病理图像识别

1. 背景
2. 数字病理图像

组织病理学切片是癌症诊断的金标准。病理医师在病人体内活体取出组织，进行切片、染色制作成切片，于显微镜下观察病变情况。随着数字病理扫描仪的研发，现有仪器能自动控制显微镜对病理切片对焦成像，并通过逐行扫描生成一张将病理切片放大40倍的数字图像（WSI）。数字病理图像可以让医生在电脑上观察病理组织，观察过程中更自如地调节放大倍数和切换观察视野。

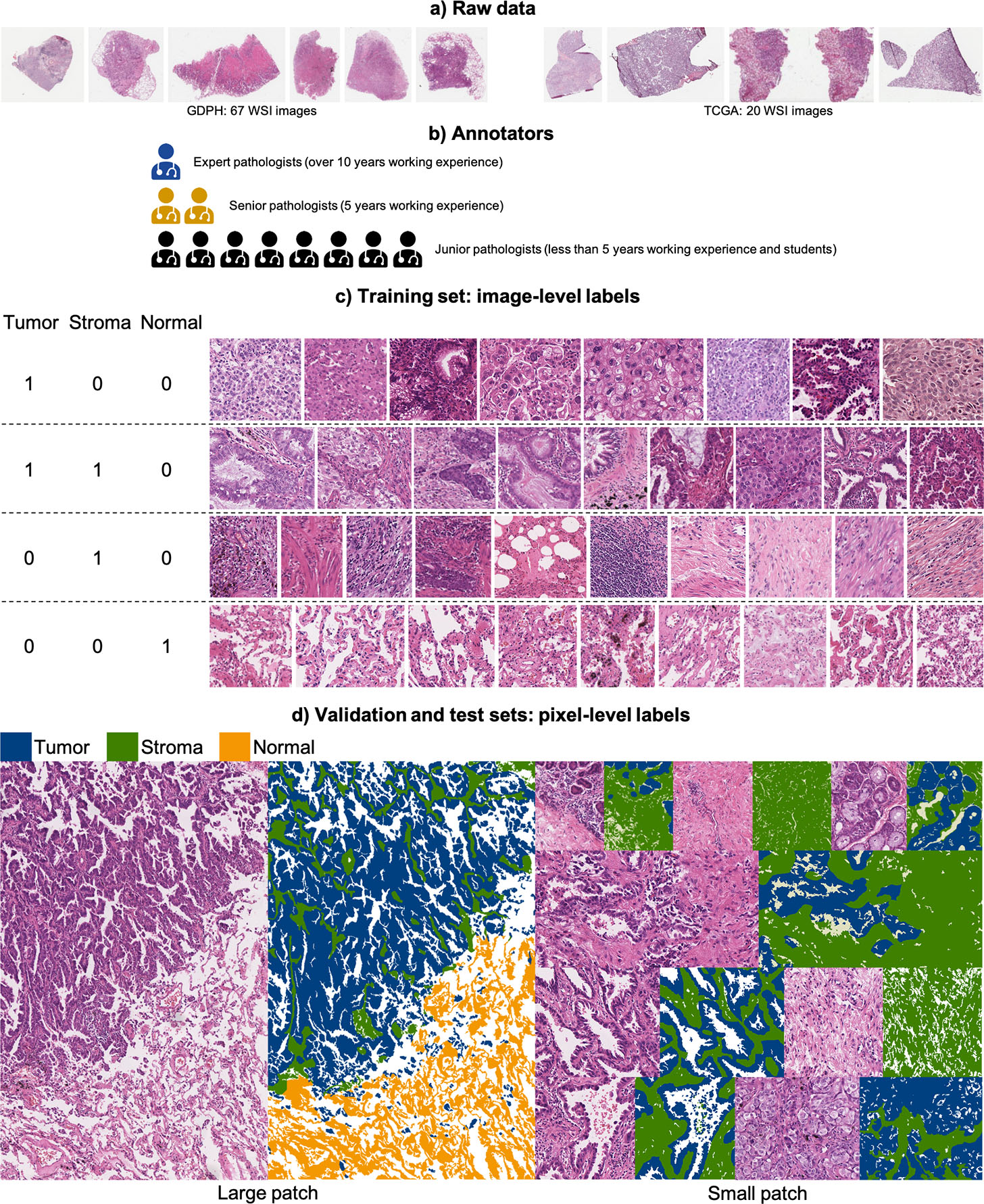


图1:数字病理图像

一张病理图像提供了大量关于肿瘤微环境（TME）的信息，一方面可以解释肿瘤的发生和发展，另一方面病理图像的诊断结果直接影响着癌症患者的治疗方案和预后措施。目前已有大量基于数字病理图像数据集的研究，探索了包括疾病分类、组织分割、染色体检测等多种任务。因此，为了进一步的临床研究，迫切需要对不同组织进行分割和区分。

1. 数据集简介（来源：<https://wsss4luad.grand-challenge.org/WSSS4LUAD/>）

肺癌是全世界癌症死亡的主要原因。在这项挑战中，共有广东省人民医院（GDPH）的67张染色切片和癌症基因组图谱（TCGA）中的20张染色切片。参赛者只得到图像级别的注释：3位数的单次编码代表肿瘤上皮组织、肿瘤基质组织和正常组织。如果图像标签为110，则表示同时含有第一、二类组织，不含第三类组织。这些标签是基于从WSI上裁剪的小图（patches）。

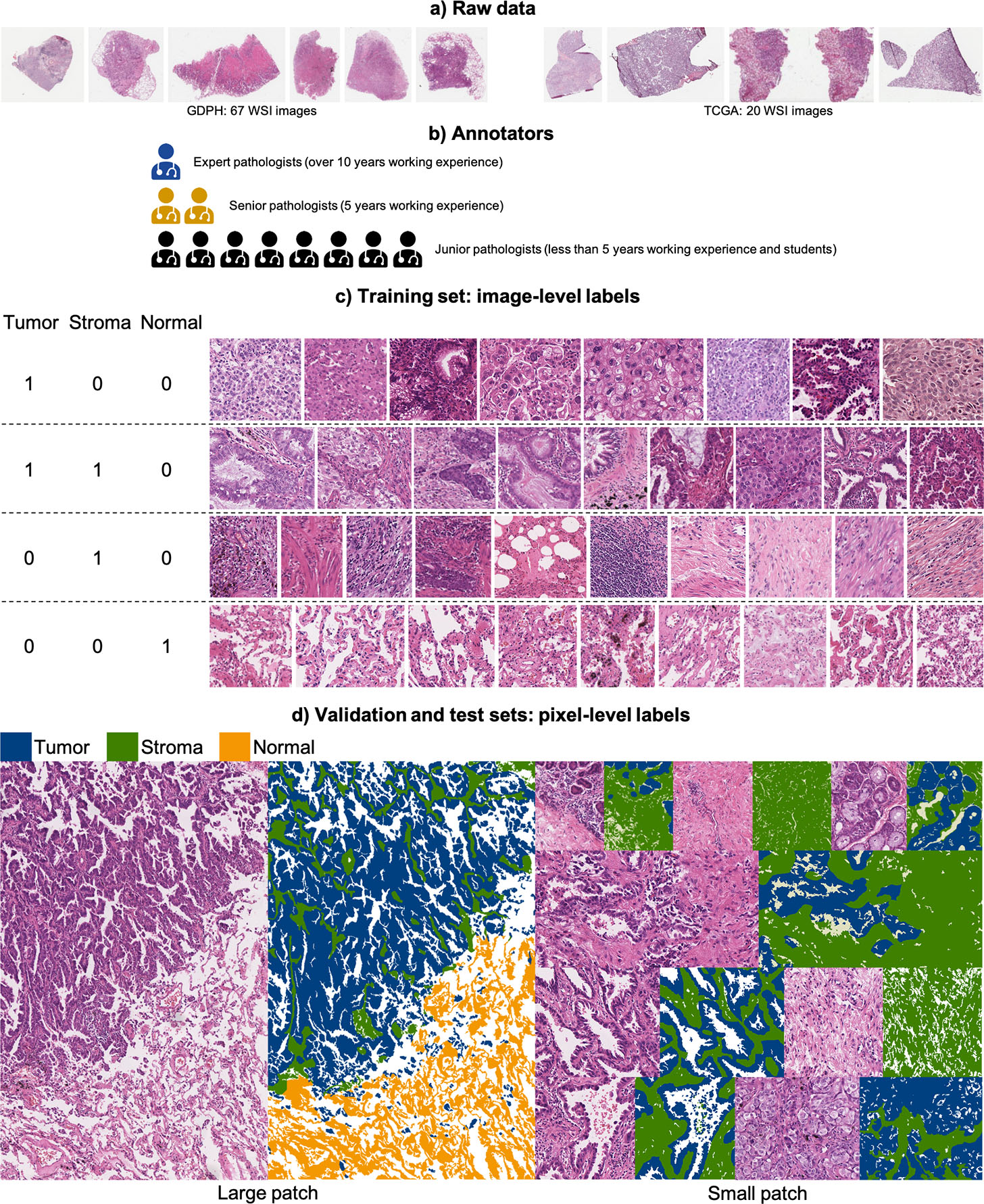


图2：训练集数据图像及标签示意

训练集中总共裁剪了 10091 个小图（patches）。其中每个图像文档的类别标示在图像文件的名称中，如'patient\_ID'+'x\_axis'+'y\_axis'+'3位one hot标签'.png’。训练集中的每个图像（patch）都是从 WSI 图像中的随机锚点 (x, y) 裁剪而来，具有随机的高度 (150~300) 和宽度 (150~300)，其中 (x, y) 是 补丁的左上角。3位one hot标签为3位零一编码，在第几位上的数字为1则代表区域内包含这一类组织（【肿瘤、基质、正常】）。数据集的统计数目如下表。

|  |  |  |  |
| --- | --- | --- | --- |
| 组织 | 肿瘤 | 基质 | 正常 |
| 数量 | 6579 | 7076 | 1832 |

该挑战同时提供了具有像素级别标注的验证集和测试集以供语义分割任务的测试。验证集：来自 GDPH 的 9 个 WSI 和来自 TCGA 的 3 个 WS。在验证集中裁剪了总共 40 个patches， 包括9个尺寸大约1500~5000\*1500~5000的和31个尺寸约200~500\*200~500的。测试集：来自 GDPH 的 9 个 WSI 和来自 TCGA 的 3 个 WSI。在测试集中裁剪了总共 80 个patches，包括14个尺寸约1500~5000\*1500~5000的和66个尺寸约200~500\*200~500的。

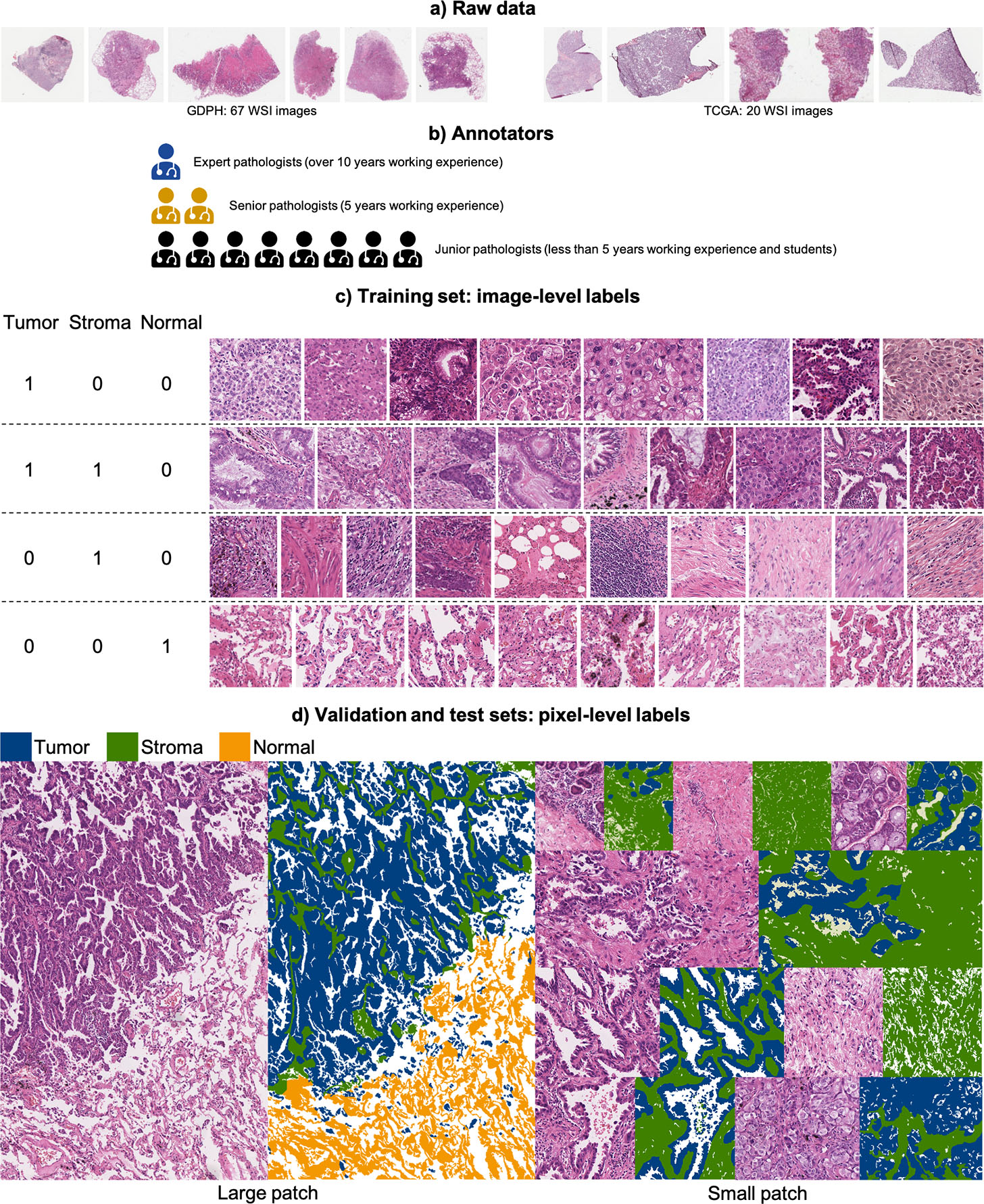


图2：测试集和验证集数据图像及标签示意

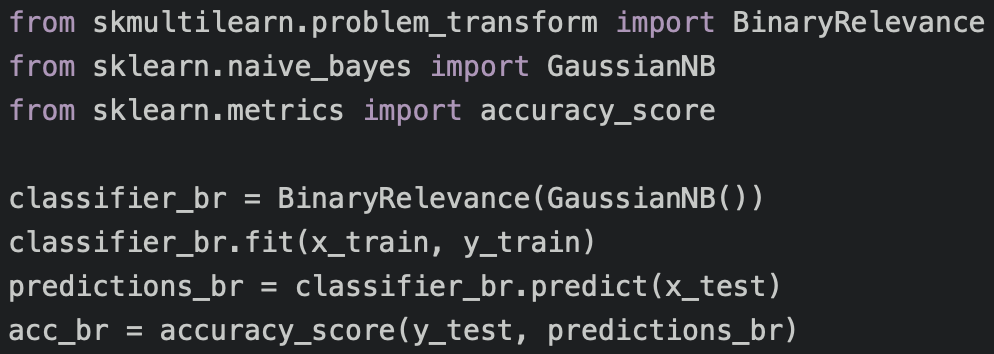
1. 任务
2. 基础任务——图像分类

将该挑战的训练集作为可用的新数据集，将该数据集再次划分为训练集、验证集和测试集，对该数据集进行图像分类任务。

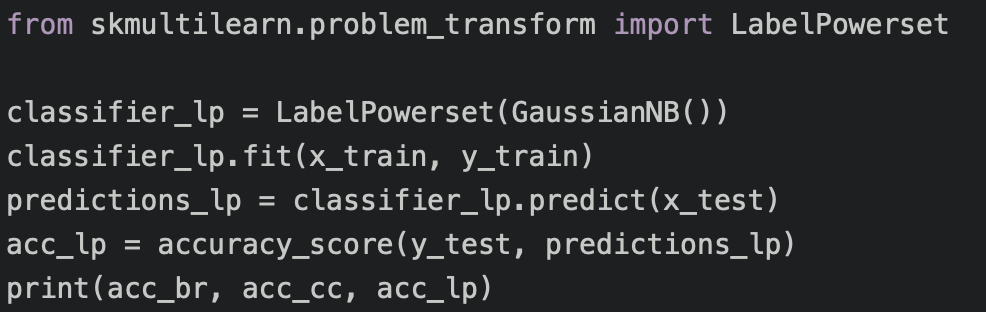
1. 进阶任务——像素级预测

受计算机视觉中弱监督语义分割（WSSS）的启发，考虑从图像级注释学习并最终能进行组织语义分割。用该挑战的验证集和测试集测试模型的表现，注意衡量指标为mIoU。

1. 线索
2. 基于sklearn 库（<https://scikit-learn.org/stable/>）的多标签分类
3. 针对每一个标签训练一个二分类器。



1. 把一个样本的标签集作为一个整体，即当做一个类别标签。这样原本的三类将变成000，001，010，011，100，101，110，111共八类。



1. 基于卷积神经网络的多标签分类

对多标签的处理方法和3.1相同，但是没有可直接用的库和函数，需要自己对数据集进行预处理。（从图像名提取分类标签、将图像划分到不同的数据集、设计被卷积神经网络引用的方式）对最普通和基础的CNN模型在最后加上一层全连接层，在计算概率的时候用sigmod将每一个标签当做一个二分类问题，loss函数则为cross\_entropy。

Pytorch平台编写CNN可参考<https://github.com/fendouai/PyTorchDocs>。

Keras平台编写CNN可参考<https://github.com/skyduy/CNN_keras>。