Mod Pathol 36 (2023) 100086

MODERN PATHOLOGY

USCAP UNITED STATES AND CANADIAN ACADEMY OF PATHOLOGY
Creating a Better Pathologist

Journal homepage: https://modernpathology.org/

Research Article

Annotating for Artificial Intelligence Applications in Digital Pathology: A Practical Guide for Pathologists and Researchers

Diana Montezuma[a,b,c,*], Sara P. Oliveira[d,e], Pedro C. Neto[d,e], Domingos Oliveira[a], Ana Monteiro[a], Jaime S. Cardoso[d,e], Isabel Macedo-Pinto[a]

[a] *IMP Diagnostics, Porto, Portugal;* [b] *Cancer Biology and Epigenetics Group, Research Center of IPO Porto (CI-IPOP)/RISE@CI-IPOP (Health Research Network), Portuguese Oncology Institute of Porto (IPO Porto)/Porto Comprehensive Cancer Center (Porto.CCC), Porto, Portugal;* [c] *Institute of Biomedical Sciences Abel Salazar (ICBAS), University of Porto, Porto, Portugal;* [d] *Telecommunications and Multimedia Unit, Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), Porto, Portugal;* [e] *Faculty of Engineering, University of Porto (FEUP), Porto, Portugal*

ARTICLE INFO

Article history:
Received 26 October 2022
Revised 24 November 2022
Accepted 14 December 2022
Available online 11 January 2023

Keywords:
annotation
artificial intelligence
computational pathology
digital pathology

ABSTRACT

Training machine learning models for artificial intelligence (AI) applications in pathology often requires extensive annotation by human experts, but there is little guidance on the subject. In this work, we aimed to describe our experience and provide a simple, useful, and practical guide addressing annotation strategies for AI development in computational pathology. Annotation methodology will vary significantly depending on the specific study's objectives, but common difficulties will be present across different settings. We summarize key aspects and issue guiding principles regarding team interaction, ground-truth quality assessment, different annotation types, and available software and hardware options and address common difficulties while annotating. This guide was specifically designed for pathology annotation, intending to help pathologists, other researchers, and AI developers with this process.

Introduction

The computational pathology field is rapidly evolving, mainly because of the swift development of artificial intelligence (AI) technology and its application to digital pathology (DP) in recent years. AI models applied to pathology images can be used in a wide range of tasks, from lesion detection and classification,[1,2] object counting,[3] immunohistochemical stain scoring[4] to even predicting patient prognosis and response to therapy (reviewed in Echle et al[5]). A requirement for successful machine learning (ML) development is access to large volumes of annotated/labeled data.[6,7] Therefore, annotation has become key. Only high-quality annotations will support the development of robust algorithms because "garbage in means garbage out" when dealing with data. Notwithstanding, access and availability of such data are limited, which represents a barrier to wider development and adoption.[7] Furthermore, even though the need for annotation guidelines has already been pointed out[8] because this field of research is relatively new, there is a general lack of available information and specific guidance regarding annotation in computational pathology. Only recently, recommendations were issued specifically for semantic annotation in pathology by the PathLake consortium.[9] In this study, in line with their objective of improving the interoperability of annotation tasks and widening the discussion on stringent annotation protocols for computational pathology,[9] we describe the essential steps to address when creating an

* Corresponding author.
E-mail address: diana.felizardo@impdiagnostics.com (D. Montezuma).

https://doi.org/10.1016/j.modpat.2022.100086

annotated pathology data set. Our contribution is mostly based on lessons learned during the development of different AI algorithms within our own projects, encompassing both the positive outturns and the challenges encountered. In this study, we aimed to present a simple and especially useful guide to help pathologists, and other researchers, initiate their annotation efforts.

Material and Methods

This study was developed as part of CADpath.AI, a nationally funded project primarily aimed at developing computer-aided diagnostic (CAD) solutions capable of diagnosing oncologic pathologies, particularly colorectal cancer and cervical cancer, through the automatic analysis of histologic samples and explaining the decision of diagnosis, fundamental for clinical acceptance.[10,11] Moreover, CADpath.AI focused on creating AI tools to facilitate the laboratory's technical checkpoints, namely a model to evaluate the number of fragments present on a whole-slide image (WSI), automatically ensuring that the material available for further analysis is consistent with that recorded during gross preparation.[12] Within our project, we have annotated/labeled more than 10,000 colorectal WSIs (for classification tasks), 2000 cervical cancer WSIs (segmentation and classification tasks), 250 breast cancer slides (immunoscore prediction task), and around 2000 variate WSIs for the development of a quality checkpoint tool (part of this work has already been published[10-12] or presented in international conferences, and some are ongoing unpublished work).

For the purposes of this guide, we performed a thorough PubMed search to identify relevant publications regarding annotation tasks in pathology. This work aimed to be a simple and helpful guide concerning annotation efforts, originating from our own experience (from both the positive outcomes and the encountered difficulties).

Results

Defining Annotation in Pathology

Annotation in pathology can have different meanings, from exhaustive drawing and delineation within a WSI up to assigning a single classification/category to the entire slide (weak annotation or labeling). The type of annotation will depend greatly on the purpose of the algorithm being developed and the type of the ML task. Most recent AI solutions for pathology rely on deep learning models, namely artificial neural networks.[7] When defining the annotation type, it is important to decide whether the WSIs will need to be fully annotated (requiring to delineate regions within the slide) or only labeled (having an identifier for the entire WSI). For weakly supervised solutions, slide-level labeling will suffice, but larger amounts of data will be necessary,[13,14] whereas more conventional supervised techniques will require a more extensive annotation strategy.[15]

In general, the annotation will depend on the type of algorithm we are developing. Semantic segmentation is used when we intend to delineate a tumor area, for example, or quantify it. In semantic segmentation, each pixel in the image is labeled.[16,17] Annotators can use a polygon line to trace the outline of each relevant class. For example, an image of a colonic wall could be separated into the classes "mucosa," "submucosa," "muscularis propria" and "adipose tissue," or a breast biopsy could be divided in "tumor" and "normal breast tissue" categories. This type of annotation can be used, for example, when designing an algorithm to detect and/or segment tumoral tissue within a slide, for example, a classifier to identify and subtype breast cancer.[18] Object detection allows the identification and counting of "objects" of interest[17]: a classic example in pathology would be an algorithm to count mitoses. This type of algorithm could alleviate manual counting, conventionally visually assessed, when grading breast cancer or neuroendocrine tumors, for instance. This task has been the object of various challenges in DP, namely the MIDOG challenge, where multiple research groups compete to develop the best performing mitoses detection algorithm.[19] Instance segmentation is when the task deals with detecting instances of objects and demarcating their boundaries: in other words, when we intend to count and delineate. Semantic segmentation treats the multiple objects of the same class as a single entity, whereas instance segmentation treats the various objects of the same class as distinct individual instances.[16] One example would be to segment and classify different nuclei types on histological slides.[20] This could be the annotation approach used to identify and quantify tumor-infiltrating lymphocytes, distinguishing them from other cells, a task with potential prognostic effect[21] (Fig. 1).

Another way of defining annotations is regarding the level of detail: case-/slide-level annotation/labeling (a diagnosis per slide/case) or more detailed, region-level annotation (marking up the areas of interest within the WSIs) or cell-level annotation (identifying the cells/nuclei of interest).[9] Annotations can be made in a pixel-level manner (exhaustive annotation strategy, in which each pixel on the image is allocated to a nonoverlapping class)[22] or the process can be more focused, making annotations in only specific areas of interest of the slide (partially annotating the slides can even be more informative, in a sense that more WSIs can be annotated, with the same effort, which would provide a more diverse training set). Furthermore, it is possible to extrapolate the results of a partial annotation for the full extent of the slide. Some existing annotation software already integrates algorithms to allow for a semi-automatic annotation strategy.[23] One example is the Aiforia software, which has an "annotation assistant" that does an automatic entire slide analysis after an initial annotation effort by the expert, indicating where to annotate further (focusing on low confidence regions identified by preexistent model) to enhance the algorithm's accuracy. Another proprietary software that facilitates and assists the annotation effort is DeePathology Studio, which works as an interactive "do it yourself" platform for AI in pathology.[24] The software is built so researchers/pathologists can easily create AI solutions, without needing programming experience because it uses built-in algorithms. Moreover, regarding open-source tools, QuickAnnotator allows for fast image annotation.[25] It is designed to improve annotation efficiency; the user annotates regions of interest (ROI) viausing a web interface, whereas a deep learning model is concurrently optimized using these annotations and applied to the ROI.[25] The user iteratively reviews the results to either accept correctly annotated regions or reject erroneously segmented structures, improving subsequent model suggestions and mitigating the human annotation effort.[25] When developing AI tools, the main issue with these preexistent support algorithms is that they can introduce biases themselves and still need to be monitored by human observers.[23] Using these systems leads to lower annotation times, reduces the risk of a cold start, avoids repetition fatigue errors, and supports the pathologist in dubious cases. Nonetheless, the behavior of these algorithms,

Figure 1.
(A) Semantic segmentation (yellow class would refer to "glandular epithelium" and blue class to "stroma"). (B) Instance segmentation (each "glandular structure" is separately identified, as an individual instance). (C) Object detection (mitoses are detected and circled in blue in a glandular epithelium).

when exposed to outlier cases, can negatively influence the pathologists' decision, which relies less on their expert insights. In addition, a small error introduced by these algorithms can be propagated to the final model, leading to undesired performance degradation. These algorithms represent a new trade-off, which should be carefully evaluated when designing a CAD system.

Addressing the Team

Building AI solutions for pathology is, first of all, an interactive and iterative process. It is interactive in the sense that it is a collaborative, 2-way, hands-on, and iterative process because it relies on testing, refining, and improving the project until the result is satisfactory. Probably, the biggest takeaway message, to one taking on an annotation endeavor, is that adequate team communication, between medical experts (namely pathologists and researchers) and ML researchers/developers is essential. Stadler et al[6] explored the best practices for constructing an annotated imaging database, and their first guiding principles reflect this importance. They emphasize that ensuring rich communication between the multidisciplinary team of experts is a prerequisite for a successful project outcome. This interaction should be based on recurrent communication with joint incentives to advance and make progress.[6] We would further accentuate the relevance of frequent and easy communication. For example, our team has set a fortnightly meeting between all project members, which has helped keep our timeline and goals on track. Moreover, annotators and ML researchers had even more regular communication to easily adjust and fine-tune data labeling to better suit the project objectives. It is important to note that annotating for AI development is a back-and-forth process and that frequent corrections and alterations to the initial designed steps will be necessary. A fluid, flexible approach is preferred and more suitable, as opposed to adhering rigidly to predefined plans.[6] Furthermore, what is perceived as simple for a medical expert can be difficult to grasp by an ML researcher. Conversely, straightforward concepts for ML experts can be challenging to understand by the clinical elements of the team. Thus, promoting effective communication across the medical and AI experts is essential to ease the process and achieve the expected results successfully (Fig. 2).

Ground-Truth Quality

A reference standard is needed for the analytical and clinical validation of an AI algorithm. In DP, there are 2 main types of reference: patient outcomes and gold standards established by pathologists.[26] Patient outcomes are probably the most valuable, but such data can take years to develop and are difficult to obtain. Hence, the most available ground truth consists of pathologists' established reference standards, which are often based on subjective interpretations.[23,27] Solutions to improve the rigor of the ground truth for subjective tasks in pathology are proposed in the recent work of Chen et al,[27] namely: to increase the number of evaluators of each case (reducing accidental errors and increasing the opinions obtained); recruiting expert evaluators (as experienced specialists will more likely identify more cases correctly); to ensure an unbiased resolution method when graders disagree (experts should perform reviews independently, and when there is disagreement, a systematic voting process or a separate arbiter could determine the final decision).

Thus, ensuring the ground truth used for the project is rigorous is a mandatory first step when developing an AI algorithm. To improve the reproducibility of the ground-truthing process in the

Figure 2.
Team collaboration. Efficient communication between the multidisciplinary team is crucial, favoring a fluid and iterative approach.

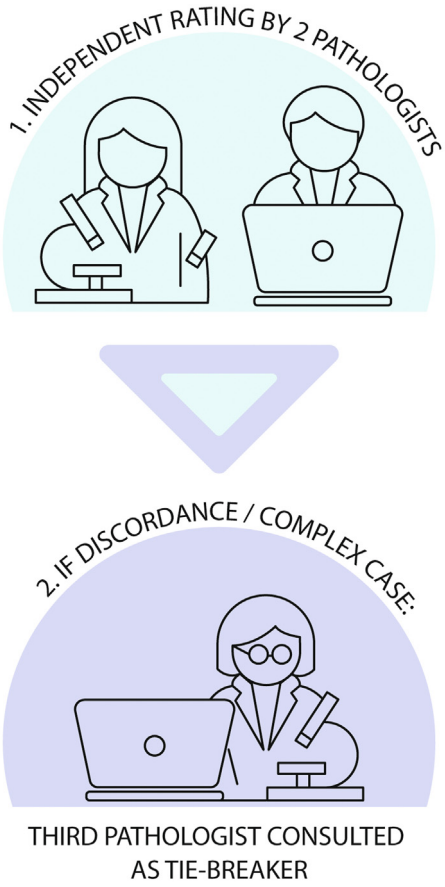

Figure 3.
A case review and ground-truthing process. Initially, each case was evaluated by 2 independent pathologists; if the classification matched, no further steps were taken. In case of discordance or in complex cases, a third pathologist was consulted as an arbiter. A similar approach was undertaken regarding full-slide annotation.

CADpath.AI project,[10,11] we initially did an exhaustive literature research concerning relevant definitions/guidelines for the proposed scope (in this case, building classification algorithms for colorectal/cervix uteri samples). A team booklet was developed with criteria to be followed in each diagnostic category in an attempt to enhance evaluation objectivity. This is in line with the recently proposed annotation workflow by the PathLake consortium, which recommends developing an annotation data dictionary as a standard reference throughout the lifecycle of the project (for a detailed description of the proposed annotation workflow, refer to Wahab et al[9]). Furthermore, regarding cases with only slide-level labeling, we defined that each WSI would be independently reviewed and rated by 2 specialist pathologists (D.M., D.O.) (because we used archived material, the initial pathology report was used as one of the inputs, reducing time and costs). If both the diagnoses were coincident, no further assessment was made; in case of divergence or for complex cases, the WSI was rechecked and rated by another pathologist (namely a subspecialized/senior pathologist, I.M.P.), working as an independent arbiter (Fig. 3). Regarding the cases that were fully annotated (with all regions of tissue within the slide delineated and tagged/classified), 2 pathologists were responsible for the task (D.M., D.O.). The strategy was to have 1 pathologist annotate the WSI, and thenceforth, it was rechecked by the other pathologist. In case of discrepancy of opinion or for complex cases, a third subspecialized/senior pathologist was consulted. Although this was a demanding approach, it allowed for better reproducibility in the ground-truth annotation. Another recommended approach to consider the inherent variability of the annotation process is to recruit multiple observers/annotators from different hospitals or laboratories.[23] During the dominance of shallow ML approaches, the reliability of the selected and extracted features was highly important. Owing to smaller data sets, less expressive algorithms, and lack of computational power, these systems had to be carefully designed.[28] Nowadays, the feature extraction process is learned and fully integrated into a deep ML system, especially in scenarios that use these guidelines to collect large data sets. On the contrary, this raises other problems concerning overfitting the training data, which can also be mitigated with the addition of samples from different distributions (another laboratory). Hence, more than ever, reliability is concerned with generalization capabilities, uncertainty handling, and adaptation.[29] All these can be tackled with an increase in the size, variability, and quality of the annotated data set following our or similar guidelines.

Choice of Hardware and Software

Regarding annotation software, there are proprietary options available (namely Leica Aperio ImageScope, Philips IntelliSite Pathology Solution, or Visiopharm's solution) and open-access alternatives (such as QuickAnnotator,[25] QuPath,[30] Cytomine,[31] ASAP,[32] DSA,[33] Orbit,[34] Omero,[35] or Sedeen Viewer[36]; for an extensive review on current systems for annotation in DP, refer to Korzynska et al[37] and to PathLake guidelines[9]). Suppose your laboratory/department has proprietary visualization software in use, it might be worth trying whether its tools are adequate for your specific annotation purposes. However, from our experience, we find that open-access solutions are better for most research endeavors in AI for DP. This is because these solutions often have more tools available and a wider range of accepted file formats. Moreover, they gather a larger volume of users, having more frequent updates and improvements. For the purposes of the practicality of this guide, we will focus on 2 available open-access software that we commonly work with (Fig. 4). QuPath is probably the most well-known and used open-source annotation software.[30,38] This application was built with the end user in mind and has a friendly user interface, allowing for easy annotation and even performing image analysis because it has its own built-in algorithms. Its annotation tools are extensive, and it is probably the most well-equipped visualization software for this purpose. It exports annotations/shapes in GeoJSON format and as labeled images, and the possibility to use scripting makes this very flexible. A great advantage of QuPath is the fact that specific documentation and useful tutorials are available (and frequently updated) by its lead developers (https://qupath.readthedocs.io/). In addition, on YouTube, several video tutorials are available, many made by QuPath users, explaining thoroughly how to best use the software.

Next, Sedeen Viewer is an image viewer built specifically for DP visualization with rich annotation capabilities. It exports its files in XML formats, which is a file format designed for both human and machine readability. XML markup language supports the organization of any arbitrary data into a hierarchy of structures, thus improving the standardization of data serialization. XML internal configuration may vary between different software brands, but these remain quite similar and easily adapted for reading by a processing script. Sedeen has a really simple user interface, which can be a plus when one is starting annotation efforts. Although it

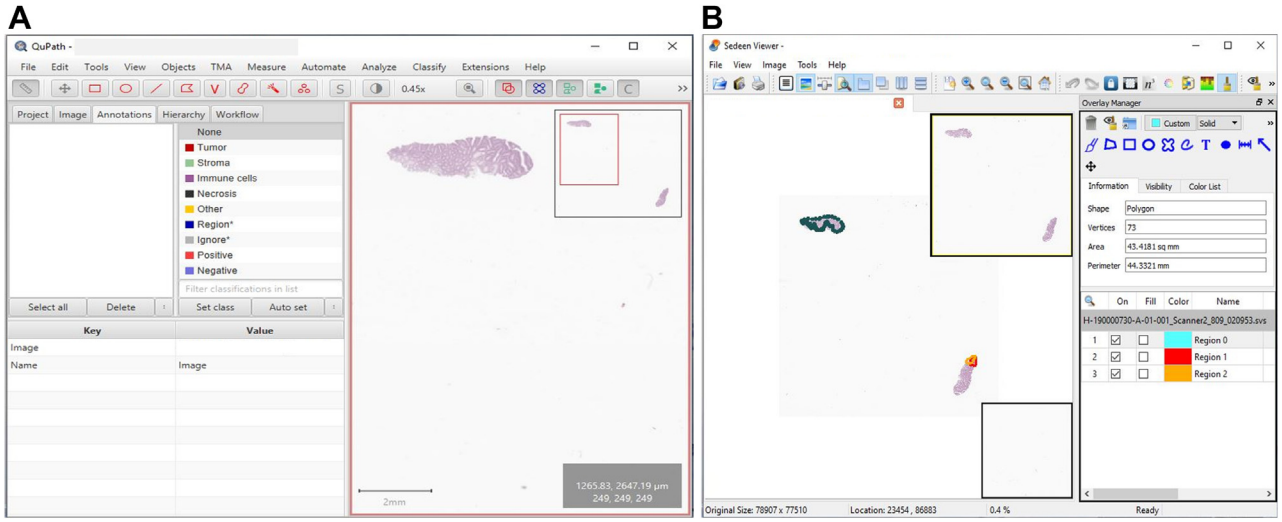

Figure 4.
Open-source annotation software interfaces: (A) QuPath; (B) Sedeen Viewer.

does not have a wide platform of users such as QuPath, with an extensive support question forum, we found that emailed questions are usually promptly responded to by the supporting team.

Concerning hardware (particularly for annotation), there are multiple different options available, namely visualization on a computer screen paired with a classic mouse or a drawing pen and pad; a computer equipped with a touch screen, tablets, or digital drawing boards (Fig. 5). From our experience, we have summarized the pros and cons of each option (Table 1), but, in the end, it is mostly a personal decision that comes down to finding the right fit regarding cost-benefit and usability. Moreover, some software options, such as QuPath, have shortcuts that affect the practicability of the hardware (for example, making the use of the mouse more convenient), and the hardware choice can also be influenced by the software being used.

Addressing Common Problems in Annotation

How to Deal With Low Image Quality/Color Discrepancies?

The boundary to decide whether an image has enough quality to be used for an AI solution or not is fuzzy.[23] As a rule of thumb, we decided that only images that were not good enough for diagnosis would be excluded from our studies, and the models would have to deal with considerable variability. In our view, this would contribute to enhancing the robustness of the algorithms. Normalization and generalization across different color patterns is an active research field of high importance, which enables algorithms to work across several scanners and data collection schemes.[39,40] Frequently, external data sets also comprise low-quality samples. Hence, these should not be completely excluded from the training data set; instead, several approaches could be followed, such as curriculum learning, annotation of the relevant areas, and quality-aware classification systems. Of utmost importance is the use of large data sets, preferably including external cases, so that the model is exposed to greater variability.[41]

What Is the Needed Level of Detail When Drawing/Contouring?

A common question that beginners face is how to delineate the tissue fragments. Should the outline of the different fragments within a WSI be extensively and rigorously delineated? In fact, annotating tissue/background borders does not need too much detail because an automatic binarization filter, such as Otsu thresholding,[42] will most likely be used to do so, with good detail and minimum tuning. These kinds of filters are able to trim the external tissue contours, saving time for the annotator (Fig. 6). On the contrary, tissue regions with different characteristics within the fragments should be carefully delineated to be clearly distinguished in the final annotation. We need to keep the different classes as well-demarcated as possible. When labeling or annotating WSI, we should avoid including 2 different classes within the same classification (this concept is further explained in Fig. 6).

Which Fragments to Annotate When the Slide Has Repeated Fragments' Cuts?

For slides with repeated consecutive sections (Fig. 7), there is no need to annotate all fragments: one can manually discard the unwanted set, or an ML model can automatically detect similar fragments and discard repeated information, as proposed by Albuquerque et al.[12] However, it is preferable to annotate the fragment with the best quality, in case of some type of artifact. Moreover, selecting one section to be annotated leads to better use of good quality tissue areas without discarding the entire sample, reduces the necessary training computation times (sometimes to less than half), and eliminates the biases toward repeated tissue samples within the training process. This latter aspect is rather important to ensure that every tissue patch has the same weight in the loss function. In other words, if an ML model is repeatedly exposed to N versions of the exact same tissue, its effect on the optimization is magnified by a factor of N. Thus, biasing the optimization algorithm toward attributing higher importance to the minimization of the error in that tissue sample when compared with the minimization of the error in other samples.

How to Transfer the Massive Amount of Data?

Transferring multiple WSIs, which easily have 1-1.5 GB each, and the correspondent annotations, between facilities, is an issue that should be addressed early on. Depending on the volume of data and proximity between institutions, one easy solution can be to simply use pen drives or external disks. However, the simplicity of this approach is accompanied by several disadvantages. First,

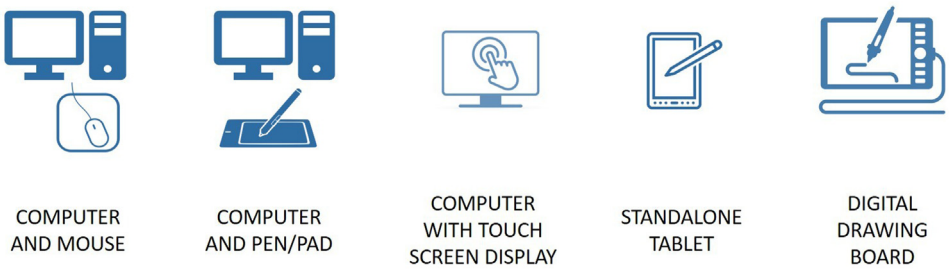

Figure 5.
Available hardware solutions for annotation.

several versions of these devices have low bandwidth; even for high-bandwidth devices, faster USB ports are not always available. Hence, they increase the sharing and access I/O overhead time. Second, as experienced during the pandemic, it might be necessary to have a distributed and remote team working on the same problem. This solution does not allow remote or multiple access points. Finally, owing to their mobility and portability, these devices are exposed to a greater risk of physical damage by several orders of magnitude. For example, our project used a remote file transfer with an encrypted connection using the sftp protocol with a certificate and user authentication. Furthermore, open science and FAIR data ("findable, accessible, interoperable and reusable" data) have been one of the main reasons for the current increase in collaborative and shareable research.[43] This philosophy leads to more complete and risk-free science because several researchers can not only review previous work but also work on it further. For this to be possible, it is necessary to share code, data, and insights.[43] Data are the most important requirement. With data, it is possible to establish benchmarks, ensure the reproducibility of a certain approach, and even allow researchers from less developed laboratories and centers to take part in the current research. But, sharing data is not as straightforward as it seems, especially in the medical domain. It is necessary to ensure that there is no possible link between the shared data and the patient, known as hard anonymization. The data quality must be retained, and data description must be of the highest quality. The sharing platform must be dedicated and not be linked to management software used in the laboratory, thus ensuring that it does not affect the ongoing laboratory work and that there is an easy access philosophy. Finally, it might be required to work together with law experts to create the necessary sharing licenses.

Designing Classification Algorithms

As stated earlier, this article derives from our own experience, and our projects have mostly focused on developing classification algorithms for colorectal and cervix uteri samples. In this study, we detail the steps we took regarding annotation for classification tasks, hoping it will help other researchers dealing with similar annotation endeavors.

Annotation for a Colorectal Neoplasia Classification Model

When designing a classifier, the first and most important step should be defining why the algorithm is important for clinical practice and how the algorithm is intended to be used: is it supposed to triage cases and run upfront before the pathologists' observation of the cases (similarly to Paige Prostate solution[44])? Or will it work as a second opinion, and the pathologist will only select a few cases to be analyzed by the algorithm? This choice will influence the definition of classes to be given by the algorithm and the following steps of the project. Next, it is important to decide which classes the classification system will be able to identify. Remember that difficult tasks for pathologists will also be a hurdle

Table 1
Different hardware solutions for annotation: pros and cons

Hardware solution	Pros	Cons
Computer + mouse	Most available and affordable option The mouse is the solution that most people will have the most upfront experience and ease to use	For extensive annotations, the mouse can be tiresome and nonergonomic
Computer + pen/pad	It is an affordable option Generally user-friendly after the initial learning curve	It requires a learning curve to be able to draw accurately; some people might not adapt
Computer with touch screen (an example would be Microsoft Surface Studio)	Drawing on top of the image is in principle the easiest strategy to annotate	Most annotation software is not optimized for touch screens yet or needs to be tweaked to work properly Can be an expensive solution To best fit the purposes of annotation, the screen needs to be movable, as drawing in a vertical position is not comfortable
Standalone touch screen tablet	Drawing on top of the image is in principle the Easiest strategy to annotate Easiest to transport	Most annotation software is not compatible with iOS or android systems. It is required to use compatible software or acquire a tablet with Windows operating system for example Many tablets can have insufficient capability to store, transfer, and run WSIs efficiently Can be an expensive solution
Digital drawing board (an example are the Wacom displays)	Drawing on top of the image is in principle the easiest strategy to annotate	Most annotation software is not optimized for touch screens yet or needs to be tweaked to work properly Can be an expensive solution

WSI, whole-slide image.

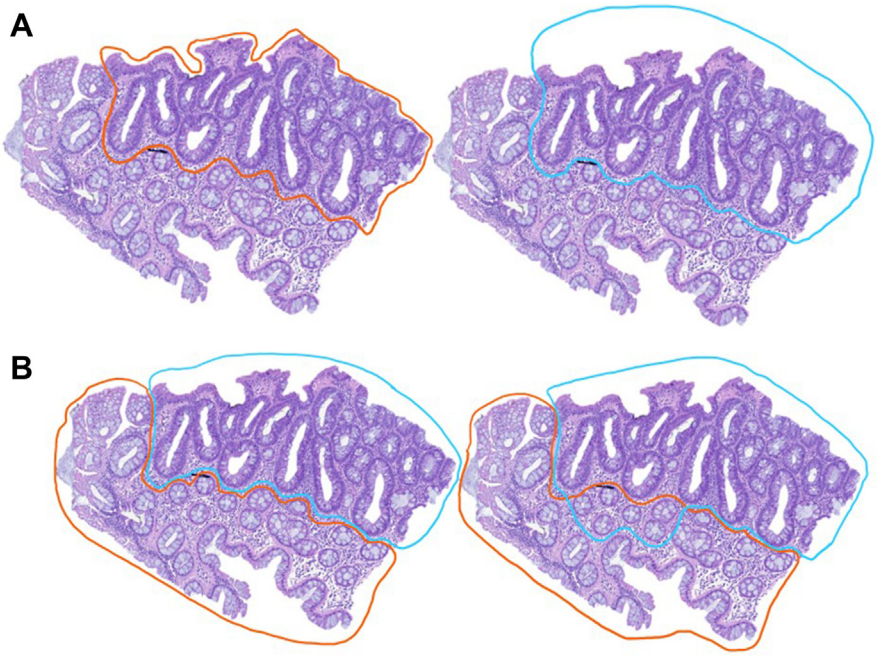

Figure 6.
(A) Contouring fragment limits (colorectal biopsy, hematoxylin and eosin). When delineating fragments, there is no need to exhaustively contour their external limits (first image, in orange) because an automatic filter might easily select the tissue areas. A simple contour (as shown in the second example, in blue) will suffice and save significant time in this task. (B) Annotating different classes/categories. Annotators should keep the different classes well-demarcated from each other. In this example, the first image shows well-defined boundaries between "normal epithelium" (orange outline) and "dysplastic epithelium" (blue outline); whether, in the following image, the annotation is inaccurate, with part of the "normal epithelium" being incorrectly encompassed within the "dysplastic" region. The same rationale applies to other tasks, such as distinguishing tumor from nontumor and epithelium from stroma.

for the algorithm, but it is in these that the added value of CAD solutions is better appreciated. In this study, we decided to develop a 3-tiered classifier for colorectal biopsies that could identify nonneoplastic mucosa, low-risk lesions (encompassing low-grade conventional adenomas) and high-risk lesions (high-grade adenomas, such as intramucosal carcinomas, and adenocarcinomas).[10,11] We aimed for it to have high sensitivity to detect the high-risk lesions. We started our effort with approximately 1100 WSIs, of which 10% were fully annotated. All have been weakly annotated, and all cases had at least a slide-level label.[10] In the fully annotated slides, all pixels (all area) within the slide had a classification. To do this, we used Sedeen software and manually delineated all "low-risk areas" (blue) and "high-risk" areas (purple). The tissue areas left unannotated were assumed as "nonneoplastic" (or, in the case normal tissue was found within a dysplastic/malignant area, a white outline would also correspond to "nonneoplastic"). Furthermore, we created a "not usable" category, delineating in gray areas to be ignored by the algorithm (namely, regions within the slide that had artifacts precluding adequate classification). A useful tip is to record the color code (RGB, hexadecimal, etc) of the used colors (this information is available in the annotation software) because each color will have different shades. It is important to always use the same to match the label recognized by a preprocessing algorithm that converts them into the desired structure to be used as ground truth for ML algorithms. The annotation effort took between 15 minutes (in small adenomas with small dysplastic areas) and several hours (large polyps with admixed areas of low- and high-grade lesions), and it was all performed using a regular mouse and computer. The choice of hardware was mostly linked to convenience and cost-benefit. Finally, we saved the annotation files in XML format and transferred the images and corresponding annotation to the ML team to process them, as explained in the section "How to Transfer the Massive Amount of Data?". The processing of annotations is

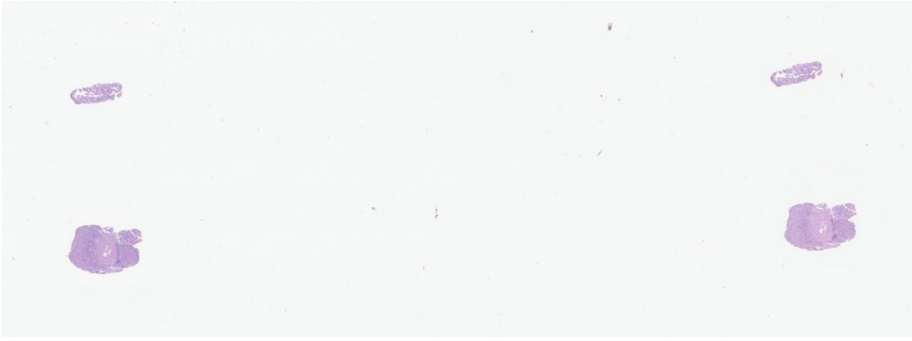

Figure 7.
Whole-slide image with consecutive sections of tissue (repeated cuts).

Table 2
Practical suggestions for Digital Pathology annotators

Carefully define your end-goal to determine which type of annotation will be needed. Also, evaluate the amount of data you have available, as this will also impact the type of machine learning strategy that can be used.
If possible, when choosing a hardware/software solution, test different options before committing to one.
Explore in which format your annotation software exports the annotation files, as this might be relevant for your further work with the machine learning research team.
Find a software which allows to easily delete annotations (although this seems a basic feature, in fact, the initial proprietary software we tested required multiple steps to delete a specific annotation, which greatly impaired the annotation flow).
To avoid discarding slides with identifiable small regions with low image quality, one can signal/annotate these areas, so they can be evaluated in the image preprocess stage, in order to be enhanced or even avoided by the classifier for better performance.
The annotation process is costly and time-consuming; to facilitate, complementary strategies may be required, such as interactive annotations (software aided) or opting for weakly supervised ML approaches.

ML, machine learning.

highly dependent on the task. In our use case, it was necessary to associate each pixel of the image with the label corresponding to its specific color. Afterward, after dividing the slide into several tiles, each tile was associated with the label with the highest presence in its pixels. In addition, slide-level labeling (or "weak" annotation) was performed in all cases, attributing a global diagnosis/classification for the entire slide. The slides were assessed by 1 of the 2 pathologists, and when the diagnosis was different from the initial report (which served as a second independent grader), the case was rechecked by a third pathologist tiebreaker (Fig. 3). Within this specific project, the following studies have leveraged the data set to 4000 cases[11] and finally to 10,500 WSIs (article in preparation) and followed the same annotation/labeling scheme. In this most recent data set, all 10,500 WSIs were weakly annotated (slide-level label) and $\approx$1000 were fully manually annotated. Importantly, from the final model, a clinical prototype was built for routine practice use, and it is currently under validation.

Annotation for a Cervix Uteri Classification Model

For this recent project, we collected 2000 WSIs (hematoxylin and eosin stained) to develop a classification algorithm to accurately detect and classify intraepithelial cervix uteri lesions (work developed during 2022 and submitted for publication in October 2022). The case set consisted of loop electrosurgical excision procedure samples and surgical specimens. The model is a 4-tier classifier: nonneoplastic, low-grade intraepithelial lesion, high-grade intraepithelial lesion, and nonrepresentative (for samples without squamous epithelium, namely with only endocervical tissue or slides containing only mucous material). Besides labeling all cases with a slide-level classification, around 200 WSIs were manually annotated. Because we needed to focus on epithelial areas, this model required a different annotation approach compared with our previous work on colorectal samples. Thus, we divided the effort into 2 phases: segmentation and classification. First, we manually contoured the squamous epithelia in red, so that a segmentation algorithm could later extract these areas for tissue classification. After, ROI were delineated by free-hand drawing or bounding boxes: characteristic low-grade intraepithelial lesion areas in light blue, high-grade intraepithelial lesion areas in purple, and nonneoplastic epithelium in orange. These annotations served not only to identify unequivocal areas of each tissue type but also to attribute a label to each epithelium area. With such an approach, although the slides were only partially annotated, the annotation process is facilitated and the annotations can be more precise. We further annotated, in yellow, areas corresponding to nonrepresentative tissue and, in black, surrounding areas to be ignored by the model (namely, regions with artifacts or significantly unfocused, which precluded adequate classification). Finally, we sent the annotations (.XML) and the images to the engineering team to be processed. All epithelium delineations were intersected with the result of Otsu thresholding, retrieving a segmentation mask with the identified epithelium tissue areas in white. Finally, the pixels that fall into the annotated interest areas were marked with the corresponding annotation color, which could later be converted into a class label.

Discussion

Only high-level quality annotations will be able to leverage AI solutions for clinical use in pathology. Although annotation particularities will vary according to the specific project goals, many common hurdles will be similar across the different DP endeavors. In Table 2, some suggestions to tackle the annotation effort are summarized. Thus, this work intends to be a practical resource to assist pathologists and other researchers in producing consistent and high-quality annotated data.

Acknowledgments

The authors thank Manuela Felizardo for creating the image of Figures 2 and 3.

Author Contributions

D.M. conceived the study. D.M., S.O., and P.N. performed literature review and wrote the original draft. D.O., A.M., J.C., and I.M.-P. reviewed and edited the manuscript. J.C. and I.M.-P. supervised the study. All authors read and agreed to the published version of the manuscript.

Data Availability

Data sharing does not apply to this article as no new data were created or analyzed in this study.

Funding

This work was financed by the ERDF—European Regional Development Fund through the Operational Programme for Competitiveness and Internationalization—COMPETE 2020 Programme within project POCI-01-0247-FEDER-045413 and by national funds through the Portuguese funding agency, FCT—Foundation for Science and Technology Portugal, within the

project LA/P/0063/2020 and the PhD grants SFRH/BD/139108/2018 and 2021.06872.BD.

Declaration of Competing Interest

The authors declare that there are no competing interests.

Ethical Approval and Consent to Participate

No ethical approval was necessary for this study as no patient information or material was directly used.

References

1. Bulten W, Kartasalo K, Chen PC, et al. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nat Med*. 2022;28(1):154–163. https://doi.org/10.1038/s41591-021-01620-2
2. Mehta S, Lu X, Wu W, et al. End-to-End diagnosis of breast biopsy images with transformers. *Med Image Anal*. 2022;79:102466. https://doi.org/10.1016/j.media.2022.102466
3. Pantanowitz L, Hartman D, Qi Y, et al. Accuracy and efficiency of an artificial intelligence tool when counting breast mitoses. *Diagn Pathol*. 2020;15(1):80. https://doi.org/10.1186/s13000-020-00995-z
4. Feng M, Deng Y, Yang L, et al. Automated quantitative analysis of Ki-67 staining and HE images recognition and registration based on whole tissue sections in breast carcinoma. *Diagn Pathol*. 2020;15(1):65. https://doi.org/10.1186/s13000-020-00957-5
5. Echle A, Rindtorff NT, Brinker TJ, Luedde T, Pearson AT, Kather JN. Deep learning in cancer pathology: a new generation of clinical biomarkers. *Br J Cancer*. 2021;124(4):686–696. https://doi.org/10.1038/s41416-020-01122-x
6. Stadler CB, Lindvall M, Lundström C, et al. Proactive construction of an annotated imaging database for artificial intelligence training. *J Digit Imaging*. 2021;34(1):105–115. https://doi.org/10.1007/s10278-020-00384-4
7. Mitchell BR, Cohen MC, Cohen S. Dealing with multi-dimensional data and the burden of annotation: easing the burden of annotation. *Am J Pathol*. 2021;191(10):1709–1716. https://doi.org/10.1016/j.ajpath.2021.05.023
8. Qayyum A, Qadir J, Bilal M, Al-Fuqaha A. Secure and robust machine learning for healthcare: a survey. *IEEE Rev Biomed Eng*. 2021;14:156–180. https://doi.org/10.1109/RBME.2020.3013489
9. Wahab N, Miligy IM, Dodd K, et al. Semantic annotation for computational pathology: multidisciplinary experience and best practice recommendations. *J Pathol Clin Res*. 2022;8(2):116–128. https://doi.org/10.1002/cjp2.256
10. Oliveira SP, Neto PC, Fraga J, et al. CAD systems for colorectal cancer from WSI are still not ready for clinical acceptance. *Sci Rep*. 2021;11(1):14358. https://doi.org/10.1038/s41598-021-93746-z
11. Neto PC, Oliveira SP, Montezuma D, et al. iMIL4PATH: a semi-supervised interpretable approach for colorectal whole-slide images. *Cancers (Basel)*. 2022;14(10):2489. https://doi.org/10.3390/cancers14102489
12. Albuquerque T, Moreira A, Barros B, et al. Quality control in digital pathology: automatic fragment detection and counting. *Annu Int Conf IEEE Eng Med Biol Soc*. 2022:588–593. https://doi.org/10.1109/EMBC48229.2022.9871208
13. Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med*. 2019;25(8):1301–1309. https://doi.org/10.1038/s41591-019-0508-1
14. Chen CL, Chen CC, Yu WH, et al. An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning. *Nat Commun*. 2021;12(1):1193. https://doi.org/10.1038/s41467-021-21467-y
15. Yakimovich A, Beaugnon A, Huang Y, Ozkirimli E. Labels in a haystack: approaches beyond supervised learning in biomedical applications. *Patterns*. 2021;2(12), 100383. https://doi.org/10.1016/j.patter.2021.100383
16. Wang S, Yang DM, Rong R, Zhan X, Xiao G. Pathology image analysis using segmentation deep learning algorithms. *Am J Pathol*. 2019;189(9):1686–1698. https://doi.org/10.1016/j.ajpath.2019.05.007
17. Yang R, Yu Y. Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis. *Front Oncol*. 2021;11, 638182. https://doi.org/10.3389/fonc.2021.638182
18. Hameed Z, Garcia-Zapirain B, Aguirre JJ, Isaza-Ruget MA. Multiclass classification of breast cancer histopathology images using multilevel features of deep convolutional neural network. *Sci Rep*. 2022;12(1):15600. https://doi.org/10.1038/s41598-022-19278-2
19. Aubreville M, Stathonikos N, Bertram CA, et al. Mitosis domain generalization in histopathology images—the MIDOG challenge. *arXiv Preprint*. 2022; arXiv:2204.03742. https://doi.org/10.48550/arXiv.2204.03742
20. Weigert M, Schmidt U. Nuclei instance segmentation and classification in histopathology images with Stardist. In: *2022 IEEE International Symposium on Biomedical Imaging Challenges (ISBIC)*. IEEE; 2022:1–4. https://doi.org/10.1109/ISBIC56247.2022.9854534
21. Shvetsov N, Grønnesby M, Pedersen E, et al. A pragmatic machine learning approach to quantify tumor-infiltrating lymphocytes in whole slide images. *Cancers (Basel)*. 2022;14(12):2974. https://doi.org/10.3390/cancers14122974
22. Lindman K, Rose JF, Lindvall M, Lundström C, Treanor D. Annotations, ontologies, and whole slide images—development of an annotated ontology-driven whole slide image library of normal and abnormal human tissue. *J Pathol Inform*. 2019;10:22. https://doi.org/10.4103/jpi.jpi_81_18
23. Homeyer A, Geißler C, Schwen LO, et al. Recommendations on compiling test datasets for evaluating artificial intelligence solutions in pathology. *Mod Pathol*. 2022;35(12):1759–1769. https://doi.org/10.1038/s41379-022-01147-y
24. Möhle L, Bascuñana P, Brackhan M, Pahnke J. Development of deep learning models for microglia analyses in brain tissue using DeePathology™ STUDIO. *J Neurosci Methods*. 2021;364:109371. https://doi.org/10.1016/j.jneumeth.2021.109371
25. Miao R, Toth R, Zhou Y, Madabhushi A, Janowczyk A. Quick Annotator: an open-source digital pathology-based rapid image annotation tool. *J Pathol Clin Res*. 2021;7(6):542–547. https://doi.org/10.1002/cjp2.229
26. Dudgeon SN, Wen S, Hanna MG, et al. A pathologist-annotated dataset for validating artificial intelligence: a project description and pilot study. *J Pathol Inform*. 2021;12:45. https://doi.org/10.4103/jpi.jpi_83_20
27. Chen PC, Mermel CH, Liu Y. Evaluation of artificial intelligence on a reference standard based on subjective interpretation. *Lancet Digit Health*. 2021;3(11):e693–e695. https://doi.org/10.1016/S2589-7500(21)00216-8
28. Kocak B, Ates E, Durmaz ES, Ulusan MB, Kilickesmez O. Influence of segmentation margin on machine learning-based high-dimensional quantitative CT texture analysis: a reproducibility study on renal clear cell carcinomas. *Eur Radiol*. 2019;29(9):4765–4775. https://doi.org/10.1007/s00330-019-6003-8
29. Tran D, Liu J, Dusenberry MW, et al. Plex: towards reliability using pretrained large model extensions. *arXiv Preprint*. 2022; arXiv:2207.07411v1. https://doi.org/10.48550/arXiv.2207.07411
30. Bankhead P, Loughrey MB, Fernández JA, et al. QuPath: Open source software for digital pathology image analysis. *Sci Rep*. 2017;7(1):16878. https://doi.org/10.1038/s41598-017-17204-5
31. Rubens U, Hoyoux R, Vanosmael L, et al. Cytomine: toward an open and collaborative software platform for digital pathology bridged to molecular investigations. *Proteomics Clin Appl*. 2019;13(1):e1800057. https://doi.org/10.1002/prca.201800057
32. Diagnostic Image Analysis Group (Computation Pathology Group). Radboud University Medical Center. ASAP. Accessed August 26, 2022. https://computationalpathologygroup.github.io/ASAP/
33. Gutman DA, Khalilia M, Lee S, et al. The digital slide archive: a software platform for management, integration, and analysis of histology for cancer research. *Cancer Res*. 2017;77(21):e75–e78. https://doi.org/10.1158/0008-5472.CAN-17-0629
34. Stritt M, Stalder AK, Vezzali E. Orbit image analysis: an open-source whole slide image analysis tool. *PLoS Comput Biol*. 2020;16(2):e1007313. https://doi.org/10.1371/journal.pcbi.1007313
35. Allan C, Burel JM, Moore J, et al. OMERO: flexible, model-driven data management for experimental biology. *Nat Methods*. 2012;9(3):245–253. https://doi.org/10.1038/nmeth.1896
36. Martel AL, Hosseinzadeh D, Senaras C, et al. An image analysis resource for cancer research: PIIP-pathology image informatics platform for visualization, analysis, and management. *Cancer Res*. 2017;77(21):e83–e86. https://doi.org/10.1158/0008-5472.CAN-17-0323
37. Korzynska A, Roszkowiak L, Zak J, Siemion K. A review of current systems for annotation of cell and tissue images in digital pathology. *Biocybern Biomed Eng*. 2021;41(4):1436–1453. https://doi.org/10.1016/j.bbe.2021.04.012
38. Bankhead P. Developing image analysis methods for digital pathology. *J Pathol*. 2022;257(4):391–402. https://doi.org/10.1002/path.5921
39. Runz M, Rusche D, Schmidt S, Weihrauch MR, Hesser J, Weis CA. Normalization of HE-stained histological images using cycle-consistent generative adversarial networks. *Diagn Pathol*. 2021;16(1):71. https://doi.org/10.1186/s13000-021-01126-y
40. Boschman J, Farahani H, Darbandsari A, et al. The utility of color normalization for AI-based diagnosis of hematoxylin and eosin-stained pathology images. *J Pathol*. 2022;256(1):15–24. https://doi.org/10.1002/path.5797
41. de Hond AAH, Leeuwenberg AM, Hooft L, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *NPJ Digit Med*. 2022;5(1):2. https://doi.org/10.1038/s41746-021-00549-7
42. Otsu N. A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern*. 1979;9:62–66.
43. European Commission. Directorate-General for Research and Innovation. Turning FAIR into reality: final report and action plan from the European Commission expert group on FAIR data, Publications Office (2018). Accessed March 23, 2022. https://data.europa.eu/doi/10.2777/1524
44. da Silva LM, Pereira EM, Salles PG, et al. Independent real-world application of a clinical-grade automated prostate cancer detection system. *J Pathol*. 2021;254(2):147–158. https://doi.org/10.1002/path.5662