子或实例标签。多实例学习中的各种方法已应用于组织病理学图像分析，包括基于增强的方法、基于[支持向量](https://link.zhihu.com/?target=https%3A//www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/support-vector-machine" \t "_blank)机的方法和基于深度学习的方法。

相比之下，半监督学习使用标记数据和未标记数据。未标记数据用于估计标记数据的真实分布。例如，如图[1](https://link.zhihu.com/?target=https%3A//www.sciencedirect.com/science/article/pii/S2001037017300867%3Fvia%253Dihub%23f0005" \t "_blank)所示，只考虑标记样本的决策边界会形成一条垂直线，但同时考虑标记和未标记样本的决策边界会形成一条斜线，这样会更准确。由于半监督学习在同一类中的样本形成具有良好判别性的集群时被认为特别有效，因此相对容易的问题可能是一个很好的目标。

**3.2.3 . 重用来自另一个任务的参数**

使用太少的训练数据执行监督学习只会导致泛化性能不足。尤其是在深度学习中尤其如此，其中要学习的参数数量非常大。在这种情况下，不是从头开始学习整个模型，学习通常从使用在另一个类似任务中优化的预训练模型的（部分）参数开始。这种学习方法称为[迁移学习](https://link.zhihu.com/?target=https%3A//www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/transfer-of-learning" \t "_blank)。在 CNN 中，最后（通常为三个）全连接层之前的层被视为特征提取器。全连接层经常被适合目标任务的新网络所取代。早期层中的参数可以按原样使用 ，或者作为初始参数，然后该网络部分或完全从目标任务[训练数据了解到（所谓的微调）。在病理图像中，没有从使用其他病理图像的任务中学习到的网络可用，因此经常使用使用 ImageNet 学习的网络，ImageNet 是一个包含大量一般图像的数据库 。例如，Xu 等人使用从在 ImageNet 上训练的 CNN 提取的特征对大脑和结肠病理图像执行分类和分割任务，并取得了最先进的性能 。尽管病理图像本身看起来与一般图像（例如猫和狗）非常不同，但它们具有共同的基本图像结构，例如线和弧。由于深度学习的早期层捕获了这些基本的图像结构，因此这种使用通用图像的预训练模型在组织病理学图像分析中效果很好。尽管如此，如果在足够数量的不同组织病理图像上预训练的模型可用，它们可能会胜过 ImageNet 预训练模型。

**3.3 . 不同级别的放大导致不同级别的信息**

组织通常由细胞组成，不同的组织表现出不同的细胞特征。在高倍场显微图像中可以很好地捕捉有关细胞形状的信息，但在低倍场中可以更好地捕捉结构信息，例如由许多细胞组成的腺体结构。由于癌组织既有细胞异型性，也有结构异型性，因此在多个放大倍数下拍摄的图像均会包含重要信息。病理学家通过改变显微镜的放大倍数来获取从细胞水平到组织水平的不同类型的信息来诊断疾病。即使在机器学习中，也存在利用不同放大倍数的图像的研究。如上所述，很难直接以原始分辨率处理图像，图像通常会调整大小以对应各种放大倍数并用作分析的输入。关于诊断，信息量最大的放大倍数仍然存在争议，但有时通过同时输入高倍率和低倍率图像来提高准确性，这可能取决于疾病和组织的类型以及机器学习算法。

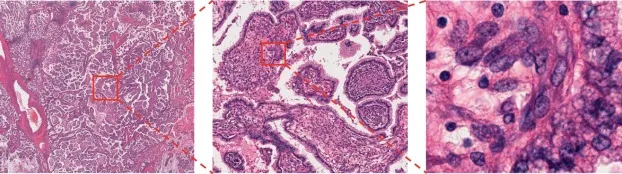


图2。同一组织病理学图像的多个放大级别。右图显示了左图红框指示的放大区域。最左边的图像清楚地显示了乳头状结构，最右边的图像清楚地显示了每个细胞的细胞核。组织病理学图像取自 TCGA。

**3.4 . WSI 作为无序纹理类图像**

病理图像在本质上不同于猫和狗，从某种意义上说，它显示出最小成分（通常是细胞）的重复模式。因此，它比物体更接近纹理。CNN 通过池化操作在一定程度上获得了平移不变性。此外，即使是普通的 CNN 也可以通过以小步幅移动组织图像来通过数据增强来学习类纹理结构。同时，一直存在的方法，其利用纹理结构更密集，如灰度共生矩阵[ [110](https://link.zhihu.com/?target=https%3A//www.sciencedirect.com/science/article/pii/S2001037017300867%3Fvia%253Dihub%23bb0560" \t "_blank) ]，局部二元模式使用CNN ，Gabor滤波器组，以及最近开发的深纹理的表示[64](https://link.zhihu.com/?target=https%3A//www.sciencedirect.com/science/article/pii/S2001037017300867%3Fvia%253Dihub%23bb0325" \t "_blank)，[112](https://link.zhihu.com/?target=https%3A//www.sciencedirect.com/science/article/pii/S2001037017300867%3Fvia%253Dihub%23bb0570" \t "_blank)]。使用 CNN 层中特征图的相关矩阵计算深层纹理表示。将 CNN 特征转换为纹理表示将导致获得关于单元位置的不变性，同时利用 CNN 学习的良好表示。深度纹理表示的另一个优点是对输入图像的大小没有限制，非常适合 WSI 的大图像尺寸。纹理和非纹理之间的界限不清楚，但单个单元格或单个结构显然不是纹理。因此，更好的方法将取决于要分析的对象。

**3.5 . 颜色变化和伪影**

WSI 是通过多个过程创建的：将病理标本切片并放置在载玻片上，用苏木精和[伊红染色](https://link.zhihu.com/?target=https%3A//www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/eosin" \t "_blank)，然后进行扫描。在每一步都可能引入与潜在生物因素无关的不良影响。例如，当组织切片被放置在载玻片上时，它们可能会弯曲和起皱；扫描过程中灰尘可能会污染载玻片；由于组织切片厚度不同，可能会出现模糊（[图3](https://link.zhihu.com/?target=https%3A//www.sciencedirect.com/science/article/pii/S2001037017300867%3Fvia%253Dihub%23f0015" \t "_blank)）；有时组织区域用颜色标记标记。由于这些伪影可能会对解释产生不利影响，因此需要使用特定算法来检测模糊和组织褶皱提出。此类算法可用于预处理 WSI。

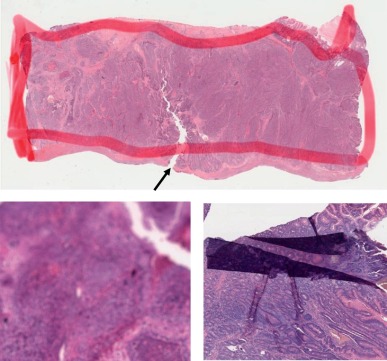


图3WSI中的工件。上：肿瘤区域用红色标记勾勒。箭头表示可能在组织制备过程中形成的撕裂。左下角：模糊的图像。右下：折叠的组织切片。组织病理学图像取自 TCGA 。

另一个严重的伪影是颜色变化，如图[4](https://link.zhihu.com/?target=https%3A//www.sciencedirect.com/science/article/pii/S2001037017300867%3Fvia%253Dihub%23f0020" \t "_blank)所示。变异来源包括染色试剂的不同批次或制造商、组织切片的厚度、染色条件和扫描仪型号。不考虑颜色变化的学习可能会降低机器学习算法的性能。如果可以将每台扫描仪获取的每一个染色组织的足够数据纳入其中，颜色变化对分类精度的影响就可以忽略不计；然而，目前这似乎不太可能。

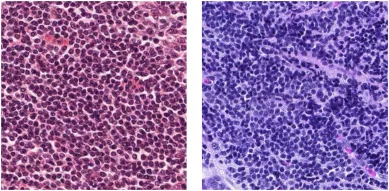


图 4组织病理学图像的颜色变化。这两个图像都显示了淋巴细胞。组织病理学图像取自 TCGA 。

为了解决这个问题，各种方法已经被提出迄今包括转换灰度，颜色正常化，色彩增强。转换为灰度是最简单的方法，但它忽略了有关病理学家常规使用的颜色表示的重要信息。相比之下，颜色归一化尝试逐个像素地调整图像的颜色值，以便源图像的颜色分布与参考图像的颜色分布相匹配。然而，由于目标图像和参考图像中细胞或组织的成分和组成比例通常不同，因此通常需要使用专用算法进行核检测等预处理来调整成分。因此，当任务中分析的 WSI 至少部分包含类似的细胞或组织成分时，颜色标准化似乎是合适的。

另一方面，颜色增强是一种通过应用随机色调、饱和度、亮度和对比度来执行的数据增强。色彩增强的优势在于无论被分析的对象如何，都可以轻松实现。颜色增强似乎适用于颜色变化较小的 WSI，因为颜色增强中过度的颜色变化可能导致最终分类器中颜色信息的丢失。由于颜色归一化和颜色增强可能是互补的，两种方法的结合可能会更好。

**4. 总结**

数字组织病理学图像识别是一个非常适合机器学习的问题，因为图像本身包含足以诊断的信息。在这篇综述中，我们提出了使用机器学习进行数字组织病理学图像分析的问题。经过目前的努力，这些问题正在逐渐得到解决，但仍有改进的余地。一旦大量注释良好的 WSI 可用，这些问题中的大多数都可能得到解决。收集来自不同机构的 WSI 以相同的标准对其进行协作注释并将这些数据公开将足以促进更复杂的数字组织病理学图像分析的发展。

最后，我们提出了一些目前尚未得到充分研究的潜在未来研究课题。