MODERN PATHOLOGY

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

Journal homepage: https://modernpathology.org/

Review Article

Built to Last? Reproducibility and Reusability of Deep Learning Algorithms in Computational Pathology

Sophia J. Wagner[a,b], Christian Matek[c,d], Sayedali Shetab Boushehri[b,c,e], Melanie Boxberg[f,g], Lorenz Lamm[a,h], Ario Sadafi[b,c], Dominik J.E. Winter[c,i], Carsten Marr[c,*], Tingying Peng[a,*]

[a] Helmholtz AI, Helmholtz Munich—German Research Center for Environmental Health, Neuherberg, Germany; [b] School of Computation, Information and Technology, Technical University of Munich, Garching, Germany; [c] Institute of AI for Health, Helmholtz Munich—German Research Center for Environmental Health, Neuherberg, Germany; [d] Institute of Pathology, University Hospital Erlangen, Erlangen, Germany; [e] Data & Analytics (D&A), Roche Pharma Research and Early Development (pRED), Roche Innovation Center Munich, Germany; [f] Institute of Pathology, Technical University Munich, Munich, Germany; [g] Institute of Pathology Munich-North, Munich, Germany; [h] Helmholtz Pioneer Campus, Helmholtz Munich—German Research Center for Environmental Health, Neuherberg, Germany; [i] School of Life Sciences, Technical University of Munich, Weihenstephan, Germany

ARTICLE INFO

Article history:
Received 10 February 2023
Revised 2 October 2023
Accepted 3 October 2023
Available online 10 October 2023

Keywords:
artificial intelligence
computational pathology
deep learning
histology/histopathology
reproducibility
reusability

ABSTRACT

Recent progress in computational pathology has been driven by deep learning. While code and data availability are essential to reproduce findings from preceding publications, ensuring a deep learning model's reusability is more challenging. For that, the codebase should be well-documented and easy to integrate into existing workflows and models should be robust toward noise and generalizable toward data from different sources. Strikingly, only a few computational pathology algorithms have been reused by other researchers so far, let alone employed in a clinical setting. To assess the current state of reproducibility and reusability of computational pathology algorithms, we evaluated peer-reviewed articles available in PubMed, published between January 2019 and March 2021, in 5 use cases: stain normalization; tissue type segmentation; evaluation of cell-level features; genetic alteration prediction; and inference of grading, staging, and prognostic information. We compiled criteria for data and code availability and statistical result analysis and assessed them in 160 publications. We found that only one-quarter (41 of 160 publications) made code publicly available. Among these 41 studies, three-quarters (30 of 41) analyzed their results statistically, half of them (20 of 41) released their trained model weights, and approximately a third (16 of 41) used an independent cohort for evaluation. Our review is intended for both pathologists interested in deep learning and researchers applying algorithms to computational pathology challenges. We provide a detailed overview of publications with published code in the field, list reusable data handling tools, and provide criteria for reproducibility and reusability.

Introduction

Technical progress has been driving digitization in pathology over the past decade. Coupled with advances in deep learning (DL) methods, computational approaches help to localize, segment, and classify single cells and tissue types in an automated manner

These authors contributed equally: Sophia J. Wagner and Christian Matek.
* Corresponding authors.
E-mail addresses: tingying.peng@helmholtz-munich.de (T. Peng), carsten.marr@helmholtz-munich.de (C. Marr).

ELSEVIER
0893-3952/$ - see front matter © 2023 United States & Canadian Academy of Pathology. Published by Elsevier Inc. All rights reserved.
https://doi.org/10.1016/j.modpat.2023.100350

Table 1
Glossary for commonly used terms in computational pathology

Digital pathology	Histologic slides are scanned and digitized so pathologists can examine the patient material on a screen instead of working on optical microscopes. Digitized slides can be stored and processed, enabling the use of computational methods in the diagnostic process.
Computational pathology	The analysis of digitized histologic slides with computational methods.[1]
Specimen	A tissue sample, for example, obtained during a biopsy or other surgical procedures, typically fixed in formalin and embedded in paraffin.
Section	A thin slice (with a typical thickness of 3-15 μm) of a specimen mounted on a microscopic slide. WSI: The digitized image of a tissue section on a microscope slide. Slides can be scanned in very high magnification resulting in images of sizes up to several gigapixels.
Patches and tiles	WSIs are split up into smaller images (eg, 512 × 512 pixels), also called patches or tiles, that can be processed by neural networks. Unlike WSIs, these smaller image data units allow for easier and parallelized image processing. The terms “tiles” and “patches” are often inconsistently used. Overall, “patches” is the more common term, but if the smaller images are used for subsequent WSI rendering by stitching them back together, they are sometimes called tiles.
Annotations	Diagnostic information on pixel or patch levels is obtained from manual expert pathologist labeling at different levels of resolution (eg, tumor regions are outlined at 2× to 5× while cell-level annotations are performed at 20× to 40×). WSI-level annotations can be all diagnostic information about the patient (eg, age, survival, staging, and grading) mainly obtained without additional expert pathologist interaction.
Supervised learning	Training procedure of a neural network, where the ground truth, that is, the correct label for the task, is available for each data point. However, in medical imaging and especially computational pathology, full expert annotations at pixel or patch level are time-consuming and, hence, rare. Pixel-level annotations are used to localize tissues in segmentation tasks, where each pixel is assigned a tissue label. Patch-level annotations are used for classification tasks, where one label is predicted for the entire input patch.
Weakly supervised learning	Due to the rareness of fully annotated WSIs, weakly supervised learning approaches, like MIL, are often used to train neural networks. With MIL, only WSI-level annotations, such as diagnostic information on the cancer type or survival, are required for classification.
CNN	A neural network that can be trained to extract features by sliding trainable filters across the image. This makes CNNs translationally invariant and, therefore, well suited for histologic data because relevant features can be found anywhere on a tile.
U-Net	A CNN with an encoder-decoder architecture for segmenting biomedical images.[2] It was adapted in many ways and ranks among the most common architectures for segmentation tasks.
Mask R-CNN	A CNN architecture for instance segmentation in object detection.[3] In contrast to U-Net, which does not distinguish between instances of a class, Mask region-CNN (Mask R-CNN) outputs a segmentation mask for each instance on the image, which makes it useful for tasks such as nuclei segmentation and cell counting.

CNN, convolutional neural network; MIL, multiple instance learning; WSI, whole slide image.

and form the research field of computational pathology[1] (see Table 1 for a glossary[1-3]). In particular, deep neural networks have recently been shown to reach the performance level of medical experts on well-defined tasks such as skin cancer diagnosis,[4] lung cancer subtype classification,[5] or the recognition of malignant white blood cells.[6]

However, despite the steady increase in publications in this field (Fig. 1) and their promising results, only a few have reached clinical implementation.[7,8] This is due to several reasons: for DL-based methods, code availability is a natural requirement for reproducibility and, in particular, reusability, which, unfortunately, is not yet current practice for most publications. Even when code is available, reproducing the original results can be challenging and requires the assistance of the original author.[9] In particular, ready-to-use scripts with sufficient instructions or intuitive demo examples are rarely published. This makes the reuse of current methods difficult for non-DL experts, especially for pathologists who are not supported by computational experts. Another reason, particularly relevant to clinical implementation, is the generalization gap of algorithms in computational pathology. Often, the published performance of DL algorithms cannot be transferred to other data sets due to differences in staining or scanner settings. Therefore, external validation of algorithms and statistical robustness analysis are essential to assess generalizability. Finally, algorithms marketed for use in a clinical setting must additionally be approved by national or international authorities such as the US Food and Drugs Administration or the European Medicines Agency, an often lengthy and complicated process involving business markets and legal issues, which is beyond the scope of this review.

Here, we focus on DL algorithms for computational pathology and their reproducibility and reusability. For the ultimate goal of *reusing* DL algorithms, the algorithms must be *reproducible* and generalizable to similar data sets (ie, *robust*) and external data sets (ie, *replicable*) (see Table 2[9-13] and Fig. 2 for definitions of reproducibility, reusability, and related terms). Because hematoxylin

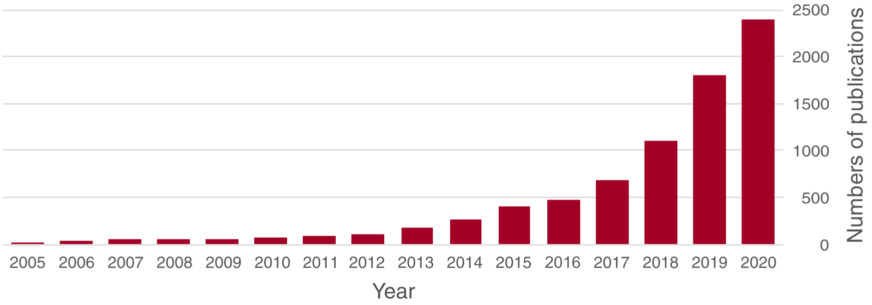

Figure 1.
The number of articles published in the field of computational pathology in PubMed (retrieved on July 22, 2021) has markedly increased in the past 15 years.

Table 2

Definitions of reproducibility, reusability, and related terms

Reproducibility	Using identical materials and procedures, the results of a study can be duplicated, and ultimately, identical conclusions can be drawn.[10] In the context of algorithms, the same result can be obtained from the same data, code, and analysis methods.[9,11]
Robustness	The same results are obtained from an algorithm despite small perturbations in the input.[12,13]
Replicability	Conclusions are stable based on independently acquired data,[9,11] that is, code and analysis methods can be employed to external data with similar results and performance. For deep learning algorithms, replicability is equivalent to model generalizability, a key requirement for the clinical application of new algorithms.
Reusability	A piece of software is considered reusable if it can be included in an existing computational pathology setup with minor efforts (eg, without the need for extensive rearrangements of the workflow or software development).

and eosin (H&E) staining is the most commonly used routine staining for cancer diagnosis,[14] we restricted this review to methods for H&E-stained whole slide image (WSI) analysis. We considered 5 computational pathology use cases and assembled a systematic overview of publications published between January 2019 and March 2021. For this, we compiled criteria for reproducibility in a practical context and examined each work with respect to these. We additionally provide an overview of current data handling tools.

Use Cases

We collected 160 studies between January 2019 and March 2021 on the automated analysis of histologic slides for cancer diagnosis and treatment (Supplementary Tables S1-S5; flowchart for the selection is presented in Supplementary Fig. S1). We split this body of literature into the following use cases: (1) stain normalization; (2) tissue type segmentation; (3) evaluation of cell-level features; (4) genetic alteration prediction; and (5) inference of grading, staging, and prognostic information (Fig. 3).

In this technical chapter, we first briefly introduce every use case, followed by an overview of the latest DL methods, focusing on works that provide code along with the publication. At the end of each section, we wrap up with an analysis of the reproducibility in the specific context.

As a prerequisite, open and publicly available data handling tools for reading, annotating, and sharing histopathological data are essential. We compiled the most common tools in Table 3[15-20] and provided a more detailed overview in Supplementary Table S6 on software features, requirements, and the possibility of extending the tool by its own code.

Stain Normalization

Most work in computational histopathology focuses on H&E-stained routine sections, with hematoxylin staining nuclei in blue-purple and eosin staining extracellular material in pink.[21] Digitized sections are prone to many image variations caused by differences in the tissue preparation technique (eg, the thickness and flatness of the sample cut), staining protocols, handling, and storage conditions. Moreover, slide scanners differ in microscope illumination, image postprocessing, or noise handling. These factors lead to significant variability in the visual appearance of WSIs, which affects subsequent analysis and may lead to poor generalizability of algorithms. Computational methods aim at reducing the effects of these variations,[22] for example, by normalizing the stain color from a predefined source domain to 1 or more target domains. These methods include improved analytical approaches, such as color deconvolution and, more recently, DL-based methods.

Color Space Methods. Color deconvolution separates the hematoxylin from the eosin component in optical density space based on a reference tile.[23] This approach has been developed further recently: adaptive color deconvolution[24] incorporated the underlying stain distribution of the target WSI instead of only a single tile. Alternatively, nonnegative matrix factorization has been used to obtain a color deconvolution matrix[25] and was optimized for graphics processing unit (GPU) usage.[26] (This approach is not DL-based. We decided to keep it in the article because non–DL-based color space methods are heavily used in DL applications of computational pathology.)

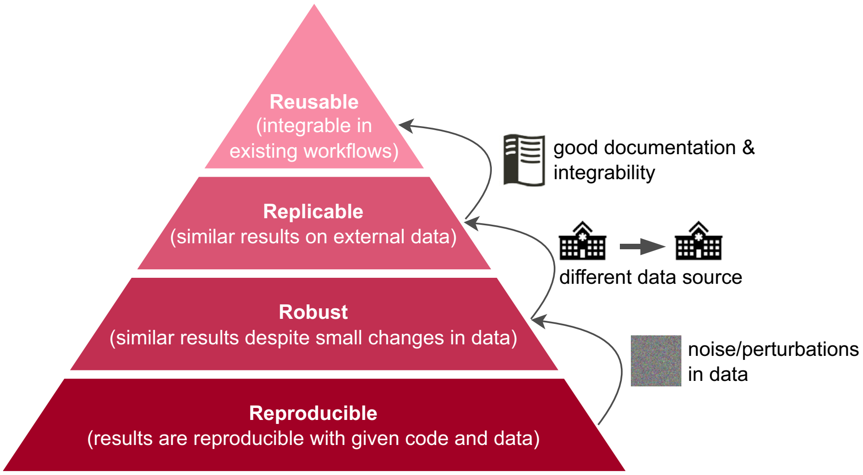
Reusable
(integrable in existing workflows)
Replicable
(similar results on external data)
Robust
(similar results despite small changes in data)
Reproducible
(results are reproducible with given code and data)
good documentation & integrability
different data source
noise/perturbations in data

Figure 2.

Reproducibility, robustness, replicability, and reusability in the context of deep learning algorithms for computational pathology.

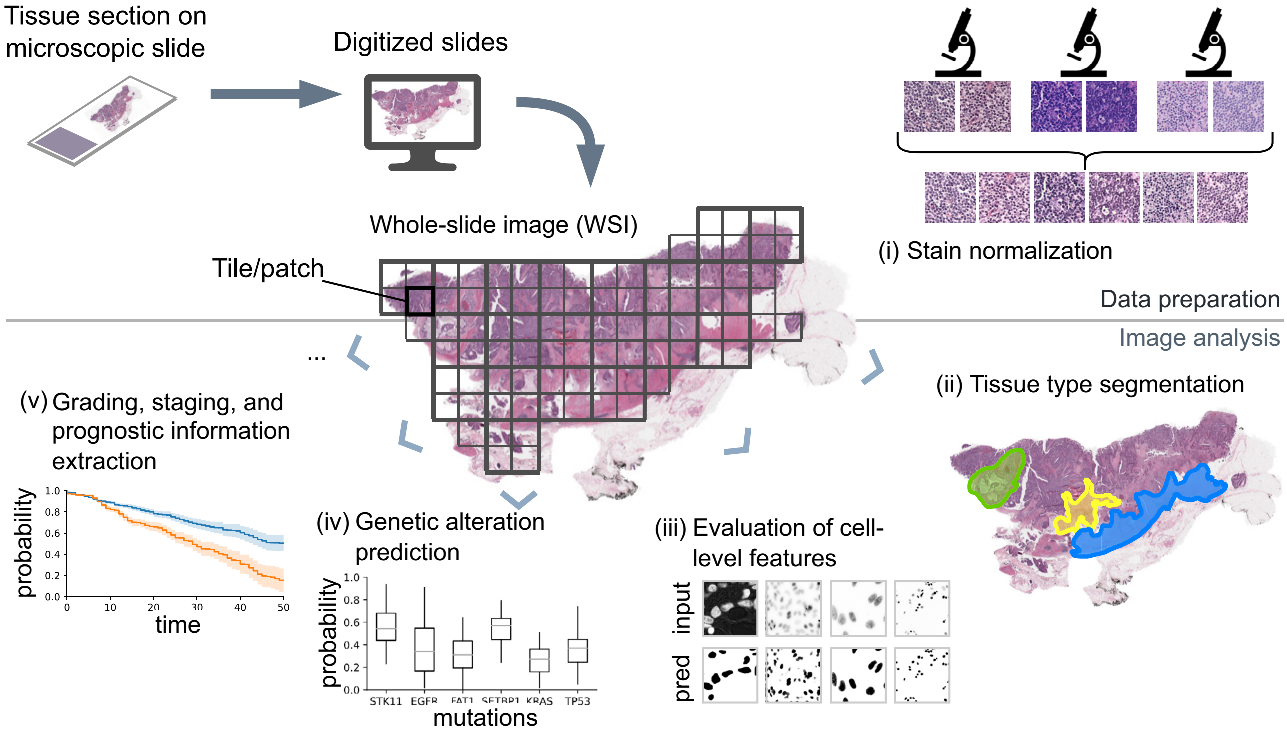

Figure 3.
Overview of the use of deep learning in computational pathology, including data handling tools for reading, annotating, and sharing whole slide images (WSIs) (Table 3), and 5 applications of deep learning methods, which are covered in the Use Cases.

Generative Models. The increasing popularity of generative adversarial networks (GANs) has led to the development of style transfer methods for stain normalization,[27] which can be trained on all target WSI tiles instead of expert-picked reference tiles.[28] StainGAN was trained with a cycle consistency loss between the source and target domains, and the generator of the target domain was used to normalize all images in that domain.[29] However, GANs may not always preserve the tissue structure.[30] To overcome this, StainNet trained a convolutional neural network (CNN) consisting of 1 × 1 convolutions, transforming the source image from its original color space to the target color space without losing structural information.[31] Alternatively, additional loss functions that compare images before and after normalization can be used to preserve the histopathological information, including texture, structure, and color features.[28]

Stain Augmentation. In contrast to normalization methods, GANs can also simulate stain variability by generating synthetic images. This renders neural networks on downstream tasks more robust and avoids loss of relevant information due to the limitations of normalization methods. Yamashita et al[32] propose data augmentation based on style transfer from artistic paintings, and Wagner et al[33] used a GAN architecture for multiple domains to synthesize realistic histologic images while preserving the tissue structure.

Stain-Aware Models. Unlike the above stain normalization methods that project the external test data to the original training domain as a preprocessing step, stain information can be incorporated directly into the model, for example, for nuclei segmentation by creating a hematoxylin-aware CNN.[34]

Tissue Type Segmentation

Accurate segmentation of a WSI into tissue types (eg, epithelial vs stromal vs lymphatic tissue) allows for quantitative follow-up analysis. Depending on the kind of available annotations, we discriminate between the following categories: methods for pixel-

Table 3
Overview of commonly used data handling tools in computational pathology

An essential prerequisite for implementing, transferring, and reusing computational pathology algorithms between researchers and different laboratories or institutions is a software structure for exchanging image data, annotations, and meta-information. With the progress of computational pathology, numerous data handling tools have been developed. In many tools, image data handling is based on system-level libraries, such as OpenSlide[15] or Open Microscopy Environment Remote Objects (OMERO).[16] They enable data to be interoperable between different vendor-specific image formats. Most of these tools also provide user interfaces for pathologists to analyze and annotate images. Annotations include class labeling or point flags, geometric shapes, and image-level labeling. While many popular image data handling and annotation tools were developed as standalone packages (eg, SlideRunner,[17] QuPath,[18] and Automated Slide Analysis Platform [ASAP]), an increasing number of recently developed packages, such as Cytomine[19] or EXpert Algorithm Collaboration Tool (EXACT),[20] allow for web-based, collaborative data handling, which is essential for distributing, exchanging, and annotating data as well as evaluating models in a multi-institutional setting.
In addition to image annotation and exchange, data handling packages allow integration with independently developed analysis algorithms at different levels. Some tools offer integrated scripting for the automation of tasks, for example, using Groovy in QuPath. Additionally, programming interfaces to popular machine learning languages, such as Python, have been developed, for example, for OMERO. Several tools, such as EXACT and CaMicroscope, offer integrated, server-side evaluation of deep learning models. A detailed overview of open and publicly available data handling tools and their respective functionalities is provided in Supplementary Table S1.

wise or patch-wise segmentation, hierarchical architectures that imitate a pathologist's workflow, and methods that use WSI-level annotations and are therefore weakly supervised.

Pixel-Wise Segmentation. Vellal et al[35] assessed the risk of breast cancer from image features, such as the percentage of fibrous stroma, epithelium, and fatty tissue. Graham et al[36] developed a rotation-invariant CNN to account for the inherent rotational symmetry of histology images and validated their application on pixel-wise gland segmentation. Jayapandian et al[37] used pixel-wise segmentation for identifying 6 tissue types in kidney biopsies, applying the same U-Net architecture for segmenting patches with different magnifications.

Patch-Wise Segmentation. Image patches can be classified separately and subsequently stitched together to create a coarse segmentation map of the entire WSI. Zhao et al[38] computed a WSI's tumor-to-stroma ratio from such segmented patches, a prognostic factor for colorectal cancer. Rączkowski et al[39] proposed an active learning framework to train a CNN for patch classification in colorectal cancer. The network's uncertainty, estimated via Monte-Carlo dropout sampling,[40] was used to detect outlier tiles in the training set and to select them afterward for reconsideration. Wang et al[41] generated spatial tissue maps by classifying single cells into tumor, stroma, and lymphocytes.

Hierarchical Segmentation. Hierarchical segmentation approaches mimic the workflow of pathologists by aggregating information from multiple scales of magnification. Schmitz et al[42] created a family of U-Net—based encoder-decoder architectures that process high- and low-resolution image tiles in separate branches, fusing them with learnable gates from 3 publicly available data sets with liver, breast, and lymph node tissues. Alternatively, HookNet[43] fused the hidden space of multiple U-Net—based models that operate on different scales to deal with high-resolution and contextual information in breast and lung cancers.

Weakly Supervised Methods. Weakly supervised methods typically require slide-level annotations only. One recent approach was training CNNs directly on the entire WSI of lung cancer sections.[44] Subsequently, class activation maps[45] highlight relevant cancerous regions that were also identified by pathologists, which can be interpreted as a confidence measure of the algorithm. Silva-Rodríguez et al[46] trained a feature extraction network on the entire downsampled WSI to classify global Gleason grades on prostate cancer. During inference, semantic segmentations are upscaled from the feature maps and achieve similar performance as fully supervised approaches while only using the global labels. Alternatively, the WSI can be split into patches and patch-wise features for WSI classification.[47] Using this approach, attention scores produce interpretable heatmaps to visualize which regions contribute to the network's prediction.

Evaluation of Cell-Level Features

The evaluation of cell-level properties is a standard task in histopathology. For example, cell density and the abundance of dividing cells in tumor tissue are critical features for tumor grading. Here, we focus on 2 of the most widely studied cell-level tasks: segmentation of nuclei and detection of mitoses.

Nuclei Segmentation. Segmentation challenges have been introduced to benchmark efforts in this field, such as the Multi-Organ Nucleus Segmentation Challenge,[48] which provided 30 images and approximately 22,000 nuclear boundary annotations in a public data set. Of the top 6 participants, 3 used U-Net—based semantic segmentation,[2] 2 used Mask region (R)-CNN—based instance segmentation,[3] and one group used stacked U-Net and R-CNN models. The 2 dominating algorithms have been further tailored toward nuclear segmentation: Cui et al[49] predicted a boundary map additionally to object segmentations to separate touching nuclei efficiently. Jin et al[50] incorporated a U-Net into a pipeline to detect lymph node metastasis in patients with breast cancer. Mask R-CNN has been combined with a deep convolutional Gaussian mixture color normalization model, which clusters pixels according to nucleus morphology.[51] Recently, other approaches, such as GANs, have been proposed, where the network is trained on unpaired data to map segmentation masks to nuclei images.[52] To ease the annotation process for nuclei segmentation, Qu et al[53] provided a DL framework trained on incomplete annotations, which are much easier to generate.

Mitosis Detection. Identifying cells in the cell cycle's mitotic phases is a diagnostically relevant task, for example, for breast cancer grading and prognosis.[54] Several challenges released public data sets and benchmarked competing approaches for mitosis detection, for example, ICPR MITOS-2012,[55] ICPR MITOS-ATYPIA-2014,[56] or Tumor Proliferation Assessment Challenge (TUPAC) 16.[54] Most recently developed mitosis detection methods can be grouped into classification, segmentation, and detection. Pati et al[57] combined a classification task with metric learning to reduce the necessary amount of labeled data for more efficient network training for patch-level classification. Another approach for mitosis detection is pixel-wise semantic segmentation. Jiménez and Racoceanu[58] showed that a U-Net—based semantic segmentation approach led to higher accuracy than that with previous classification approaches. Lafarge et al[59] proposed a special Euclidean motion group convolution to achieve translation and rotation invariance, which was integrated into a U-Net architecture and improved the model's robustness. Many other recent publications on mitosis detection were based on object detection,[60-63] where only a weak centroid annotation that marks the center of the mitotic figure is required, compared to pixel-wise annotations for segmentation approaches. An alternative approach applied a cascade network, combining a first-stage object detection to identify mitosis candidates and a second-stage classification network for refinement.[64]

Genetic Alteration Prediction

As genetic alterations can carry crucial predictive and prognostic information for cancer diagnosis, they have become increasingly relevant to the diagnostic workup and the selection of therapeutic pathways.[65] Therefore, patients are profiled for genetic alterations or other biomarkers that characterize their disease, for example, colorectal cancer,[66] to obtain better-targeted therapies. However, using molecular assays to determine the mutational spectrum of malignant cells is expensive and time-consuming. Furthermore, DNA or RNA extracted from small samples may not suffice quantitatively for a comprehensive analysis. RNA in older samples may already be degraded and, hence, not qualified for analysis. Techniques such as whole-genome sequencing require fresh tissue and are thus not applicable to formalin-fixed paraffin-embedded tissue usually used for histologic sample preparation. Therefore, algorithm-based prognostic stratification and mutation prediction from H&E-stained WSIs offer an attractive, cost- and time-effective, as well as tissue-sparing addition to existing molecular characterization methods.

Image-Based Mutation Prediction. Although some genetic alterations, such as mutations, copy number variations, and translocations, can be relevant for disease characterization, most work has so far focused on mutation prediction from imaging data. Coudray et al[5] found that 6 commonly mutated genes in lung adenocarcinoma (*STK11*, *EGFR*, *FAT1*, *SETBP1*, *KRAS*, and *TP53*) can be predicted from WSI images. Since then, image-based mutation prediction has been applied to various types of cancers, such as melanoma,[67,68] breast cancer,[69-71] lung cancer,[72,73] colorectal cancer,[74-76] bladder cancer,[77] and thyroid carcinoma.[78] Several recent studies attempt a pan-cancer approach that predicts genetic alteration across multiple tissue types from WSIs directly.[79,80]

Modeling Strategies for Mutation Prediction. Most approaches rely on standardized processing pipelines from preprocessing (Stain Normalization) and region of interest extraction (Tissue Type Segmentation) to model training and evaluation. Common CNN or transformer-based models are used as network architectures for per-tile prediction, where all tiles from a patient's WSI inherit the same label. This label can be both continuous (eg, tumor mutational burden) or categorical (eg, the mutation status of a selected gene or the microsatellite status).[5,79] The final prediction at WSI level is an aggregation of tile labels using, in the simplest case, majority voting (for categorical targets) or averaging (for continuous targets). Alternatively, Cao et al[76] employed multiple instance learning to weight the tile-level prediction by importance, which achieved better accuracy than standard supervised learning methods. Fu et al[81] classified each tile into different malignant and nonmalignant tissue types in a pan-cancer fashion and, subsequently, predicted driver gene mutations. Almost all studies mentioned above, except for Bychkov et al,[71] trained their networks on publicly available data from The Cancer Genome Atlas (TCGA).[82]

Grading, Staging, and Prognostic Information Extraction

A typical goal of histopathology is to recognize and evaluate primary lesions, determine their histopathologic subtype and grade (as defined by respective World Health Organization classifications for different tumor entities), and derive therapeutically relevant information from these features. In the context of computational pathology, this set of tasks can be addressed by determining features known to possess prognostic or predictive value. Alternatively, extracting prognostic or predictive information directly from imaging data, molecular properties, or clinical data can be attempted. Both methodologies have recently been applied across a variety of entities.

Inference of Known Factors and Biomarkers. Computational pathology approaches to extract known markers and scores include determination of the Gleason score in prostate cancer,[83-85] grading of gliomas,[86,87] and automated evaluation of mitoses[88,89] or tumor-infiltrating lymphocytes in breast[90] and head and neck cancer.[91] Machine learning methods have also been applied in disease staging, for example, to assess the degree of spread to the lymph nodes, either by highlighting areas suspicious for lymphovascular invasion, as in the case of testicular cancer,[92] or by predicting the risk of lymph node metastasis from the primary lesion, as in the case of bladder cancer.[93] Several studies inferred molecular properties with a prognostic value from H&E, such as microsatellite instability in gastrointestinal cancer[94] or the molecular subtype of invasive bladder cancer.[77]

Image-Based Biomarkers. Finally, prognostic factors can be derived directly from histopathology images[38] or in combination with other clinical or molecular data, for example, from the genome or transcriptome.[38,95,96] This route can be followed without referring to previously known prognostic factors. Hence, although these approaches may first rely on computational predictions alone, they may also lead to identifying novel prognostic or predictive factors that lend themselves to direct human evaluation, which can be identified through explainability methods.[97] Examples of this approach include automated quantification of intratumoral stroma in rectal cancer,[98] evaluation of nuclear morphology for survival prediction in lung cancer,[99] DL-based prognosis in nasopharyngeal cancer,[98,100] and survival prediction in colorectal cancer.[101,102]

Materials and Methods

To assess reproducibility and reusability in computational pathology, we scanned whether and how code was publicly available, evaluated criteria for data access, and checked if the statistical variance of the reported findings was provided. In Supplementary Table S2, we list all 41 publications (out of 160) together with the following evaluation criteria that we used for code, data, and statistical variance. (1) Inspired by the FAIR principles demanding that data should be Findable, Accessible, Interoperable, and Reusable,[103] we surveyed whether the code was made publicly available. Additionally, we noted the platform used for sharing and the programming language and machine learning frameworks employed. Furthermore, we checked for instructions for running the code, whether the code was minimally documented, and if a pretrained model was available for direct application. (2) We evaluated data access for the 41 publications with available code and checked whether the data set and required annotations were publicly available. Additionally, we recorded what kind of data had been used (eg, tiled WSIs vs entire WSIs). We also reported whether the appropriate preprocessing steps were provided and the training-validation-test split used for model development and evaluation. In terms of replicability, we specified what kind of test set had been used, whether it was similar to the training set or covered an independent cohort. (3) Finally, we checked if any measure of statistical variance of the reported findings was provided. This is one way to tackle the difficulties concerned with reproducing the results, which can be introduced on multiple levels: computer-level inaccuracies, such as floating-point numbers that can be rounded differently on different machines, architectures, or execution environments,[104] or algorithm-level stochasticity due to the stochastic behavior of optimization techniques. One way of dealing with this is to statistically analyze the experimental results (eg, by determining the SD of the results) and perform a sufficient exploration of hyperparameters.[9] A straightforward evaluation approach is to repeat an experiment multiple times or over several folds of cross-validation and report mean and SD across experiments.

Results

Code Availability

In our study, 41 of the 160 publications (26%) made their code publicly available (Fig. 4A). Interestingly, the ratio of publications with code differs across the 5 use cases (Fig. 4B). For stain normalization, we retrieved 29 research articles, where only 7 (24%) of them provided code with their method. In the field of tissue type segmentation and localization, only 12 of the total 51 investigated articles (24%) had their code publicly available, and

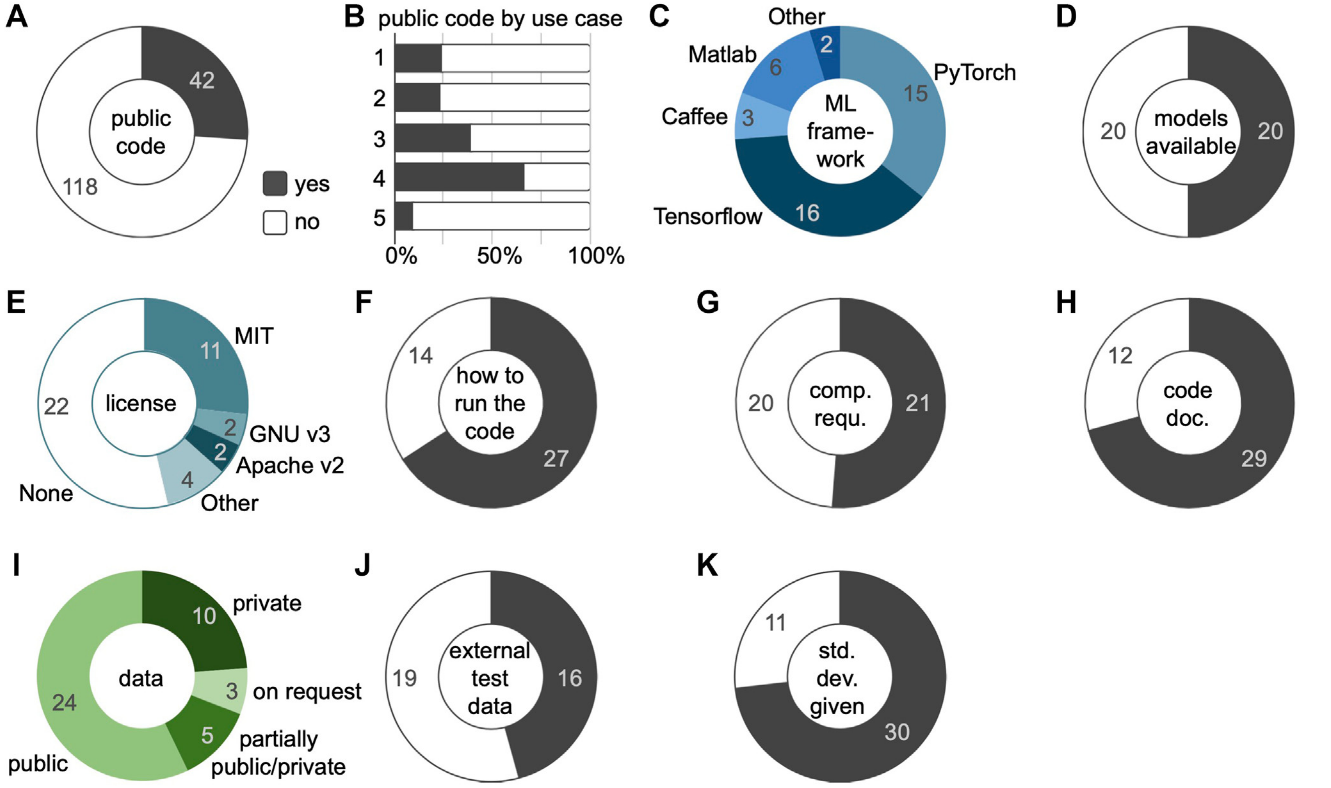

Figure 4.
Analysis of our systematic literature search of computational pathology articles. (A) A quarter of the methods are published with code (26%). (B) The ratio differs across the use cases. (C) Most works (74%) used PyTorch or TensorFlow as machine learning (ML) frameworks. (D) Half of the publications with code release their final model weights. (E) Half of the code is published under a license. (F) More than half of the code includes instructions on how to run the code. (G) The computational requirements are mentioned in half of the repositories. (H) Almost three-quarters documented their code inline. (I) Large public data sets are mainly used and sometimes complemented by private cohorts. (J) Almost half of the publications used an independent cohort for evaluation. (K) The majority analyzes their results statistically.

only 3 publications provided the pretrained model weights. Among the 28 research articles that we screened for the evaluation of cell-level features, 10 articles (39%) provided code with the publications. The code of 5 of these 10 could be run in Google Colab and thus was directly applicable. For genetic alteration prediction, 8 of the 13 articles (61%) have provided their code along with their method. In survival analysis, only 4 of the 38 studies (11%) published their code. Interestingly, genetic alteration prediction has the highest ratio of published code. One reason could be that a key publication included a well-documented codebase[5]; most subsequent publications were aware of this work and, therefore, matched this standard by publishing their code, including documentations. Hence, the level of reusability in one field may depend on the preceding publications. This also strengthens the role of the publisher in the context of reproducibility and reusability in computational pathology. For the 41 articles that published their code, we checked the evaluation criteria for code, data, and statistical variance detailed in Materials and Methods and detailed them in Supplementary Table S2.

Machine Learning Frameworks, Model Weights, and License

Almost 75% of the methods were implemented in Python using the open and publicly available machine learning frameworks TensorFlow (https://www.tensorflow.org/) (38%) and PyTorch (https://pytorch.org/) (36%; Fig. 4C). Pretrained model weights were only available for half of the publications that also provided code (Fig. 4D); this makes reproducing experimental results difficult, particularly because the preprocessing steps are rarely available. Also, without model weights, a direct application without retraining the model is not possible, hampering its use by pathologists not specialized in DL. Less than half of the repositories were published under a license (46%), which is the legal requirement for reusing code since code without any license falls under the default copyright laws. This shows that the importance of licensing is not yet widely recognized in the research field representing a hurdle toward clinical implementation.

Documentation

Detailed documentation is an essential prerequisite for any step toward reproducing or reusing a published codebase. In the repositories analyzed in our study, 27 repositories (66%) contained instructions on how to run the code or train the DL model. However, only 21 repositories (51%) defined the computational requirements or provided a computational environment for running the code (Fig. 4G). Examining the main code files, 29 codebases (71%) provided at least a minimal level of documentation in the form of inline comments (Fig. 4H).

Data Sets

More than half of all methods (57%) were evaluated on publicly available data sets (Fig. 4I). Most studies (eg, all 13 studies reviewed for genetic alteration prediction) developed their methods based on TCGA[82]; it contains data from multiple institutions; thus, it can be split into training and test sets by cohort level. Complementing TCGA with external, mostly private, data as an independent evaluation cohort was also a common practice (Fig. 4J). Nevertheless, TCGA, with a few hundred slides for each cancer type, is insufficient to represent all cancer heterogeneities. The reliance of computational pathology on relatively few publicly available data sets renders their selection strategy and processing critical. Batch effects can be detected by DL models and lead to overestimation of the model's performance.[105] Therefore, we strongly encourage the development of more publicly available multi-institutional data sets.

Statistical Variance

We believe that a thorough evaluation of sources and magnitude of variability, both on an algorithmic and a data level, is essential to making modern computational pathology algorithms more reusable and generalizable. Almost three-quarters (73%) of the methods analyzed their results statistically, in which we considered all kinds of statistical notions to be statistical analysis (Fig. 4K). Many different sources of variability can be relevant to the performance of computational pathology algorithms. It is, therefore, difficult to devise a single strategy for quantifying all sources.

Discussion

It has been increasingly recognized that computational reproducibility and reusability are an essential part of good scientific practice for DL applications.[106-109] Especially for the interdisciplinary field of computational pathology, both are critical requirements for enabling wider use of algorithms and the translation to clinical application.

In our survey of recently published computational pathology DL approaches, however, we found that there is still a long way to go. For stain color normalization, for example, techniques to reduce the color and intensity variations in histologic images from different laboratories can render a downstream task algorithm more generalizable. Although neural network—based stain normalization techniques have evolved considerably in recent years (Stain Normalization), their use in downstream applications is still limited, probably because pretrained stain normalization models are rarely available and, in most cases, code is not shared. Instead, we observe that easy-to-use algorithms not based on DL are typically applied. The lack of reusability hinders the practical application of innovative network-based methods.

Even if the code is shared, supporting documentation is necessary to reuse the code but is often missing. Especially for researchers without a computational background, even a well-documented codebase might not lower the hurdle sufficiently to adopt existing DL algorithms. Build-in plugins or models integrated into commonly used data handling tools are desirable from a clinical perspective.

Furthermore, convincing experiments on external cohorts are often missing, lowering the chances of successful reuse and translation. Most state-of-the-art methods in computational pathology are based on DL algorithms and typically require large amounts of labeled training data. Making these data available is as crucial as providing a well-documented code. We acknowledge that data and appropriate annotations cannot be publicly shared in some cases, for example, due to legal or ethical constraints. Here, reasonable compromises, such as partial data sharing or evaluation using public data sets,[108] should be considered.

Finally, reusable code is likely to have an impact on the adoption of computational methods not only in a research context but also in the diagnostic routine. Although the full translation of research algorithms to a diagnostic setting is a complex process beyond the scope of the present discussion, shared and well-documented code can be of key importance for generating initial results on routine data. These could serve as an initial test for users from the diagnostic routine to evaluate whether a given algorithm might be usable and helpful in their specific setting, hence providing motivation for users to initiate the path toward integration of algorithms into their diagnostic routine. An increase of reusable and reproducible algorithms in a research context is likely to foster the adoption of DL algorithms in a routine setting.

Despite the lack of reproducibility and reusability in many computational pathology approaches, we hope the field will profit from the surging discussions, for example, in computer vision.[9] As a step in this direction, large conferences, such as Medical Image Computing and Computer-Assisted Intervention since 2021, started to employ reproducibility checklists for authors in their submission form that will be publicly available upon acceptance of the article. We encourage the scientific community to recognize the long-term value of reproducibility and reusability and to foster their realization in computational pathology.

Author Contributions

S.J.W., C.Matek, T.P., and C.Marr designed the concept of the review. M.B. and C.Matek supported with domain-related advice. S.J.W., L.L., and D.J.E.W. compiled the evaluation criteria for reproducibility and reusability. S.J.W. and C.Matek assessed current data handling tools. For the use cases, S.J.W., D.J.E.W., and S.S.B. reviewed the publications for stain normalization, L.L. and S.J.W. for tissue type segmentation, D.J.E.W. and T.P. for evaluation of cell-level features, T.P. for genetic alteration prediction, and A.S. and C.Matek for grading, staging, and prognostic information inference. S.J.W. analyzed the results. All authors contributed to constructive discussions. S.J.W. wrote the manuscript with C.Matek, C.Marr, and T.P. and input from all other authors.

Data Availability

Not applicable.

Funding

S.J.W., L.L., and S.S.B. are supported by the Helmholtz Association under the joint research school "Munich School for Data Science—MUDS." S.J.W. is supported by the Add-on Fellowship of the Joachim Herz Foundation. C.Marr has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement No. 866411).

Declaration of Competing Interest

None reported.

Ethics Approval and Consent to Participate

Not applicable.

Supplementary Material

The online version contains supplementary material available at https://doi.org/10.1016/j.modpat.2023.100350

References

1. Fuchs TJ, Buhmann JM. Computational pathology: challenges and promises for tissue analysis. *Comput Med Imaging Graph*. 2011;35(7-8):515–530.
2. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*. Springer International Publishing; 2015:234–241.
3. He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017:2961–2969.
4. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115–118. https://doi.org/10.1038/nature21056
5. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat Med*. 2018;24(10):1559–1567.
6. Matek C, Schwarz S, Spiekermann K, Marr C. Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. *Nat Mach Intell*. 2019;1(11):538–544.
7. 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
8. van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: the path to the clinic. *Nat Med*. 2021;27(5):775–784.
9. Pineau J, Vincent-Lamarre P, Sinha K, et al. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). Preprint. Posted online March 27, 2020. arXiv 2003.12206. https://doi.org/10.48550/arXiv.2003.12206
10. Goodman SN, Fanelli D, Ioannidis JPA. What does research reproducibility mean? *Sci Transl Med*. 2016;8(341):341ps12.
11. Artner R, Verliefde T, Steegen S, et al. The reproducibility of statistical results in psychological research: an investigation using unpublished raw data. *Psychol Methods*. 2021;26(5):527–546. https://doi.org/10.1037/met0000365
12. Oala L, Fehr J, Gilli L, et al. ML4H auditing: from paper to practice. In: Alsentzer E, McDermott MBA, Falck F, Sarkar SK, Roy S, Hyland SL, eds. *Proceedings of the Machine Learning for Health NeurIPS Workshop*. Vol 136. Proceedings of Machine Learning Research; 2020:280–317.
13. Li JZ. Principled approaches to robust machine learning and beyond. Massachusetts Institute of Technology. Accessed June 28, 2021. https://dspace.mit.edu/handle/1721.1/120382?show=full
14. Rosai J. Why microscopy will remain a cornerstone of surgical pathology. *Lab Invest*. 2007;87(5):403–408.
15. Goode A, Gilbert B, Harkes J, Jukic D, Satyanarayanan M. OpenSlide: a vendor-neutral software foundation for digital pathology. *J Pathol Inform*. 2013;4:27.
16. Allan C, Burel JM, Moore J, et al. OMERO: flexible, model-driven data management for experimental biology. *Nat Methods*. 2012;9(3):245–253.
17. Aubreville M, Bertram C, Klopfleisch R, Maier A. SlideRunner. In: *Bildverarbeitung Für Die Medizin 2018*. Springer Berlin Heidelberg; 2018:309–314.
18. Bankhead P, Loughrey MB, Fernández JA, et al. QuPath: open source software for digital pathology image analysis. *Sci Rep*. 2017;7(1), 16878.
19. Marée R, Rollus L, Stévens B, et al. Collaborative analysis of multi-gigapixel imaging data using Cytomine. *Bioinformatics*. 2016;32(9):1395–1401.
20. Marzahl C, Aubreville M, Bertram CA, et al. EXACT: a collaboration toolset for algorithm-aided annotation of images with annotation version control. *Sci Rep*. 2021;11(1):4343.
21. Chan JKC. The wonderful colors of the hematoxylin-eosin stain in diagnostic surgical pathology. *Int J Surg Pathol*. 2014;22(1):12–32.
22. Chen JM, Li Y, Xu J, et al. Computer-aided prognosis on breast cancer with hematoxylin and eosin histopathology images: a review. *Tumour Biol*. 2017;39(3), 1010428317694550.
23. Macenko M, Niethammer M, Marron JS, et al. A method for normalizing histology slides for quantitative analysis. In: *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE; 2009. https://doi.org/10.1109/ISBI.2009.5193250
24. Zheng Y, Jiang Z, Zhang H, Xie F, Shi J, Xue C. Adaptive color deconvolution for histological WSI normalization. *Comput Methods Programs Biomed*. 2019;170:107–120.
25. Vahadane A, Peng T, Sethi A, et al. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Trans Med Imaging*. 2016;35(8):1962–1971.
26. Anand D, Ramakrishnan G, Sethi A. Fast GPU-enabled color normalization for digital pathology. In: *2019 International Conference on Systems, Signals and Image Processing (IWSSIP)*. 2019:219–224.
27. Tschuchnig ME, Oostingh GJ, Gadermayr M. Generative adversarial networks in digital pathology: a survey on trends and future potential. *Patterns (N Y)*. 2020;1(6), 100089.
28. Liang H, Plataniotis KN, Li X. Stain Style Transfer of histopathology images via structure-preserved generative learning. In: *Machine Learning for Medical Image Reconstruction*. Springer International Publishing; 2020:153–162.
29. Shaban MT, Baur C, Navab N, Albarqouni S. Staingan: stain style transfer for digital histological images. In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. 2019:953–956.
30. Cohen JP, Luck M, Honari S. Distribution matching losses can hallucinate features in medical image translation. In: *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018*. Springer International Publishing; 2018:529–536.
31. Kang H, Luo D, Feng W, et al. StainNet: a fast and robust stain normalization network. *Front Med (Lausanne)*. 2021;8:746307.
32. Yamashita R, Long J, Banda S, Shen J, Rubin DL. Learning domain-agnostic visual representation for computational pathology using medically-irrelevant style transfer augmentation. *IEEE Trans Med Imaging*. 2021;40(12):3945–3954. https://doi.org/10.1109/TMI.2021.3101985
33. Wagner SJ, Khalili N, Sharma R, et al. Structure-preserving multi-domain stain color augmentation using style-transfer with disentangled representations. *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference*. 2021. https://doi.org/10.1007/978-3-030-87237-3_25
34. Zhao B, Chen X, Li Z, et al. Triple U-net: hematoxylin-aware nuclei segmentation with progressive dense feature aggregation. *Med Image Anal*. 2020;65, 101786.
35. Vellal AD, Sirinukunwattan K, Kensler KH, et al. Deep learning image analysis of benign breast disease to identify subsequent risk of breast cancer. *JNCI Cancer Spectr*. 2021;5(1):kaa119.
36. Graham S, Epstein D, Rajpoot N. Dense steerable filter CNNs for exploiting rotational symmetry in histology images. *IEEE Trans Med Imaging*. 2020;39(12):4124–4136.
37. Jayapandian CP, Chen Y, Janowczyk AR, et al. Development and evaluation of deep learning-based segmentation of histologic structures in the kidney cortex with multiple histologic stains. *Kidney Int*. 2021;99(1):86–101.
38. Zhao K, Li Z, Yao S, et al. Artificial intelligence quantified tumour-stroma ratio is an independent predictor for overall survival in resectable colorectal cancer. *EBioMedicine*. 2020;61:103054.
39. Rączkowski A, Możejko M, Zambonelli J, Szczurek E. ARA: accurate, reliable and active histopathological image classification framework with Bayesian deep learning. *Sci Rep*. 2019;9(1):14347.
40. Gal Y, Ghahramani Z. Bayesian convolutional neural networks with Bernoulli approximate variational inference. Preprint. Posted online June 6, 2015. arXiv 1506.02158. https://doi.org/10.48550/arXiv.1506.02158
41. Wang S, Wang T, Yang L, et al. ConvPath: a software tool for lung adenocarcinoma digital pathological image analysis aided by a convolutional neural network. *EBioMedicine*. 2019;50:103–110.
42. Schmitz R, Madesta F, Nielsen M, et al. Multi-scale fully convolutional neural networks for histopathology image segmentation: from nuclear aberrations to the global tissue architecture. *Med Image Anal*. 2021;70:101996.
43. van Rijthoven M, Balkenhol M, Siliņa K, van der Laak J, Ciompi F. HookNet: multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images. *Med Image Anal*. 2021;68:101890.
44. 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.
45. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016:2921–2929.
46. Silva-Rodríguez J, Colomer A, Naranjo V. WeGleNet: a weakly-supervised convolutional neural network for the semantic segmentation of Gleason grades in prostate histology images. *Comput Med Imaging Graph*. 2021;88, 101846.
47. Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat Biomed Eng*. 2021;5(6):555–570. https://doi.org/10.1038/s41551-020-00682-w
48. Kumar N, Verma R, Anand D, et al. A multi-organ nucleus segmentation challenge. *IEEE Trans Med Imaging*. 2020;39(5):1380–1391.
49. Cui Y, Zhang G, Liu Z, Xiong Z, Hu J. A deep learning algorithm for one-step contour aware nuclei segmentation of histopathology images. *Med Biol Eng Comput*. 2019;57(9):2027–2043.
50. Jin YW, Jia S, Ashraf AB, Hu P. Integrative data augmentation with U-net segmentation masks improves detection of lymph node metastases in

breast cancer patients. *Cancers*. 2020;12(10). https://doi.org/10.3390/cancers12102934

51. Jung H, Lodhi B, Kang J. An automatic nuclei segmentation method based on deep convolutional neural networks for histopathology images. *BMC Biomed Eng*. 2019;1:24.
52. Mahmood F, Borders D, Chen RJ, et al. Deep adversarial training for multi-organ nuclei segmentation in histopathology images. *IEEE Trans Med Imaging*. 2020;39(11):3257–3267.
53. Qu H, Wu P, Huang Q, et al. Weakly supervised deep nuclei segmentation using partial points annotation in histopathology images. *IEEE Trans Med Imaging*. 2020;39(11):3655–3666.
54. Veta M, Heng YJ, Stathonikos N, et al. Predicting breast tumor proliferation from whole-slide images: the TUPAC16 challenge. *Med Image Anal*. 2019;54:111–121.
55. Roux L, Racoceanu D, Loménie N, et al. Mitosis detection in breast cancer histological images an ICPR 2012 contest. *J Pathol Inform*. 2013;4(1):8. https://doi.org/10.4103/2153-3539.112693
56. Roux L, Racoceanu D, Capron F, et al. MITOS & ATYPIA—detection of mitosis and evaluation of nuclear atypia score in breast cancer histological images. An ICPR 2014 Contest. 22nd International Conference on Pattern Recognition. Image Pervasive Access Lab(IPAL) Lab; June 27, 2014; Singapore.
57. Pati P, Foncubierta-Rodríguez A, Goksel O, Gabrani M. Reducing annotation effort in digital pathology: a co-representation learning framework for classification tasks. *Med Image Anal*. 2021;67:101859.
58. Jiménez G, Racoceanu D. Deep learning for semantic segmentation vs. classification in computational pathology: application to mitosis analysis in breast cancer grading. *Front Bioeng Biotechnol*. 2019;7:145.
59. Lafarge MW, Bekkers EJ, Pluim JPW, Duits R, Veta M. Roto-translation equivariant convolutional networks: application to histopathology image analysis. *Med Image Anal*. 2021;68:101849.
60. Lei H, Liu S, Xie H, Kuo JY, Lei B. An improved object detection method for mitosis detection. *Conf Proc IEEE Eng Med Biol Soc*. 2019;2019:130–133.
61. Sohail A, Khan A, Wahab N, Zameer A, Khan S. A multi-phase deep CNN based mitosis detection framework for breast cancer histopathological images. *Sci Rep*. 2021;11(1):6215.
62. Wollmann T, Rohr K. Deep consensus network: aggregating predictions to improve object detection in microscopy images. *Med Image Anal*. 2021;70,102019.
63. Lei H, Liu S, Elazab A, Gong X, Lei B. Attention-guided multi-branch convolutional neural network for mitosis detection from histopathological images. *IEEE J Biomed Health Inform*. 2021;25(2):358–370.
64. Mahmood T, Arsalan M, Owais M, Lee MB, Park KR. Artificial intelligence-based mitosis detection in breast cancer histopathology images using faster R-CNN and deep CNNs. *J Clin Med*. 2020;9(3):749. https://doi.org/10.3390/jcm9030749
65. Ashley EA. Towards precision medicine. *Nat Rev Genet*. 2016;17(9):507–522.
66. Singh MP, Rai S, Pandey A, Singh NK, Srivastava S. Molecular subtypes of colorectal cancer: an emerging therapeutic opportunity for personalized medicine. *Genes Dis*. 2021;8(2):133–145.
67. Zhang H, Kalirai H, Acha-Sagredo A, Yang X, Zheng Y, Coupland SE. Piloting a deep learning model for predicting nuclear BAP1 immunohistochemical expression of uveal melanoma from hematoxylin-and-eosin sections. *Transl Vis Sci Technol*. 2020;9(2):50.
68. Kim RH, Nomikou S, Dawood Z, et al. *A Deep Learning Approach for Rapid Mutational Screening in Melanoma*. Cold Spring Harbor Laboratory; 2019: 610311. https://doi.org/10.1101/610311
69. Lu Z, Xu S, Shao W, et al. Deep-learning-based characterization of tumor-infiltrating lymphocytes in breast cancers from histopathology images and multiomics data. *JCO Clin Cancer Inform*. 2020;4:480–490.
70. Anand D, Kurian NC, Dhage S, et al. Deep learning to estimate human epidermal growth factor receptor 2 status from hematoxylin and eosin-stained breast tissue images. *J Pathol Inform*. 2020;11:19.
71. Bychkov D, Linder N, Tiulpin A, et al. Deep learning identifies morphological features in breast cancer predictive of cancer ERBB2 status and trastuzumab treatment efficacy. *Sci Rep*. 2021;11(1):4037. https://doi.org/10.1038/s41598-021-83102-6
72. Wang S, Rong R, Yang DM, et al. Computational staining of pathology images to study the tumor microenvironment in lung cancer. *Cancer Res*. 2020;80(10):2056–2066.
73. Yu KH, Wang F, Berry GJ, et al. Classifying non-small cell lung cancer types and transcriptomic subtypes using convolutional neural networks. *J Am Med Inform Assoc*. 2020;27(5):757–769.
74. Jang HJ, Lee A, Kang J, Song IH, Lee SH. Prediction of clinically actionable genetic alterations from colorectal cancer histopathology images using deep learning. *World J Gastroenterol*. 2020;26(40):6207–6223.
75. Echle A, Grabsch HI, Quirke P, et al. Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning. *Gastroenterology*. 2020;159(4):1406–1416.e11.
76. Cao R, Yang F, Ma SC, et al. Development and interpretation of a pathomics-based model for the prediction of microsatellite instability in colorectal cancer. *Theranostics*. 2020;10(24):11080–11091.
77. Woerl AC, Eckstein M, Geiger J, et al. Deep learning predicts molecular subtype of muscle-invasive bladder cancer from conventional histopathological slides. *Eur Urol*. 2020;78(2):256–264.
78. Tsou P, Wu CJ. Mapping driver mutations to histopathological subtypes in papillary thyroid carcinoma: applying a deep convolutional neural network. *J Clin Med*. 2019;8(10):1675. https://doi.org/10.3390/jcm8101675
79. Kather JN, Heij LR, Grabsch HI, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat Cancer*. 2020;1(8):789–799.
80. Noorbakhsh J, Farahmand S, Foroughi Pour A, et al. Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images. *Nat Commun*. 2020;11(1):6367.
81. Fu Y, Jung AW, Torne RV, et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat Cancer*. 2020;1(8):800–810. https://doi.org/10.1038/s43018-020-0085-8
82. Gutman DA, Cobb J, Somanna D, et al. Cancer Digital slide archive: an informatics resource to support integrated in silico analysis of TCGA pathology data. *J Am Med Inform Assoc*. 2013;20(6):1091–1098.
83. Steiner DF, Nagpal K, Sayres R, et al. Evaluation of the use of combined artificial intelligence and pathologist assessment to review and grade prostate biopsies. *JAMA Netw Open*. 2020;3(11):e2023267.
84. Bulten W, Balkenhol M, Belinga JA, et al. Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. *Mod Pathol*. 2021;34(3):660–671.
85. Nagpal K, Foote D, Liu Y, et al. Erratum: publisher correction: development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ Digit Med*. 2019;2:113.
86. Truong AH, Sharmanska V, Limbäck-Stanic C, Grech-Sollars M. Optimization of deep learning methods for visualization of tumor heterogeneity and brain tumor grading through digital pathology. *Neurooncol Adv*. 2020;2(1): vdaa110.
87. Rathore S, Niazi T, Iftikhar MA, Chaddad A. Glioma grading via analysis of digital pathology images using machine learning. *Cancers*. 2020;12(3):578. https://doi.org/10.3390/cancers12030578
88. Chang MC, Mrkonjic M. Review of the current state of digital image analysis in breast pathology. *Breast J*. 2020;26(6):1208–1212.
89. 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.
90. Balkenhol MC, Ciompi F, Świderska-Chadaj Ż, et al. Optimized tumour infiltrating lymphocyte assessment for triple negative breast cancer prognostics. *Breast*. 2021;56:78–87.
91. Shaban M, Khurram SA, Fraz MM, et al. A novel digital score for abundance of tumour infiltrating lymphocytes predicts disease free survival in oral squamous cell carcinoma. *Sci Rep*. 2019;9(1), 13341.
92. Ghosh A, Sirinukunwattana K, Khalid Alham N, et al. The potential of artificial intelligence to detect lymphovascular invasion in testicular cancer. *Cancers*. 2021;13(6):1325. https://doi.org/10.3390/cancers13061325
93. Harmon SA, Sanford TH, Brown GT, et al. Multiresolution application of artificial intelligence in digital pathology for prediction of positive lymph nodes from primary tumors in bladder cancer. *JCO Clin Cancer Inform*. 2020;4:367–382.
94. Kather JN, Pearson AT, Halama N, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med*. 2019;25(7):1054–1056.
95. Hao J, Kosaraju SC, Tsaku NZ, Song DH, Kang M. PAGE-Net: interpretable and integrative deep learning for survival analysis using histopathological images and genomic data. *Pac Symp Biocomput*. 2020;25:355–366.
96. Failmezger H, Muralidhar S, Rullan A, de Andrea CE, Sahai E, Yuan Y. Topological tumor graphs: a graph-based spatial model to infer stromal recruitment for immunosuppression in melanoma histology. *Cancer Res*. 2020;80(5):1199–1209.
97. Tosun AB, Pullara F, Becich MJ, Taylor DL, Fine JL, Chennubhotla SC. Explainable AI (xAI) for anatomic pathology. *Adv Anat Pathol*. 2020;27(4): 241–250.
98. Geessink OGF, Baidoshvili A, Klaase JM, et al. Computer aided quantification of intratumoral stroma yields an independent prognosticator in rectal cancer. *Cell Oncol (Dordr)*. 2019;42(3):331–341.
99. Alsubaie NM, Snead D, Rajpoot NM. Tumour nuclear morphometrics predict survival in lung adenocarcinoma. *IEEE Access*. 2021;9:12322–12331. https://doi.org/10.1109/ACCESS.2021.3049582
100. Zhang F, Zhong LZ, Zhao X, et al. A deep-learning-based prognostic nomogram integrating microscopic digital pathology and macroscopic magnetic resonance images in nasopharyngeal carcinoma: a multi-cohort study. *Ther Adv Med Oncol*. 2020;12, 1758835920971416.
101. Kather JN, Krisam J, Charoentong P, et al. Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. *PLoS Med*. 2019;16(1), e1002730.
102. Abbet C, Zlobec I, Bozorgtabar B, Thiran JP. Divide-and-rule: self-supervised learning for survival analysis in colorectal cancer. *Medical Image Computing and Computer Assisted Intervention—MICCAI*. 2020;2020:480–489. https://doi.org/10.1007/978-3-030-59722-1_46
103. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3: 160018.

104. Hill DRC. Repeatability, Reproducibility, Computer Science and High Performance Computing: stochastic simulations can be reproducible too. In: *2019 International Conference on High Performance Computing Simulation*. HPCS; 2019:322−323.

105. Howard FM, Dolezal J, Kochanny S, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nat Commun*. 2021;12(1):4423. https://doi.org/10.1038/s41467-021-24698-1

106. Stodden V, McNutt M, Bailey DH, et al. Enhancing reproducibility for computational methods. *Science*. 2016;354(6317):1240−1241.

107. Hutson M. Artificial intelligence faces reproducibility crisis. *Science*. 2018;359(6377):725−726.

108. Haibe-Kains B, Adam GA, Hosny A, et al. Transparency and reproducibility in artificial intelligence. *Nature*. 2020;586(7829):E14−E16.

109. Carpenter AE, Kamentsky L, Eliceiri KW. A call for bioimaging software usability. *Nat Methods*. 2012;9(7):666−670.