Lab Invest 105 (2025) 102203

LABORATORY INVESTIGATION

journal homepage: https://laboratoryinvestigation.org/

Research Article

Annotation Practices in Computational Pathology: A European Society of Digital and Integrative Pathology (ESDIP) Survey Study

Diana Montezuma[a,b,c,*], Sara P. Oliveira[a,d,*], Yuri Tolkach[a,e], Peter Boor[a,f], Alex Haragan[a,g], Rita Carvalho[a,h], Vincenzo Della Mea[a,i], Tim-Rasmus Kiehl[a,j], Sabine Leh[a,k,l], Mustafa Yousif[a,m], David Ameisen[a,n], Mircea-Sebastian Șerbănescu[a,o], Norman Zerbe[a,f,j], Vincenzo L'Imperio[a,p,q]

[a] *European Society of Digital and Integrative Pathology (ESDIP), Lisboa, Portugal;* [b] *Research & Development Unit, IMP Diagnostics, Porto, Portugal;* [c] *Cancer Biology and Epigenetics Group, Research Center of Portuguese Oncology Institute of Porto/RISE@Research Center of Portuguese Oncology Institute of Porto (Health Research Network), Portuguese Oncology Institute of Porto/Porto Comprehensive Cancer Centre Raquel Seruca, Porto, Portugal;* [d] *Department of Molecular Pathology, The Netherlands Cancer Institute, Amsterdam, The Netherlands;* [e] *Institute of Pathology, University Hospital Cologne, Cologne, Germany;* [f] *University Clinic Aachen, RWTH University Aachen, Aachen Germany;* [g] *Department of Pathology, Royal Liverpool University Hospital, Liverpool, United Kingdom;* [h] *Department of Pathology, Medizinische Versorgungszentren (MVZ) Helios Klinikum Emil von Behring, Berlin, Germany;* [i] *Department of Mathematics, Computer Science and Physics, University of Udine, Udine, Italy;* [j] *Charité-Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Medical Informatics, Berlin, Germany;* [k] *Department of Pathology, Haukeland University Hospital, Bergen, Norway;* [l] *Department of Clinical Medicine, University of Bergen, Bergen, Norway;* [m] *Department of Pathology, University of Michigan Medical School, Ann Arbor, Michigan;* [n] *Imginit SAS, Paris, France;* [o] *Department of Medical Informatics and Biostatistics, University of Medicine and Pharmacy of Craiova, Romania;* [p] *School of Medicine and Surgery, University of Milano-Bicocca, Milan, Italy;* [q] *Department of Pathology, Fondazione IRCCS San Gerardo dei Tintori, Monza, Italy*

ARTICLE INFO

Article history:
Received 2 August 2024
Revised 2 November 2024
Accepted 19 November 2024
Available online xxx

Keywords:
annotation
digital pathology
European Society of Digital and Integrative Pathology
survey

ABSTRACT

Integrating digital pathology and artificial intelligence (AI) algorithms can potentially improve diagnostic practice and precision medicine. Developing reliable, generalizable, and comparable AI algorithms depends on access to meticulously annotated data. However, achieving this requires robust collaboration among pathologists, computer scientists, and other researchers to ensure data quality and consistency. The lack of standardization and scalability is a significant challenge when generating annotations and annotated data sets. Recognizing these limitations, the Scientific Committee of the European Society of Digital and Integrative Pathology (ESDIP) performed a comprehensive international survey to understand the current state of annotation practices and identify actionable areas to address critical needs in the annotation process. The analysis and summary of the survey results provide several insights for all stakeholders involved in data preparation and ground truthing, ultimately contributing to the advancement of AI in computational pathology.

Introduction

The adoption of digital pathology in routine practice is a prerequisite for the development and application of artificial intelligence (AI) algorithms for a wide array of tasks, such as tumor

These authors contributed equally to this work: Diana Montezuma and Sara P. Oliveira.
* Corresponding authors.
E-mail addresses: diana.felizardo@impdiagnostics.com (D. Montezuma), s.oliveira@nki.nl (S.P. Oliveira).

https://doi.org/10.1016/j.labinv.2024.102203

detection, subtyping, and grading,[1,2] and classification of non-oncological diseases or biomarker evaluation such as (positive) cell counting.[3] Many promising diagnostic, prognostic, or predictive AI algorithms are being developed and tested for various diseases and clinical situations. As these technologies become increasingly available in pathology diagnostics, it is necessary to ensure their reliability, generalizability, and comparability. The development of AI algorithms depends on access to large volumes of annotated/labeled data.[4,5] Manual annotation and labeling, however, can be both time-consuming and laborious and currently lack any standardization concerning the principle of annotations, their precision, or the structure of classes. Initial efforts to harmonize annotation practices for research have been made.[6-9] However, practicable recommendations are lacking. To understand the current state of annotation practices, the Scientific Committee of the European Society of Digital and Integrative Pathology (ESDIP) conducted an international survey focusing on manual annotations as these are currently most widely used for the development of clinical-grade digital pathology tools using state-of-the-art supervised principles. Thus, the analysis and summary of the survey results are intended to serve as an empirical basis and a starting point for stakeholders to guide large-scale standardization and regulatory efforts for data preparation and ground truthing in computational pathology.

Materials and Methods

Survey

The authors, an international group of pathologists and computer scientists from the Scientific Committee and the Board of the ESDIP, created an online questionnaire using the Google Forms web platform (Google) to assess the current annotation strategies for computational pathology. The survey comprised 40 comprehensive questions on numerous aspects of the annotation process (Supplementary Material.pdf), which are divided into the following 4 sections: general information, information about the research group and focus, details about the annotation process, and lessons learned. The objective was to obtain insights into current annotation practices and unveil heterogeneity/disparities of the different approaches to propose eventual solutions to commonly encountered problems. The survey was initially taken in September 2023 by 6 Scientific Committee members, who provided their input. This was followed by iterative revisions through email exchanges to improve the questions and structure. Shortly thereafter, a pilot test was conducted, and several members of the ESDIP Board, pathology and computational experts, completed the survey, which allowed for its further refinement. The survey plan was initially presented during the first EurAsia Academy on Digital Pathology web meeting (www.eurasiadp.org/) and then circulated among pathologists and researchers using ESDIP's mailing list, ESDIP's social media accounts, the authors' social media accounts (Facebook, LinkedIn, and X), and via personal contacts to encourage participation. The survey aimed to include researchers involved not only in human but also in preclinical, veterinary, and toxicologic pathology. Participants were asked to complete the survey only if they currently have or have had computational pathology projects involving manual annotations. Our survey methodology adhered to the General Data Protection Regulation, ensuring compliance with data protection and privacy standards throughout data collection. Participants who agreed to be mentioned are acknowledged in the collaborators' list of this article. The survey ran for 2 months (October 1, 2023, to November 30, 2023).

Statistical Analysis

After closing the survey, results were extracted from Google Forms and added to Google Sheets for preliminary filtering of the received answers. In total, there were 144 respondents. Seven replies were excluded (6 repeat respondents and 1 unrelated to pathology annotation; in the case of repeated submissions from the same participant, only the last submission was considered), leading to a final pool of 137 participants. Regarding free text input questions, blank or invalid replies (such as "."; "-"; "..."; or similar) were not recorded. Valid replies were accounted for and underwent analysis summarized in the "Results" section. Results are reported as absolute numbers and percentages, which can exceed 100% for some questions when aggregated, as each participant could give more than 1 answer. In detail, the percentage was based on the number of participants for each question, calculated using the following formula: the number of participants selecting a choice/total number of participants on that question $\times$ 100. In instances of conflicting results between open- and closed-ended questions, priority was given to replies obtained from the latter. Descriptive statistics were performed with Excel 2016 (Microsoft).

Results

Characteristics of the Participants

The majority of the respondents were board-certified pathologists (n = 81; 59.1%), aged 30 to 55 years (n = 110; 80.3%), and based in Europe (n = 94; 68.6%). Most were directly involved in the manual annotation process for projects of their research group (n = 87; 63.5%), with approximately half of them working for academic/noncommercial purposes (n = 71; 51.8%), as shown in Figure 1.

Research Groups and Interests

In the majority of cases, people responsible for the annotation process, that is, those who defined the annotation principles and controlled the final result/quality, were board-certified pathologists (n = 120; 87.6%), followed by pathology residents (n = 59; 43.1%), medical students (n = 32; 23.4%), biomedical scientists (n = 29; 21.2%), computer scientists/engineers (n = 30; 21.9%), and other persons (n = 10; 7.3%). The pathology expertise of annotators might be of utmost importance, especially while developing diagnostic tools. In addition, diagnostic tools in many domains might require subspecialization from annotators/pathologists involved (eg, genitourinary pathologists as annotators in developing prostate algorithms). Interestingly, over half of the research groups (n = 81; 59.1%) declared they had no subspecialists involved in the annotation process or any aspect of the project. The employment of subspecialists in the annotation process was reported by 21 (19.7%) research groups. Finally, among those who selected more than 1 answer, 43 teams (31.4%) included both subspecialized (in all or just some applications) and non-subspecialized annotators in their average projects. Some participants declared that the annotation load was split between pathologists and annotators without a pathologic background. In such situations, pathologists were involved in more complex tasks (eg, lymph nodes, easily detectable even without pathologic expertise, were handled by less-specialized annotators while germinal centers were annotated by a pathology specialist).

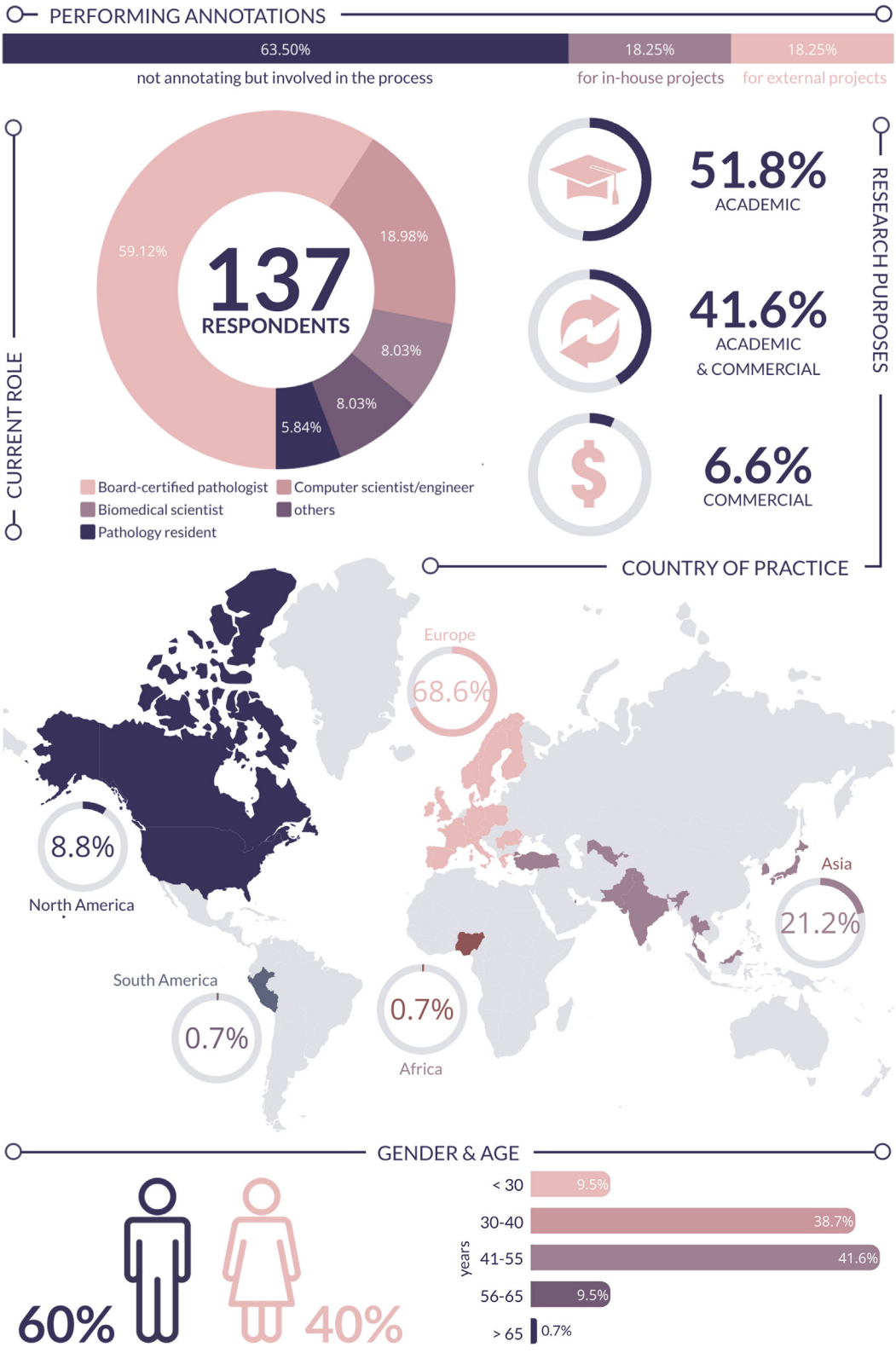

Figure 1.
A summary of the characteristics of the survey's participants.

Further details on the training of annotators were provided by 74 participants (54%), with many responses with more than 1 statement (Supplementary Material 2-Table S2.1). In this subset of respondents, pathologists had major involvement in the specific training of the annotators (n = 48; 62.3%), namely by guiding/instructing the annotators and reviewing the quality of the final

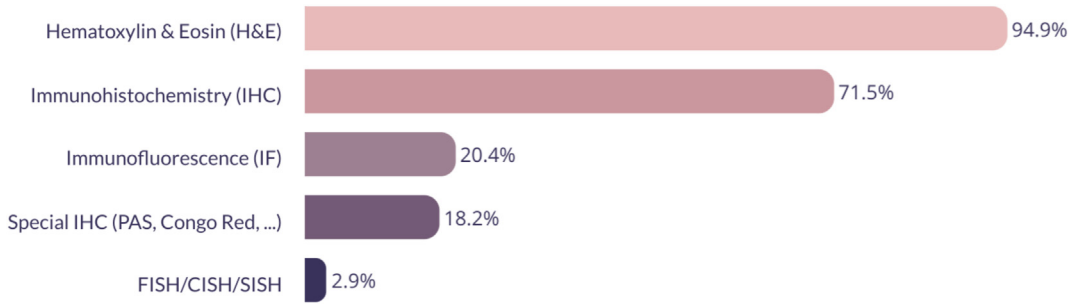

Figure 2.
Distribution of the preferred substrate for annotations among respondents. CISH, chromogenic in situ hybridization; FISH, fluorescence in situ hybridization; PAS, periodic acid Schiff; SISH, silver-enhanced in situ Hybridization.

product. Some respondents indicated that a minimum exposure time to the specific pathology subfield is required for adequate annotation quality (eg, at least 5 years). Some respondents made use of particular training on the annotation software/platform to be used (n = 13; 16.9%). Nine participants (11.7%) mentioned team meetings and reaching a consensus as another approach to standardizing the annotations. Other options included taking specific courses (n = 3), watching tutorials/PowerPoint presentations (n = 3), and following inhouse guidelines (n = 3). Less frequently mentioned options (2 or fewer respondents) were the following: "training from vendors," online resources and literature search, "Pathotrainer software from Pathomation," and international guidelines (eg, College of American Pathologists). Finally, 3 people did not use any specific training.

The projects of the vast majority of respondents focused on human pathology (n = 134 [97.8%] vs n = 16 [11.7%] in veterinary/toxicologic pathology), primarily reflecting the affiliation of study respondents. The projects were commonly based on hematoxylin and eosin–stained histologic slides (n = 130; 94.9%), with, however, a substantial number of projects involving immunohistochemistry (IHC) (n = 98; 71.5%), immunofluorescence (n = 28; 20.4%), or other forms of slide preparation (detailed in Fig. 2).

The Annotation Process

Annotation Substrate and Objects

The majority of respondents and their teams worked directly with whole-slide images (WSIs) (n = 100; 73%), with only 37 (27%) working with image patches (Fig. 3).

The annotations might be performed at different levels, for example, at the slide level (slide labels; eg, tumor present or not), tissue level (eg, tumor tissue, stromal tissue, mucosa, etc), and single-cell level (Fig. 4). The most frequently mentioned annotation approach was at the tissue level (n = 117; 85.4%), such as for training algorithms involving segmentation of different tissue types, for example, for diagnostic purposes. Around half of the participants (n = 65; 47.4%) worked at only 1 level of analysis, that is, only the tissue level (n = 46; 33.6%), cell level (n = 18; 13.1%), or slide level (n = 1; 0.7%). The annotation objects of the remaining half of the respondents (n = 72; 52.6%) were variable, with 6 respondents (4.4%) reporting other objects than mentioned above, such as artifacts within a slide or asbestos bodies.

Regarding the annotation detail (using a scale ranging from 1, ie, imprecise/rough annotations, up to 5, ie, very precise annotations with well-defined contours and without class overlap), most participants (n = 90; 65.7%) stated that the annotations were very detailed (4 or 5). Only 12 (8.8%) of the participants referred to a low level of detailing while annotating for their projects (1 or 2).

For a typical histologic slide included in a project (eg, tissue sample size, 15 x 15 mm or more), only 8 (5.8%) of the participants stated that they were annotating 100% of the tissue, with the majority (n = 109; 79.6%) reporting less than 50% tissue coverage. Moreover, for 28 (20%) respondents, the decision on the annotation area ranging from smaller regions of interest up to the entire WSI was met based on the following factors: the specific question of the project (n = 12; 43%), annotation objects (eg, normal tissue vs tissue with pathologic changes; n = 4 [14%]), sample size (eg, surgical resection specimen vs biopsy/cytology; n = 6 [21%]), or other characteristics of the sample (such as stain type or quality; n = 1 [4%]). According to these respondents, the final decision regarding the level of annotation relied on a combination of these factors.

Annotation Effort

Most participants (n = 103; 75.2%) replied that typically, 1 project includes up to 500 WSI/patches to be annotated, with the trend being independent of the type of annotation source (WSI vs patches), as shown in Figure 4.

Based on the free text input answers received (n = 21; 15%), the following different factors influenced the number of WSI/patches annotated per project: the project type/question (n = 10), the sample size (n = 2), the software/platform used (n = 1), or the availability of slides/cases for the specific project (n = 1) with the remainder (n = 7) advocating a combination of the reasons mentioned above.

The time necessary to complete the annotation of 1 project sample, according to respondents, ranged from an average of less than 15 minutes (n = 46; 33.1%), 15 to 30 minutes (n = 38; 27.9%), 30 to 60 minutes (n = 31; 22.8%), 1 to 3 hours (n = 12; 8.8%), or more than 3 hours (n = 5; 3.7%), with 5 (3.7%) participants reporting a variable timeline (n = 5; 3.7%). The majority of projects (n = 100; 73%) included less than 5 annotation categories/classes, 27 (19.7%) included 5 to 10 categories, and 10 (7.3%) respondents indicated using more than 10 classes.

Annotation Software and Hardware

The most commonly used software was QuPath (https://qupath.github.io/)[10] (n = 80; 59.3%), followed by ImageScope from Leica (n = 31; 23%), HALO from Indica Labs (n = 18; 13.3%), and Cytomine[11,12] from Cytomine Corporation (n = 16; 11.9%).

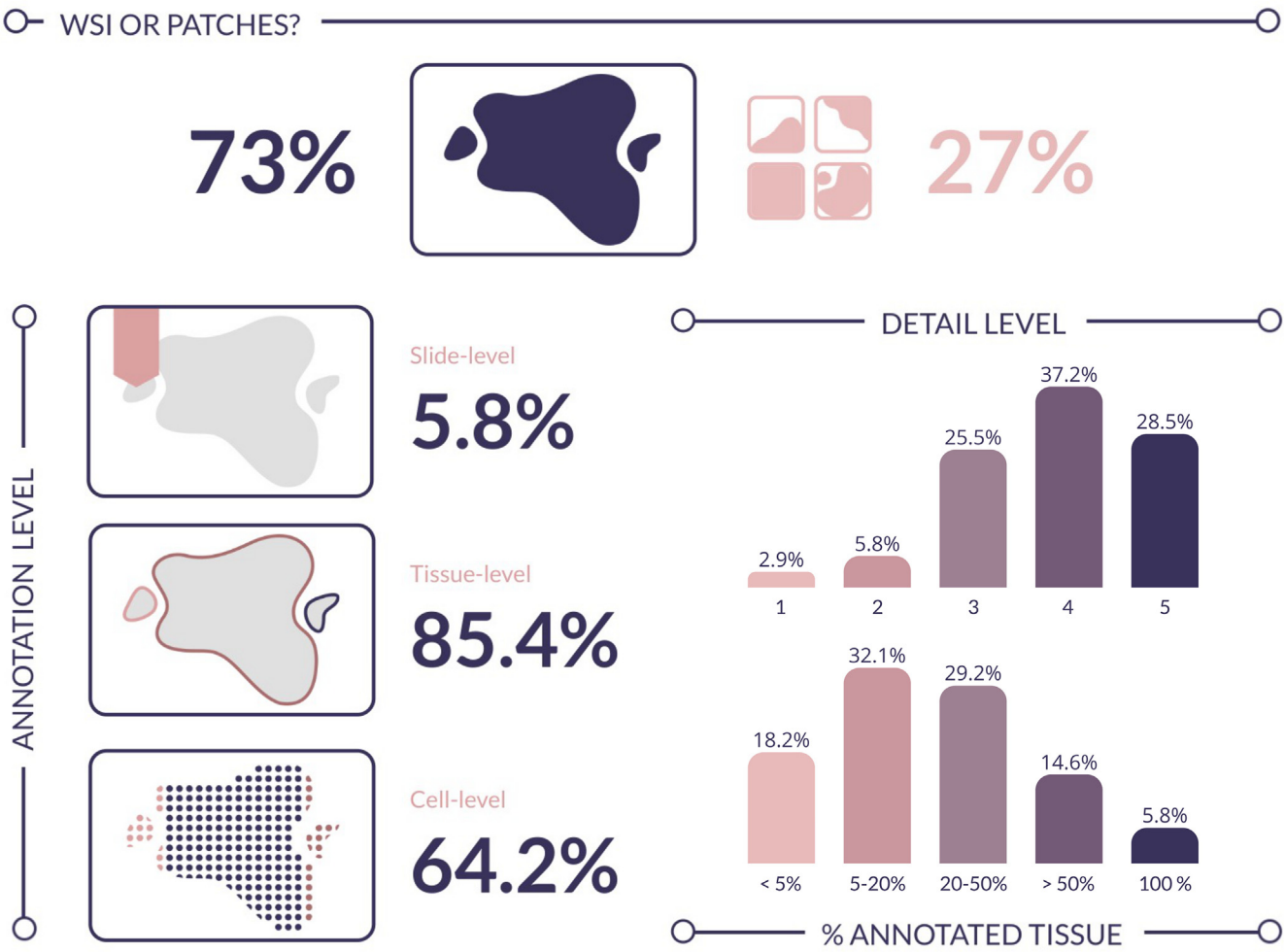

Figure 3.
Distribution of participants' preferences in percentages regarding annotation sources (WSI vs patches) and detail level (slide level, tissue level, and cell level) among participants. The upper bar chart on the right side of the graphic demonstrates the increasing complexity of annotation detail (from imprecise/rough annotation, corresponding to 1, to very precise with well-defined contours and without class overlap, corresponding to 5). WSI. whole-slide image.

Sixteen (11.9%) respondents used inhouse developed solutions (Fig. 5). Just over half of the participants (n = 70; 51.9%) used more than 1 annotation software type in their projects. A detailed list of the software mentioned, along with reference links and relative usage percentages, is reported in Supplementary Material 3.

Concerning the hardware used for annotation, a large majority of respondents (n = 134; 97.8%) used a computer mouse (conventional, gaming, or other, such as ergonomic or vertical), 39 (28.5%) used a pen (either on a digital screen or on a drawing pad), and 11 (8%) performed the annotations on a tablet. Thirty-five (25.5%) respondents mentioned using more than 1 option (Fig. 6).

Further feedback on hardware recommendations/preferences was provided by 34 (24.8%) participants. Nineteen (55.9%) recommended using a pen on a digital screen (mostly Wacom tablets), and 7 (20.1%) suggested using a gaming or ergonomic mouse. Other recommendations to optimize the annotation process were using both the mouse and the pen, having a high-quality and large screen, using shortcuts on the keyboard, and choosing the right software, depending on the task.

Sixty-eight participants (50%) did not use any kind of automation for annotations, 15 (11%) used QuPath-based algorithms, 13 (8.8%) used custom algorithms combined with the annotation software, and 5 (3.7%) used commercial or other open-source tools, such as QuickAnnotator (open source; https://github.com/choosehappy/QuickAnnotator), NuClick (open source; https://github.com/navidstuv/NuClick), or Aiforia's annotation assistant (commercial). Fourteen (10.3%) respondents used other options for annotation automation, and 22 (16.2%) used automation only for selected projects. Moreover, 29 participants (21.2%) gave additional details on their experience with automated tools (Supplementary Material 2-Table S2.2). Thirteen of these emphasized the use of software tools that facilitate the annotation effort (such as QuPath magic wand/brush; HALO's option that allows converting the tissue masks into annotations; Patchsorter (https://github.com/choosehappy/PatchSorter) scheme to annotate "large numbers of objects with similar features simultaneously"; or Aiforia's annotation assistant). Ten respondents mentioned using pretrained algorithms or active learning methods to perform the annotations (that were subsequently adjusted/corrected). Three respondents also cited the Segment Anything Model[13,14] extension for QuPath as being useful, as well as Otsu's thresholding for foreground segmentation. The technique of restaining hematoxylin and eosin stains with an IHC marker with further registration of the images to generate

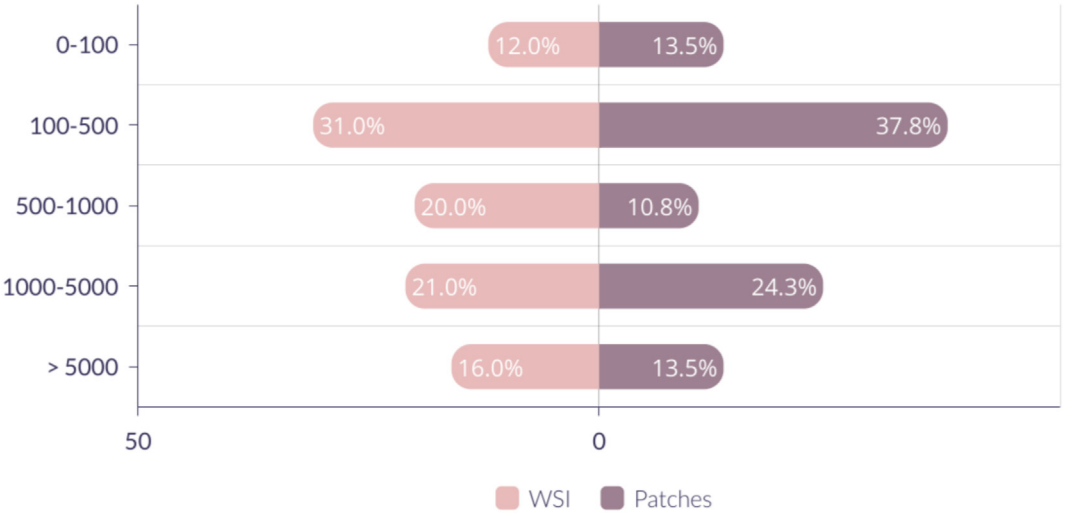

Figure 4.
The number of images annotated in a typical project by participants (%) either using WSI or patches. WSI. whole-slide image.

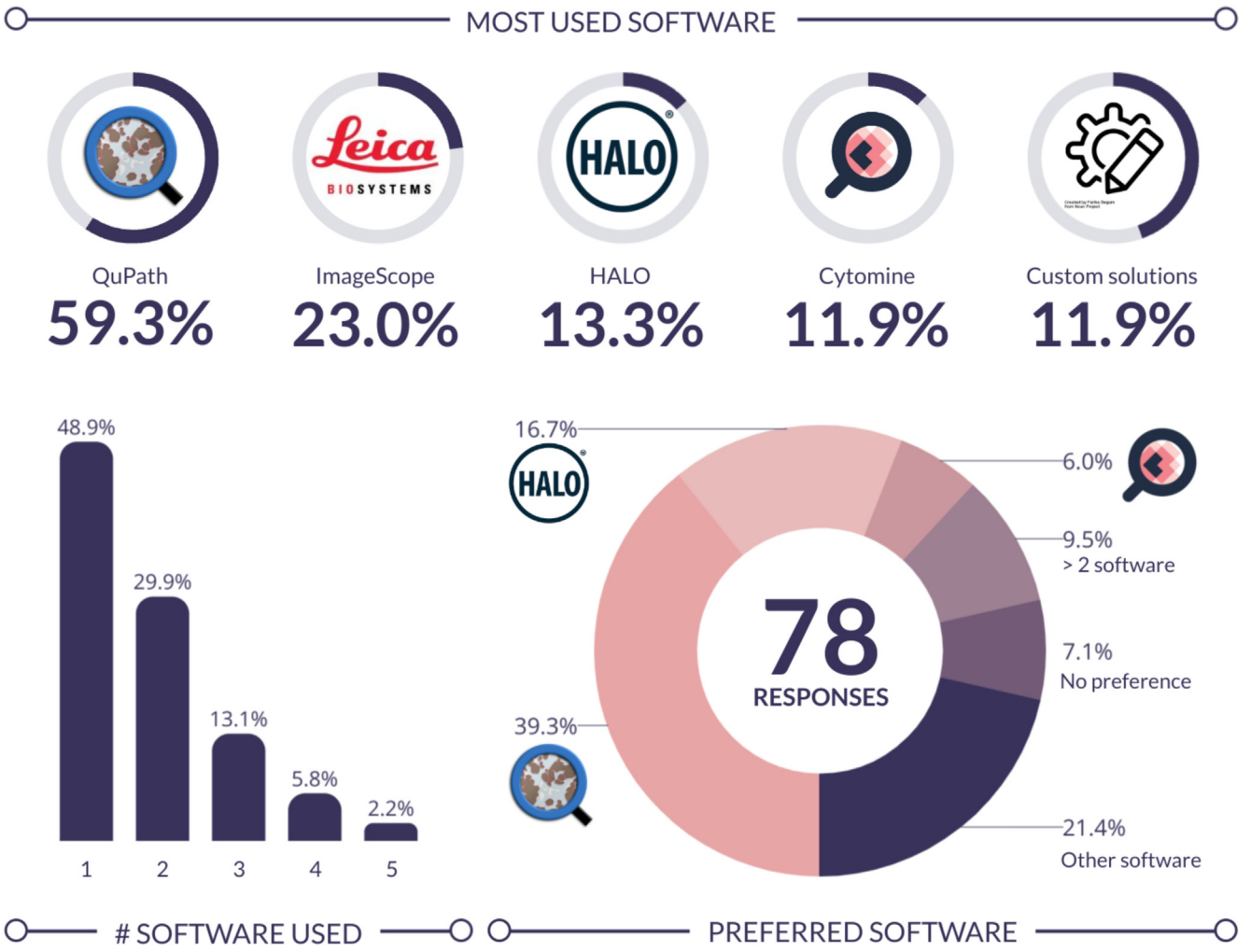

Figure 5.
In the upper panel, the most used software for annotation by participants (%) are shown. In the bottom left panel, the number of software used (from 1-5 different software) is shown. In the bottom right panel, the distribution in percentage of preferences on the annotation software among 78 participants who answered the dedicated free text input question.

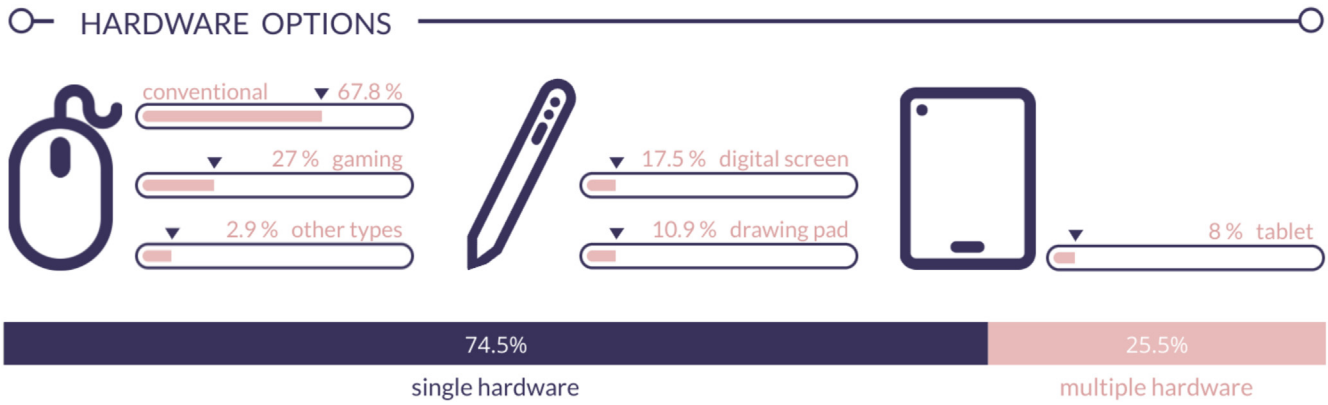

Figure 6.
Preferences of participants on the hardware used for the annotation process. On the left are statistics for the mouse (conventional, gaming, or other specific types) used, followed by those related to using pens on either digital screens or drawing pads (in the middle), and tablet usage on the right.

automatic annotations was also mentioned as helpful. A general comment was that the automation methodology might depend on the project.

Image and Annotation File Formats

The most common WSI file format used was TIFF/TIF (n = 94; 69.6%), followed by SVS (n = 79; 58.5%), MRXS (3DHISTECH) (n = 54; 40%), and NDPI (Hamamatsu) (n = 41; 30.4%), as depicted in Figure 7. Twenty respondents (14.8%) mentioned other slide formats such as iSyntax (Philips), OME.TIF, or CZI (Zeiss). Only 16 (11.0%) indicated using the DICOM file format. Two participants (1.5%) did not know the image file format used in their projects.

This general trend is confirmed in the question on the WSI file format preferred by annotators, with TIF, or its variants (n = 31; 22.6%), and SVS (a TIFF variant from Leica; n = 27; 19.7%) still receiving the majority of votes. DICOM was preferred by 11 respondents, who mentioned that it would be beneficial to have this format to be more widely disseminated, in contrast to proprietary formats. These were followed by NDPI, MRXS, JPG, and PNG (favored by 6, 5, 4, and 3 participants, respectively). Two participants also mentioned other formats (namely OME-NGFF, a next-generation file format for bioimaging data,[15] and VSI from Olympus). Some of these formats were cited for their relatively lower "usability" and reliability ("sometimes working and sometimes not"), especially in terms of readability by software such as QuPath and in file transfer process (eg, MRXS for 15 participants and SVS, iSyntax, CZI, and TIF for other 12 participants). Forty-five respondents (32.8%) answered they did not know or did not have a preference.

For storing ready-to-use annotations, the most commonly used file format was GeoJSON (n = 35; 40.2%), followed by XML (n = 33; 37.9%), JSON (n = 25; 28.7%), and QPDATA (n = 22; 25.3%). Other answers (n = 17; 10.5%) included PNG, TIF, or CSV files to store the annotations (Fig. 7). Fifty participants (36.5%) did not know the format in which the annotations were exported in their projects.

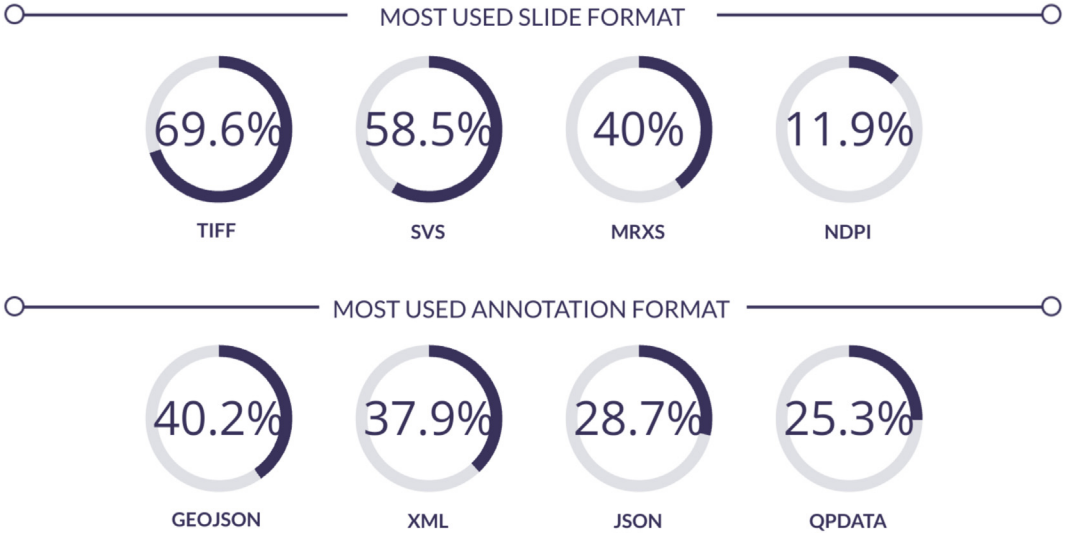

Figure 7.
Percentages of the most used slide formats (upper panel) and the most used annotation formats (lower panel).

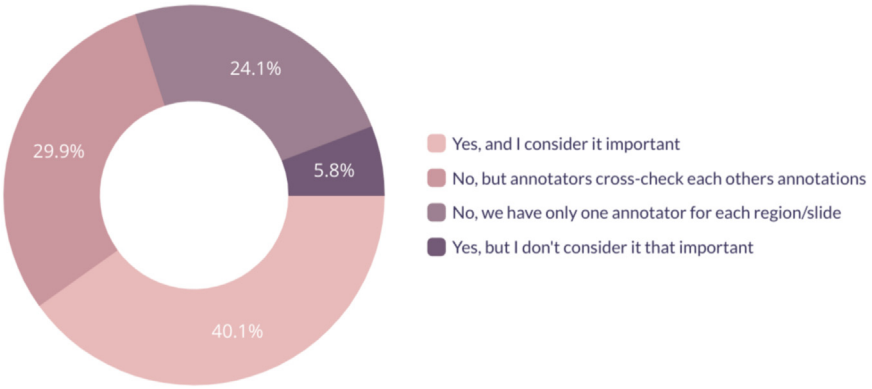

Figure 8.
Distribution of the participants' preferences in percentages on using multiple sources for ground truth.

Annotation Guidelines and Quality Control

The vast majority (n = 119; 86.9%) of the respondents defined instructions for annotations, with 83 (60.6%) mentioning the use of structured guidelines and 36 (26.3%) defining at least some standards. Nine (6.6%) respondents noted that the details of such instructions depended upon the specific project.

We received details on quality control practices from 136 participants (Supplementary Materials 2-Table S2.3). Thirty-three (24.3%) participants had no quality control (or were unsure) for the annotation process. Twenty (14.7%) performed quality control but did not provide any details on their method. For most groups, quality control consisted of a cross-check of the annotations by an independent analyst, for example, by a pathologist/expert (n = 35; 25.7%; some cases or all) or by other annotators (n = 19; 14%). Twenty respondents (14.7%) mentioned that the cases were discussed and reviewed in a team or that they had multiple annotators working on the same slides/regions of interest (with a final comparison of annotations). Three respondents (2.2%) stated that the quality control was performed by the same person performing the annotations (eg, reviewing the slides but not cross-checking). Thirteen teams (9.6%) used statistical analysis or engineering strategies to perform quality control (eg, using outlier results to detect incorrect annotations, concordance measures to select the best annotations, using software to check for conflicting annotations, etc).

Ground-Truth Sources

Regarding the use of multiple sources for ground truth, the responses were fairly evenly split (Fig. 8), with 63 (46%) of participants using multiple annotations and 74 (54%) using single annotations. Among the latter, 41 (55.4%) mentioned that annotators cross-checked annotations made by others.

On the decision of the subject and detail of the annotation, 133 participants gave additional information, and in most cases (n = 69; 51.9%), this came from the team/multidisciplinary discussion (that could integrate different professionals, namely pathologists, clinical physicians, computer scientists, students, etc). In contrast, for the remaining 54 (40.6%) respondents, the decision was taken by a single person, which could be the project lead, principal investigator, pathologist, medical advisor, head of the laboratory, or annotator, especially in small teams. Eight (6%) people said it would depend on the project. Last, 4 participants added that this decision also had input from the "clients"/"vendors"/"customer" (in 1 case, the client was the sole responsible party for decision-making and, in the other 3, the clients were part of the team deciding the annotation protocol). One person replied that they did not know who was responsible for these decisions.

Lessons Learned

We asked about the most significant challenges regarding the annotation process and received inputs from all the 137 respondents. The time allocated to annotation was the most frequently mentioned issue (n = 103; 75.2%), followed by difficulty establishing a ground truth (n = 72; 52.5%). The lack of complete optimization of software and/or hardware for the annotation process (n = 69; 50.3%) and the lack of guidelines (n = 48; 35.0%) were also seen as important challenges. Five people mentioned "interobserver variability" as a difficulty, and 2 indicated that the annotation process itself could be a hurdle, as it could be "repetitive" and "very mechanical." Other less common (expressed by 2 or fewer respondents each) concerns included the difference in file formats to work with, size/quality of the slides, regulatory effort related to data to annotate, security restrictions that impair sharing cases with external observers, difficulties in communication between pathologists and other team members, as well as the need for more automation.

We also asked the participants (n = 130) what their suggestions would be to facilitate the annotation efforts and how they saw the evolution of the annotation process in the near future. The necessity for more automation to assist annotations ("annotation process should be facilitated by a semi-automatic approach," "implementing innovative annotator assistant tools") was the most frequently mentioned topic (n = 51; 39.2%), along with the need for more advanced software ("more reliable," "interoperable," and better accessible/open access, with user support) and hardware (n = 29; 22.3%). Many respondents also highlighted the need for internationally recognized guidelines/recommendations for annotations in computational pathology (n = 22; 16.9%) and specific training (n = 5). Eight respondents argued that pathologists/experts should be the annotators or at least be involved in the annotation process. Regarding the tips and advice to facilitate current annotation processes, some of the recommendations were as follows: to use segmentation tools such as the

Table 1
A summary of suggestions for annotators

Tips and advice to facilitate the current annotation process
• Use automatic segmentation tools, such as the SAM extension for QuPath. • Take advantage of software tools that aid annotation. • Establish a solid ground truth (foster teamwork and involve experts in the field). • Employ IHC, special stains, or molecular studies to enhance the ground truth or assist annotation. • Facilitate effective communication between ML experts and the clinical team. • Develop a comprehensive project roadmap, prioritizing defining desired outcomes and determining annotation requirements and strategies. • Engage pathologists in the annotation process while also leveraging nonpathologists for specific tasks. • Whenever possible, "employ multiple annotators and evaluate interobserver concordance." • Select user-friendly software options. • Opt for ergonomic hardware solutions. • Solutions such as a drawing table with a pen, a gaming/ergonomic mouse, and keyboard shortcuts can be helpful. • "Setting the stage": incorporate regular breaks into your schedule and set clear expectations, noting that annotations can consume significant time. Identify your most productive hours and structure your work around them. Ensure your workspace is ergonomic and create a favorable atmosphere with background music or other preferred ambience.
Recommendations to improve the annotation process in the future
• Availing enhanced software and hardware options for annotation tasks (emphasizing interoperability, accessibility, and support). • Implementing automated annotation processes (eg, "implementing innovative annotator assistant tools, such as SAM or others that can significantly shorten the annotation process"). • Leveraging foundation models. • Establishing internationally recognized guidelines/recommendations for annotation in computational pathology to increase reproducibility across different settings. • Implementing training programs focusing on annotation tasks. • Availing enhanced ground-truth accuracy (through the integration of IHC staining, combining scanning data with molecular insights, and other advanced techniques). • Enabling better definitions among pathologists/experts to refine ground truths. In addition, exploration of unsupervised methodologies and outcome-driven approaches instead of solely relying on pathologist-driven ground truths. • Increasing acceptance of computational pathology within the practicing pathology community. • Improving access to collective annotation data sets to foster collaboration and data sharing. • Optimizing the costs associated with digital pathology solutions to enable broader implementation of computational pathology projects. • Standardizing file formats, namely DICOM, to streamline data exchange and interoperability. • Enhancing image resolution and overall quality to ensure reliability in computational pathology analyses.

DICOM, digital imaging and communications in medicine; IHC, immunohistochemistry; ML, machine learning; SAM, segment anything model.

Segment Anything Model for "speeding up the annotation process"; to use "sparse annotations to speed up and avoid irrelevant annotations"; to use IHC/special stains or other molecular studies to improve the ground truth or to assist annotation; to involve pathologists in the process; to choose "user-friendly software"; and to select ergonomic and easy-to-use hardware (namely "gaming mouse" and keyboard shortcuts). These observations were confirmed in another dedicated question, namely on tips/recommendations for annotators, with replies from 36 participants. The responses highlighted additional insights, such as taking breaks while annotating, keeping focused, and having a favorable work environment (eg, "music and podcasts running in the background"). Table 1 summarizes the various suggestions received.

Discussion

The survey provided multifaceted details on participants' demographic profile, their professional settings, and the technologies used for annotations. Regarding software, QuPath[10] was the most widely used and preferred choice. Respondents highlighted its perceived advantages of open access, the possibility of adding third-party extensions, various tools to facilitate annotation, an active support forum, and the possibility of community-driven development.[10,16] The survey data highlight the diverse array of hardware preferences among participants, reflecting the variety of devices available to navigate virtual slides and annotate images.[17] The findings underscore the importance of accommodating diverse hardware preferences and optimizing annotation tools to enhance productivity and user experience in computational pathology annotation tasks.[8]

According to the responses from 135 participants, the preferred image format was TIF and its variants, followed by SVS. This is likely attributable to the broader use of these formats consequent to a longer stay in the field of computational pathology, as well as their user-friendly nature.[18] On the other hand, different reasons have been advocated for the lower usage of some file formats alternative to TIF and SVS, such as the relatively lower readability by common software or the greater complexity in the file transfer phase. The need to standardize image format seems paramount in advancing computational pathology, similar to radiology/imaging, in which the DICOM format has greatly facilitated data interoperability, collaboration, and the development of advanced imaging analysis algorithms.[19,20] Pathology is still lagging in the standardization process, which is demonstrated by the multiple file formats being used. Although some of the participants advocated for broader implementation of the DICOM format, in line with some authors,[21] an alternative possible solution to promote standardization could also be the adoption of software and viewers able to manage different file formats, similar to already mentioned QuPath that profits from OpenSlide (https://openslide.org/) and Bio-Formats (https://www.openmicroscopy.org/bio-formats/) libraries. This would be in line with the trend of adopting format and vendor-neutral solutions to promote interoperability in digital and computational pathology.[22] Compared with the image format, the awareness about which was high among participants, there was less knowledge regarding the annotation file format. This unawareness could be due to the high proportion of pathologists in the survey. Understanding the annotation file format is useful to ensure compatibility across tools, maintain data integrity, and support the long-term accessibility and usability of the research findings. This underscores the need for a more integrated interdisciplinary discussion with

Table 2
Recommendations to address critical needs in computational pathology

Recommendations for the CPath community

• Establishing guidelines and recommendations
  • Internationally recognized guidelines for standardized annotation processes are needed to enhance reproducibility across different clinical and research environments; eg, annotation protocols, quality standards, and best practices to improve consistency and reproducibility across different projects and settings.
• Enhancing data set sharing
  • Expand access to collective annotation data sets to widen collaborative research and innovation; eg, crowdsourced annotation schemes or multimodal approaches.
• Establishing dedicated training programs
  • Create specialized training modules on annotation techniques, ensuring that both pathologists and computational experts are well prepared to meet high-quality annotation standards; eg, workshops on technical aspects of software/platforms for annotation, domain-specific courses in pathology, or hybrid programs for annotating teams.
• Promoting CPath acceptance
  • Encourage multidisciplinary communication and acceptance of CPath within the practicing pathology community; eg, workshops on technical aspects of software/platforms for annotation, domain-specific courses in pathology, or hybrid programs for annotating teams.
• Improving ground truth accuracy
  • Promote research and collaboration for consistent definitions; eg, consider integrating advanced methods (eg, IHC and molecular insights) to refine ground-truth standards, crowdsourced annotation schemes, or multimodal approaches.

Recommendations for the technology developers/industry stakeholders

• Software and hardware improvements
  • Interoperability, accessibility, and support in annotation tools should be prioritized; eg, improved software interfaces or adapted tools for devices such as tablets.
• Improving annotation tools
  • Further develop AI-driven and automation-supported annotation tools to streamline the annotation process. Foundation models can be leveraged to assist annotation tasks; tailored segmentation algorithms or annotator assistance tools.
• File standardization
  • Implementing universal file standards to facilitate seamless data exchange and optimize interoperability.
• Improving cost efficiency and accessibility
  • Designing cost-effective software and hardware solutions to expand access to CPath tools, enabling broader adoption.
• Advancing image quality standards
  • Enhancing image resolution and overall quality to ensure reliability in CPath analyses.

Cpath, computational pathology; IHC, immunohistochemistry.

medical and computer science professionals, promoting the exchange of information and knowledge to improve the post-annotation phases of the projects.

The prevalence of participants employing quality assessment on the annotation process, especially through review by pathologists, suggests reliance on collaborative validation processes to ensure dependability in annotation tasks. However, it is notable that a significant proportion of respondents reported no quality control measures or did not describe their methods, highlighting potential gaps in quality assurance practices within the field.[23,24] In addition, a smaller group employed statistical or engineering strategies, underscoring the diversity of approaches employed to mitigate errors and enhance the quality of annotations.[25]

Although the majority of participants indicated that decisions regarding what and how to annotate were made through multidisciplinary collaboration, a not negligible subset of respondents stated that this responsibility fell upon a single individual, such as the project lead or a pathologist. Moreover, a few respondents mentioned project dependency or input from clients/vendors, mirroring the relatively low percentage of participants annotating for commercial projects. These results highlight the diverse decision-making structures in annotation teams within computational pathology, with some favoring collaborative approaches and others relying on a single individual/expert authority or project-specific considerations. It appears from many of the replies that the teams involved in the annotation are small, which may impair true multidisciplinary cross-communication. Together with the lack of standards and guidelines, this represents a significant hurdle toward robust and standardized annotations.

The primary challenges in the annotation process mentioned by the participants included time constraints, difficulty establishing a ground truth, suboptimal software/hardware, the absence of standardized guidelines, and concerns regarding interobserver variability. Moreover, the time-consuming nature of the annotation process was also mentioned. The reliability/reproducibility of pathologists' ground truth has been a matter of debate for the machine learning community.[26] In fact, inter- and intraobserver variability is inherent to the assessment process in pathology. Interestingly, some authors argue that, despite these limitations, there might not be more reliable sources than the pathologists' evaluations for diagnostic/classification purposes.[26] In the same way as pathologists deal with uncertainty (requesting a second opinion from colleagues, performing ancillary tests, etc), different approaches have been suggested to limit subjectivity in the ground-truthing process for medical applications to increase the number of graders in each case, to recruit expert evaluators/panels, and to ensure an unbiased resolution method when graders disagree.[27] These strategies should be (need to be) implemented when establishing an annotation project in pathology.

Multiple respondents have stated that it would be helpful to have internationally recognized guidelines/recommendations for annotation in computational pathology to increase reproducibility across different settings. In addition, the need for better training in annotation tasks, in the form of workshops/courses, was also mentioned by various participants in the survey. These statements and results of the study should serve as a basis for such developments. To facilitate this, we propose 7 actionable areas for addressing critical needs in the annotation process in computational pathology (Table 2), aiming to mobilize the community toward collaborative efforts and impactful improvements in the field.

This study has some limitations. One is the limited number of respondents, with most being European, which may not fully capture global annotation practices. Although substantial efforts were made to recruit a diverse sample, increasing participation further remains a challenge that could impact the generalizability of our

findings. Expanding participation in future studies, particularly by engaging more institutions and international societies, would be valuable to ensure broader applicability and enrich our understanding of computational pathology annotation practices worldwide. In addition, most respondents were involved in the annotation but were not the annotators themselves, which may hinder valuable information. Last, many replies are free text input and subjective (with some answers contradicting others). Importantly, preferences regarding software/hardware reflect participants' choices, and this work does not aim to entail a comprehensive description of all available options. Finally, the current study focuses on providing a snapshot of existing practices rather than tracking changes over time, reflecting current annotation approaches. Moving forward, we intend to investigate how these practices are evolving, as this remains a critical area of interest, particularly in light of the rapid advancements in computational pathology. Despite these limitations, this survey indicates the nature of annotation practices and the diverse perspectives on the topic.

The results of the present survey demonstrate a heterogeneous landscape of annotation practices with minimal standardization in computational pathology. This is likely to impact the performance, quality, and generalizability of AI models. Key questions remain, such as how to standardize file formats to improve interoperability, how to enable wide annotated data sharing, and how to develop effective guidelines. Moving forward, the ESDIP recommends specific actionable areas that can be addressed, namely prioritizing the implementation of targeted training initiatives and encouraging interdisciplinary collaboration, to promote the progressive standardization of the annotation process. This will not only enhance the accuracy and consistency of computational pathology but also ultimately advance innovation and improve patient outcomes.

Acknowledgments

Participants who agreed to be mentioned are acknowledged and listed in Supplementary Material 4. This list includes only participants who agreed to be mentioned. Participants who chose to be mentioned but did not provide the necessary information (name and affiliation) were contacted by email. Those who did not reply are not included.

Author Contributions

D.M., S.P.O., and V.L. conceived and designed the study; developed the methodology of the study and wrote the original draft; and performed acquisition, analysis, and interpretation of data, and statistical analysis. P.B., A.H., R.C., V.D.M., T.-R.K., S.L., M.Y., D.A., M.-S.S., N.Z., and Y.T. reviewed and revised the paper. N.Z. and Y.T. provided technical and material support. All authors read and approved the final paper.

Data Availability

The data sets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Funding

The authors received no specific funding for this work.

Declaration of Competing Interest

D. Montezuma is an employee at IMP Diagnostics. D. Ameisen is a shareholder of ImginIT SAS.

Ethics Approval and Consent to Participate

The present study did not require ethical approval or consent to participate.

Supplementary Material

The online version contains supplementary material available at https://doi.org/10.1016/j.labinv.2024.102203

References

1. Bulten W, Kartasalo K, Chen PHC, et al. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nat Med*. 2022;28:154–163.
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.
3. Pantanowitz L, Hartman D, Qi Y, et al. Accuracy and efficiency of an artificial intelligence tool when counting breast mitoses. *Diagn Pathol*. 2020;15:80.
4. 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.
5. 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.
6. 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.
7. Elfer K, Gardecki E, Garcia V, et al. Reproducible reporting of the collection and evaluation of annotations for artificial intelligence models. *Mod Pathol*. 2024;37(4):100439. https://doi.org/10.1016/j.modpat.2024.100439
8. Montezuma D, Oliveira SP, Neto PC, et al. Annotating for artificial intelligence applications in digital pathology: a practical guide for pathologists and researchers. *Mod Pathol*. 2023;36(4):100086.
9. Lindman K, Rose JF, Lindvall M, et al. 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.
10. Bankhead P, Loughrey MB, Fernández JA, et al. QuPath: open-source software for digital pathology image analysis. *Sci Rep*. 2017;7(1), 16878.
11. Marée R, Rollus L, Stévens B, et al. Collaborative analysis of multi-gigapixel imaging data using Cytomine. *Bioinformatics*. 2016;32(9):1395–1401.
12. 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.
13. Kirillov A, Mintun E, Ravi N, et al. Segment anything. Preprint. Posted online April 5, 2023. arXiv 2304.02643. https://doi.org/10.48550/arXiv.2304.02643
14. Sugawara K. Training deep learning models for cell image segmentation with sparse annotations. Preprint. Posted online June 13, 2023. bioRxiv 2023.06.13.544786. https://doi.org/10.1101/2023.06.13.544786
15. Moore J, Allan C, Besson S, et al. OME-NGFF: a next-generation file format for expanding bioimaging data-access strategies. *Nat Methods*. 2021;18:1496–1498.
16. Kaczmarzyk JR, O'Callaghan A, Inglis F, et al. Open and reusable deep learning for pathology with WSInfer and QuPath. *NPJ Precis Oncol*. 2024;8:9.
17. Mateos EA, Caballero-Alemán F, Albarracín-Ferrer M, et al. Research on devices for handling whole slide images on pathology workstations. An ergonomic outlook. *Diagn Pathol*. 2016;2(1):232.
18. Pantanowitz L, Sharma A, Carter AB, Kurc T, Sussman A, Saltz J. Twenty years of digital pathology: an overview of the road travelled, what is on the horizon, and the emergence of vendor-neutral archives. *J Pathol Inform*. 2018;9:40.
19. Larobina M. Thirty years of the DICOM standard. *Tomography*. 2023;9(5):1829–1838.
20. Herrmann MD, Clunie DA, Fedorov A, et al. Implementing the DICOM standard for digital pathology. *J Pathol Inform*. 2018;9:37.
21. Clunie DA. DICOM format and protocol standardization-a core requirement for digital pathology success. *Toxicol Pathol*. 2021;49(4):738–749.
22. Romberg D, Strohmenger K, Jansen C, et al. EMPAIA app interface: an open and vendor-neutral interface for AI applications in pathology. *Comput Methods Programs Biomed*. 2022;215:106596.

23. Chen Y, Zee J, Smith A, et al. Assessment of a computerized quantitative quality control tool for whole slide images of kidney biopsies. *J Pathol*. 2021;253(3):268–278.
24. Rajaganesan S, Kumar R, Rao V, et al. Comparative assessment of digital pathology systems for primary diagnosis. *J Pathol Inform*. 2021;12:25.
25. Schömig-Markiefka B, Pryalukhin A, Hulla W, et al. Quality control stress test for deep learning-based diagnostic model in digital pathology. *Mod Pathol*. 2021;34(12):2098–2108.
26. Gonzalez R, Saha A, Campbell CJV, Nejat P, Lokker C, Norgan AP. Seeing the random forest through the decision trees. Supporting learning health systems from histopathology with machine learning models: challenges and opportunities. *J Pathol Inform*. 2023;15:100347.
27. Chen PHC, Mermel CH, Liu Y. Evaluation of artificial intelligence on a reference standard based on subjective interpretation. *Lancet Digit Health*. 2012;3(11):e693–e695.