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

Sophia J. Wagner[1,2,*], Christian Matek[3,4,*], Sayedali Shetab Boushehri[3,5,6], Melanie Boxberg[7,8], Lorenz Lamm[1,9], Ario Sadafi[2,3], Dominik J. E. Waibel[3,10], Carsten Marr[3,†], Tingying Peng[1,†]

[1] Helmholtz AI, Helmholtz Munich – German Research Center for Environmental Health, Neuherberg, Germany
[2] Technical University of Munich, Department of Informatics, Garching, Germany
[3] Institute of AI for Health, Helmholtz Munich – German Research Center for Environmental Health, Neuherberg, Germany
[4] University Hospital Erlangen, Institute of Pathology, Erlangen, Germany
[5] Technical University of Munich, Department of Mathematics, Garching, Germany
[6] Data Science, Pharmaceutical Research and Early Development Informatics (pREDi), Roche Innovation Center Munich (RICM), Penzberg, Germany
[7] Technical University Munich, Institute of Pathology, Munich, Germany
[8] Institute of Pathology Munich-North, Munich, Germany
[9] Helmholtz Pioneer Campus, Helmholtz Munich – German Research Center for Environmental Health, Neuherberg, Germany
[10] Technical University of Munich, School of Life Sciences, Weihenstephan, Germany
* Equal contribution
† Corresponding authors

Keywords. Reproducibility. Reusability. Artificial Intelligence. Deep Learning. Computational Pathology. Cancer. Histology/Histopathology

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 in existing workflows, and models should be robust towards noise and generalizable towards 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 five use cases: stain normalization, tissue type segmentation, evaluation of cell-level features, genetic alteration prediction, and direct extraction of grading, staging, and prognostic information. We compiled criteria for data and code availability, and for statistical result analysis and assessed them in 161 publications. We found that only one quarter (42 out of 161 publications) made code publicly available and thus fulfilled our minimum requirement for reproducibility and reusability. Among these 42 papers, three quarters (30 out of 42) analyzed their results statistically, less than half (20 out of 42) have released their trained model weights, and only about a third (16 out of 42) used an independent cohort for evaluation.

This review highlights candidates for reproducible and reusable algorithms in computational pathology. It is intended for both pathologists interested in deep learning, and researchers applying deep learning algorithms to computational pathology challenges. We provide a list of reusable data handling tools and a detailed overview of the publications together with our criteria for reproducibility and reusability.

1 Introduction

Technical progress has been driving digitization in pathology over the past decade. Coupled with advances in deep learning methods, computational approaches help to localize, segment, and classify single cells and tissue types in an automated manner - and form the research field of computational pathology (see Box 1 for a glossary; (Fuchs and Buhmann, 2011). 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 (Esteva et al., 2017), lung cancer subtype classification (Coudray et al., 2018), or the recognition of malignant white blood cells (Matek et al., 2019)

However, despite the steady increase in the number of publications in this field (Fig. 1) and their promising results, only a few have reached clinical implementation (Echle et al., 2020a; van der Laak et al., 2021). This is due to several reasons: For deep learning-based methods, code availability is a natural requirement for reproducibility, 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 (Pineau et al., 2020). In particular, ready-to-use scripts with sufficient instructions or intuitive demo examples are rarely published. This makes the reuse of recent methods difficult for non-deep-learning experts, especially for pathologists who are not supported by computational experts. Another reason, which is particularly relevant to clinical implementation, is the generalization gap of algorithms in computational pathology. Often, the published performance of deep learning algorithms cannot be transferred to other datasets, due to differences in staining or scanner settings. Therefore, external validation of algorithms and statistical robustness analysis are key to assess generalizability. Finally, any algorithm employed in a clinical setting must additionally be approved by national or international authorities such as the United States Food and Drugs Administrative (FDA) or the European Medicines Agency (EMA), an often lengthy and complicated process involving business markets and legal issues, which is beyond the scope of this review.

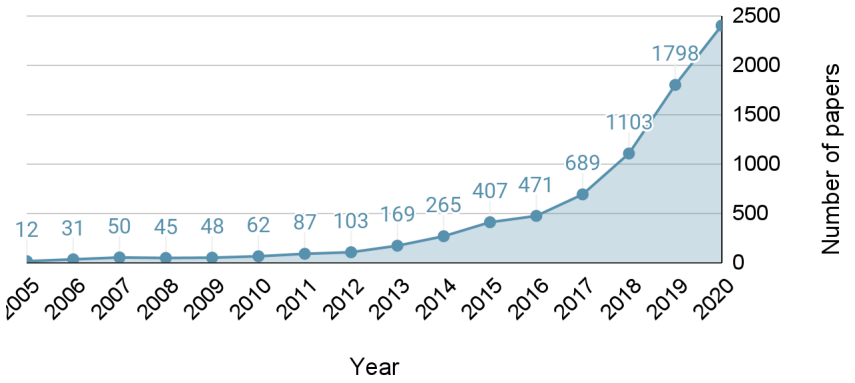


Figure 1: The number of papers published in the field of computational pathology in PubMed [retrieved on 22nd July '21] has markedly increased in the last 15 years.

Here, we focus on deep learning algorithms for computational pathology and their reproducibility and reusability. For the ultimate goal of *reusing* other deep learning algorithms, the algorithms must be *reproducible* and generalizable to similar datasets (i.e., *robust*) and external datasets (i.e., *replicable*) (see Box 2 and Fig. 2 for a definition of reproducibility and reusability and related terms). Because hematoxylin-and-eosin (H&E) staining is the most commonly used routine staining for cancer diagnosis and is often

referred to as the baseline staining (Rosai, 2007), we restrict this review to methods for H&E-stained whole-slide image (WSI) analysis. We considered five 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.

Box 1: Glossary

Digital pathology: Histological slides are scanned and digitized, such that pathologists can examine the patient material on a computer 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 histological slides with computational methods (Fuchs and Buhmann, 2011).

Specimen: A tissue sample, e.g., obtained during a biopsy or other surgical procedures, typically fixed in formalin and embedded in paraffin (FFPE).

Section: A thin slice (with a typical thickness of 3-15 µm) of a specimen mounted on a microscopic slide**.**

Whole-slide image (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 giga-pixels.

Tiles and patches: WSIs are split up into smaller images (e.g. 512x512 pixels), also called patches*,* that can be processed by neural networks. If the patches are used for subsequent WSI-rendering by stitching them back together, they are often called tiles. Unlike WSIs, these smaller units of image data allow for easier and parallelized image processing.

Annotations: Diagnostic information on pixel-level or patch-level are obtained from manual expert pathologist labeling. WSI-level annotations can be all diagnostic information about the patient (e.g., age, survival, staging, grading) mostly obtained without additional expert pathologist interaction.

Supervised learning: Training procedure of a neural network, where the ground truth, i.e., the correct label for the task, is available for each data point. However, in medical imaging and especially in computational pathology, full expert annotations at pixel-level or patch-level are very 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 multiple-instance learning (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.

Convolutional neural network (CNN): A neural network that can be trained to extract features by sliding learnable filters across the image. This makes CNNs translationally invariant and therefore well suited for histological data since important features can be found anywhere on a tile.

U-Net: A powerful CNN with an encoder-decoder architecture used for segmenting biomedical images (Ronneberger et al., 2015). It was adapted in many different ways and ranks among the most common architectures for segmentation tasks.

Mask R-CNN: A CNN architecture for instance segmentation in object detection (He et al., 2017). In contrast to U-Net that does not distinguish between instances of a class, 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.

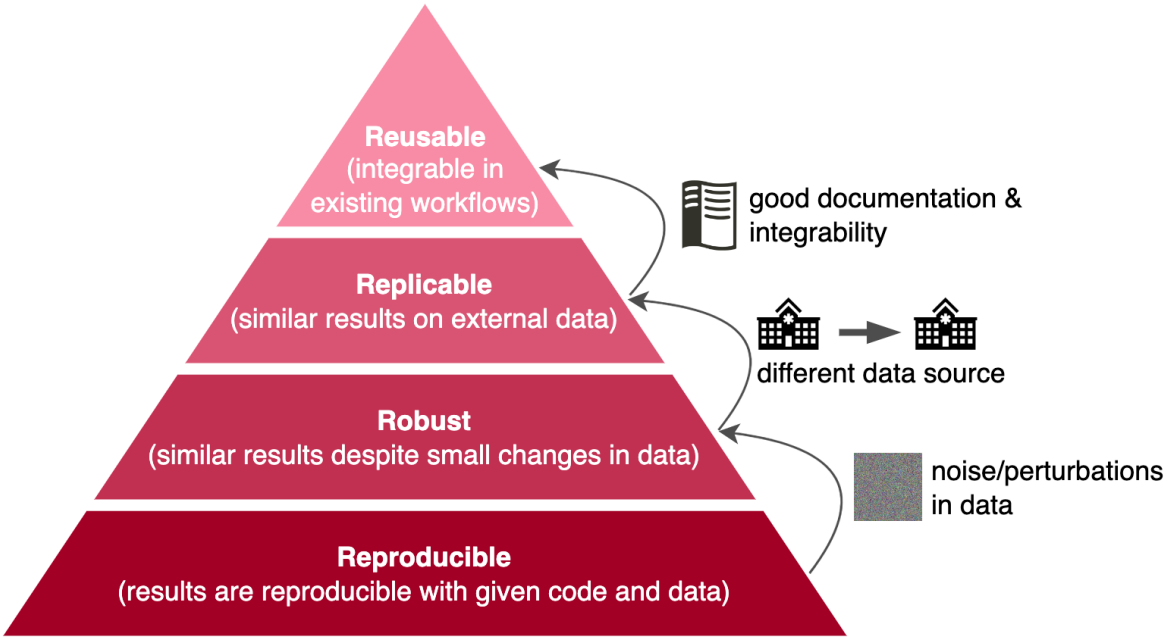


Figure 2: Reproducibility, robustness, replicability, and reusability in the context of deep learning algorithms for computational pathology.

Box 2: Definitions

Reproducibility: Using identical materials and procedures, the results of a study can be duplicated, and ultimately, identical conclusions can be drawn (Goodman et al., 2016). In the context of algorithms, the same result can be obtained from the same data, code, and analysis methods (Artner et al., 2020; Pineau et al., 2020).

Robustness: The same results are obtained from an algorithm despite small perturbations in the input (Li, 2018; Oala et al., 2020).

Replicability: Conclusions are stable based on independently acquired data (Artner et al., 2020; Pineau et al., 2020), i.e., 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, which is 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 (e.g., without the need for extensive rearrangements of the workflow).

2 Use Cases

We collected 161 papers from January 2019 and March 2021 on the automated analysis of histological slides for cancer diagnosis and treatment (Supplementary Table T1-5). We split this body of literature into the following use cases: (i) stain normalization, (ii) tissue type segmentation, (iii) evaluation of cell-level features, (iv) genetic alteration prediction, and (v) direct extraction of grading, staging, and prognostic information (Fig. 3).

In this technical chapter, we first provide a brief introduction of every use case, followed by an overview of the latest deep learning methods with a focus 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 to any use case, open and publicly available data handling tools for reading, annotating, and sharing histopathological data are essential. We compiled the most common tools in Box 3 and provide a more detailed overview in the supplementary material (Supplementary Table T0) on software features, requirements, and the possibility to extend the tool with its own code.

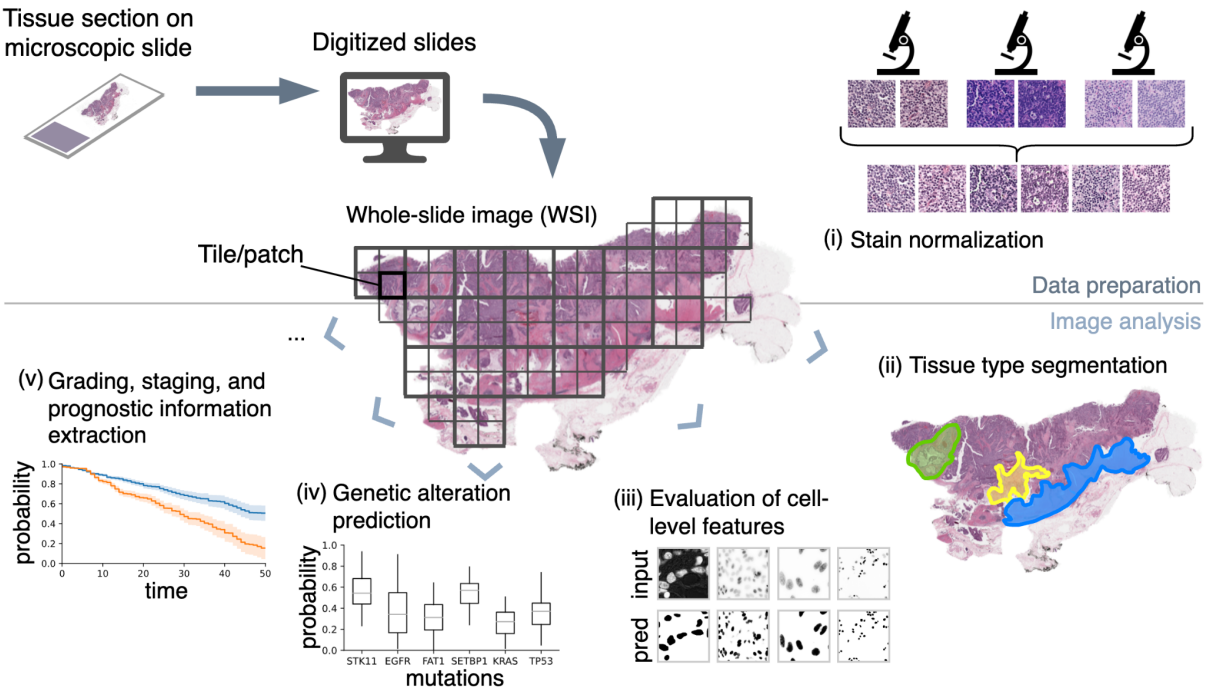


Figure 3: Overview of the use of deep learning in computational pathology including data handling tools for reading, annotating, and sharing WSIs (Box 3) and five applications of deep learning methods, which are covered in Section 2.1 - 2.5

Box 3: Data Handling Tools

A key prerequisite for implementing, transferring, and reusing computational pathology algorithms between researchers and different labs or institutions is a software structure that allows for the exchange of 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 like OpenSlide (Goode et al., 2013) or OMERO (Allan et al., 2012). They enable data to be interoperable between different vendor-specific image formats. Most of these tools also provide user interfaces for

pathologists to annotate images. Annotations can 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 (e.g., SlideRunner (Aubreville et al., 2018), QuPath (Bankhead et al., 2017), ASAP[1]) an increasing number of recently developed packages, such as Cytomine (Marée et al., 2016) or EXACT (Marzahl et al., 2021) 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 automation of the tasks, e.g., using Groovy in QuPath. Additionally, programming interfaces to popular machine learning languages such as Python have been developed, e.g., for OMERO. Several tools, such as EXACT and CaMicroscope[2], 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 T1.

2.1 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 (Chan, 2014). Digitized sections are prone to a multitude of image variations, caused by differences in the tissue preparation technique (e.g. the thickness and flatness of the sample cut), staining protocols, handling, and storage conditions. Moreover, slide scanners differ in microscope illumination, image post-processing, or noise handling. These factors lead to large variability in the visual appearance of WSIs that affects subsequent analysis and may lead to poor generalizability of algorithms. Computational methods aim at reducing the effects of these variations (Chen et al., 2017), e.g., by normalizing the stain color from a predefined source domain to one or more target domains, to arrive at a comparable visual appearance. These methods include improvements of analytical approaches, such as color deconvolution, and more recently, deep learning-based methods.

Color space methods. Color deconvolution separates the hematoxylin from the eosin component in optical density space based on a reference tile (Macenko et al., 2009). This approach has been developed further recently: Adaptive color deconvolution (Zheng et al., 2019) incorporated the underlying stain distribution of the target WSI instead of only a single tile. There, the authors assumed that each pixel is assigned to one stain and fitted a deconvolution matrix to a group of randomly sampled pixels from the source WSI using gradient-based optimization. Alternatively, non-negative matrix factorization has been used to obtain a color deconvolution matrix (Vahadane et al., 2016) and was optimized for GPU usage (Anand et al., 2019).

Generative models. The increasing popularity of Generative Adversarial Networks (GANs) has led to the development of style-transfer methods for stain normalization (Tschuchnig et al., 2020), which can be trained on all target WSI tiles instead of expert-picked reference tiles (Liang et al., 2020). StainGAN was trained with a cycle-consistency loss between the source and target domain, and the generator of the target domain was used to normalize all

[1] https://computationalpathologygroup.github.io/ASAP/
[2] https://github.com/camicroscope/caMicroscope

images in that domain (Shaban et al., 2019b). However, GANs may not always preserve the tissue structure (Cohen et al., 2018). To overcome this, StainNet trained a convolutional neural network (CNN) consisting of 1x1-convolutions and thereby transformed the source image from its original color space via intermediate color spaces to the target color space without losing structural information (Kang et al., 2020). However, this approach relies on image pairs from two different domains, which is challenging as paired images rarely exist. Alternatively, specific loss functions that compare images before and after normalization can be used to preserve the histopathological information including texture, structure, and color features added to the traditional GAN-loss that learns the stain distribution of a reference dataset (Liang et al., 2020). In contrast to normalization methods, GANs can also be used to simulate stain variability by generating synthetic images. This renders neural networks on downstream tasks more robust and avoids losing relevant information due to limitations of normalization methods. Yamashita et al. (2021) propose data augmentation based on style transfer from artistic paintings by replacing the low-level texture content with the style of artistic images. Since this produces unrealistic images for augmentation, Wagner et al. (2021) use a GAN architecture for multiple domains to synthesize realistic histological 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 pre-processing step, stain information can be incorporated directly into the model, e.g., for nuclei segmentation by creating a hematoxylin-aware CNN (Zhao et al., 2020a). This approach is based on a U-Net (Ronneberger et al., 2015) and has three branches: one for processing the input image, one for processing the hematoxylin component (retrieved from a standard color deconvolution method), and one for feature aggregation of the other branches to finally output segmentation maps.

2.2 Tissue Type Segmentation

Accurate segmentation of a WSI into tissue types (e.g., 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-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. (2021) assessed the risk of breast cancer from image features, such as the percentage of fibrous stroma, epithelium, and fatty tissue. To extract these features, they trained a 21-layer convolutional network inspired by VGG (Simonyan and Zisserman, 2014) and U-Net (Ronneberger et al., 2015) for pixel-wise segmentation of large 2048x2048 pixel tiles. Graham et al. (2020) 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. (2021) used pixel-wise segmentation for identifying six 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. From such segmented patches, Zhao et al. (2020b) computed a WSI's tumor-stroma-ratio (TSR), a prognostic factor for colorectal cancer. Here, a pretrained VGG (Simonyan and Zisserman, 2014) was fine-tuned for classifying tiles into nine tissue types to determine the TSR.

Rączkowski et al. (2019) proposed an active learning framework to train a CNN, inspired by ResNet (He et al., 2016) and DarkNet (Redmon and Farhadi, 2017) for patch classification in colorectal cancer. The network's uncertainty, estimated via Monte-Carlo dropout sampling (Gal and Ghahramani, 2015), was used to detect outlier tiles in the training set and to select them afterward for reconsideration. Wang et al. (2019) generated spatial tissue maps by classifying single cells into tumor, stroma, and lymphocytes. For this, they first extracted nuclei positions. Then, small patches centered around these nuclei were extracted and classified into the three classes using a CNN. The resulting classified positions can be used to extract spatial statistics or to generate segmentation masks using a kernel smoothing algorithm.

Hierarchical segmentation approaches mimic the workflow of pathologists by aggregating information from multiple scales of magnification. Schmitz et al. (2021) created a family of U-Net-based encoder-decoder architectures that process high- and low-resolution image tiles in separate branches from three publicly available datasets with liver, breast, and lymph node tissue. Additionally, they proposed a gate that decides whether to include the global context optimized by a classification loss to ease gradient flow through the deeper layers of the encoder for the global context. Alternatively, HookNet (van Rijthoven et al., 2021) 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 cancer.

Weakly supervised methods typically require slide-level annotations only. One recent approach was training CNNs directly on the entire WSI of lung cancer sections (Chen et al., 2021). Subsequently, class activation maps (Zhou et al., 2016) highlight relevant cancerous regions that were also identified by pathologists, which can be interpreted as a confidence measure of the algorithm. Similarly, Silva-Rodríguez et al. (2021) trained a feature extraction network on the entire, downsampled WSI for the classification of 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 used for WSI classification (Lu et al., 2021). Using this approach, attention scores produce interpretable heatmaps to visualize which regions contribute to the network's prediction.

2.3 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 two important features for tumor grading. A multitude of deep learning approaches has been proposed for these and related tasks (Lagree et al., 2021). Here we focus on two of the most widely studied cell-level tasks, namely segmentation of nuclei and detection of mitoses.

Nuclei segmentation. To benchmark efforts in this field, segmentation challenges have been introduced, like the Multi-Organ Nucleus Segmentation Challenge (Kumar et al., 2020), which provided 30 images and around 22,000 nuclear boundary annotations in a public dataset. Out of the top six participants, three used U-Net-based semantic segmentation (Ronneberger et al., 2015), two used Mask R-CNN-based instance segmentation (He et al., 2017), and one group used stacked U-Net and R-CNN models. The two dominating algorithms have been further tailored towards nuclear segmentation: Cui et al. (2019) predicted a boundary map additionally to object segmentations to separate touching nuclei efficiently. Jin et al. (2020) incorporated a U-Net into a pipeline to detect lymph node

metastasis in breast cancer patients, integrating multiple segmentation channels for nuclei, mitosis, tubule, etc. in the U-Net input. The Mask R-CNN has successfully been combined with a deep convolutional gaussian mixture color normalization model, which clusters pixels according to nucleus morphology (Jung et al., 2019), where the authors performed multiple interferences and post-processing steps to boost segmentation performance. Recently, other approaches such as GANs have been proposed (Mahmood et al., 2020a), where the network is trained on unpaired data to map segmentation masks to nuclei images. To ease the annotation process for nuclei segmentation, Qu et al. (2020) provided a deep learning framework trained on incomplete annotations, which are much easier to generate. Their two-stage approach first detected nuclei locations based on the partial annotations, before self-training with background propagation was applied to boost nuclei detection. In the second stage, a segmentation model was trained with this data in a weakly supervised fashion.

Mitosis detection. Identifying cells that are in the mitotic phases of the cell cycle is a diagnostically relevant task e.g. for breast cancer grading and prognosis (Veta et al., 2019). Several challenges released public datasets and benchmarked competing approaches for mitosis detection, e.g., ICPR MITOS-2012 (Roux et al., 2013), ICPR MITOS-ATYPIA-2014 (Roux et al., 2014), or TUPAC 16 (Veta et al., 2019). Most of the recently developed mitosis detection methods can be grouped into three categories: classification, segmentation, and detection. As one of the first deep learning applications in the medical field, Cireşan et al. (2013) trained a network to classify each pixel in an image by considering a tile centered around that pixel whilst tiles are extracted in a sliding-window fashion. This process was accelerated by precomputing the hematoxylin fraction of the H&E staining, which highlighted nuclei as mitosis candidates and hence restricted tile extraction (Saha et al., 2018). More recently, Pati et al. (2021) combined a classification task with metric learning to reduce the necessary amount of labeled data for more efficient network training. Another approach for mitosis detection is pixel-wise semantic segmentation. Jiménez and Racoceanu (2019) showed that a U-net-based semantic segmentation approach led to higher accuracy than previous classification approaches. Lafarge et al. (2021) 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 papers on mitosis detection were based on object detection (Lei et al., 2019, 2021; Sohail et al., 2021; Wollmann and Rohr, 2021), 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 (Mahmood et al., 2020b).

2.4 Genetic alteration prediction

As genetic alterations can carry crucial predictive and prognostic information, they have become increasingly relevant to the diagnostic workup and the selection of therapeutic pathways (Ashley, 2016). Therefore, patients are profiled for genetic alterations that characterize their disease, e.g., in colorectal cancer (Singh et al., 2021), 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, and 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 on formalin-fixed paraffin-embedded tissue that is usually used for histological sample preparation. Therefore, algorithm-based prognostic stratification and mutation prediction from H&E-stained WSIs offers an attractive, cost- and time-effective as well as tissue-sparing addition to existing molecular characterization methods.

Image-based mutation prediction. While some genetic alterations, such as mutations, copy number variations, and translocations can be relevant for disease characterization, the majority of work has so far focussed on mutation prediction from imaging data. Coudray et al. (2018) were the first to address this topic. They found that six 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 (Kim et al., 2019; Zhang et al., 2020b), breast cancer (Anand et al., 2020; Bychkov et al., 2021; Lu et al., 2020), lung cancer (Wang et al., 2020; Yu et al., 2020), colorectal cancer (Cao et al., 2020; Echle et al., 2020b; Jang et al., 2020), bladder cancer (Woerl et al., 2020), and thyroid carcinoma (Tsou and Wu, 2019). Several recent studies attempt a pan-cancer approach that predicts genetic alteration across multiple tissue types from WSIs directly (Kather et al., 2020; Noorbakhsh et al., 2020).

Modeling strategies for mutation prediction. Most approaches rely on standardized processing pipelines from pre-processing (see Section 2.1) and region of interest extraction (Section 2.2) to model training and evaluation. As network architectures, common CNN models such as Inception (Szegedy et al., 2016) or ResNet (He et al., 2016) are used for per-tile prediction, where all tiles from a patient WSI inherit the same label. This label can be both continuous (e.g., tumor mutational burden) or categorical (e.g., the mutation status of a selected gene, or the microsatellite status) (Coudray et al., 2018; Kather et al., 2020). 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. (2020) employ multiple instance learning (MIL) since it does not need instance labels for each tile and achieves better accuracy than standard supervised learning methods. Fu et al. (2020) did not train an end-to-end neural network for direct mutation prediction but rather classified each tile into different malignant and non-malignant tissue types. This classification network was used to extract features in a pan-cancer fashion and, subsequently, to predict driver gene mutations. Almost all studies mentioned above, except for Bychkov et al. (2021), trained their networks on publicly available data from The Cancer Genome Atlas (TCGA) (Gutman et al., 2013).

2.5 Grading, Staging, and Prognostic Information Extraction

A typical goal of the analysis of pathology slides is not only to recognize and evaluate primary lesions but also to determine their histopathologic subtype and grade (as defined by respective WHO classifications for different tumor entities) and to derive therapeutically relevant information from these features. In the context of computational pathology, this set of tasks can be addressed through the determination of features that are known to possess prognostic or predictive value. Alternatively, it can be attempted to extract prognostic or predictive information directly from imaging data, molecular properties, or clinical data. 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 (Bulten et al., 2021; Nagpal et al., 2019; Steiner et al., 2020), grading of gliomas (Rathore et al., 2020; Truong et al., 2020), and automated evaluation of mitoses (Chang and Mrkonjic, 2020; Pantanowitz et al., 2020) or tumor-infiltrating lymphocytes in breast (Balkenhol et al., 2021) and head and neck cancer (Shaban et al., 2019a). Machine learning methods have also been applied in disease staging, e.g., 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 (Ghosh et al., 2021) or by predicting the risk of lymph node metastasis from the primary lesion in the case of bladder cancer (Harmon et al., 2020). Several studies inferred molecular properties with a prognostic value from H&E, such as microsatellite instability in gastrointestinal cancer (Kather et al., 2019a), or the molecular subtype of invasive bladder cancer (Woerl et al., 2020).

Image-based biomarkers. Finally, prognostic factors can be derived directly either from histopathology images (Zhao et al., 2020b) or in combination with other clinical or molecular data, e.g., from the genome or transcriptome (Failmezger et al., 2020; Hao et al., 2020; Zhao et al., 2020b). Potentially, this route can be followed without referring to previously known prognostic factors. Hence, while these approaches may first rely on computational predictions alone, they may also lead to the identification of novel prognostic or predictive factors that lend themselves to direct human evaluation, which can be identified through explainability methods (Tosun et al., 2020). Examples of this approach include automated quantification of intratumoral stroma in rectal cancer (Geessink et al., 2019), evaluation of nuclear morphology for survival prediction in lung cancer (Alsubaie et al., 2021), deep learning-based prognosis in nasopharyngeal cancer (Geessink et al., 2019; Zhang et al., 2020a), and survival prediction in colorectal cancer (Abbet et al., 2020; Kather et al., 2019b).

3 Methods

To assess reproducibility and reusability in computational pathology, we (i) monitored whether and how code was publicly available, (ii) evaluated criteria for data access, and (iii) checked if the statistical variance of the reported findings was provided. In Supplementary Table T2, we list all 42 publications (out of 161) together with the following evaluation criteria that we used for code, data, and statistical variance:

a) Inspired by the FAIR principles demanding that data should be findable, accessible, interoperable, and reusable (Wilkinson et al., 2016), we surveyed whether the code was made publicly available, in addition to also noting the platform that was used for sharing and the programming language and machine learning frameworks that were employed. Further, we checked for instructions for running the code, whether the code was minimally documented, and if a pretrained model was available for direct application.
b) For the 42 publications with available code, we evaluated data access and checked whether the dataset and required annotations were publicly available. Additionally, we recorded what kind of data had been used (e.g., tiled WSIs versus entire WSIs). We also reported whether the relevant pre-processing steps were provided as well as the training – validation – test split that was 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 whether it covered an independent cohort.

c) 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 like floating-point numbers that can be rounded differently on different machines, architectures, or execution environments (Hill, 2019), 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 and perform a sufficient exploration of hyperparameters (Pineau et al., 2020). A straightforward evaluation approach is to repeat the experiment multiple times and report mean and standard deviation across experiments, or over several folds of cross-validation.

4 Results

Code availability. In our study, 42 out of 161 publications (26%) made their code publicly available (Fig. 3a). Interestingly, the ratio of publications with code differs across the five use cases: For stain normalization, we retrieved 29 research papers, where only 7 (24%) of them provided code with their method. In the field of tissue type segmentation and localization, only 12 out of our total 51 investigated papers (24%) had their code publicly available and only three publications provided the pretrained model weights. Among the 28 research papers that we screened for the evaluation of cell-level features, 11 papers (39%) provided code with the publications. The code of 5 of these 11 could be run in Google Colab and thus was directly applicable. For genetic alteration prediction, 8 out of 13 papers (61%) have provided their code along with their method. In survival analysis, only 4 out of 38 studies (11%) published code. Interestingly, genetic alteration prediction has the highest ratio of published code. One reason for this could be that a key publication for genetic alteration prediction published in Nature Medicine included a well-documented codebase (Coudray et al., 2018); most subsequent publications build on this work and could therefore publish their code more easily. 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 42 papers that published their code, we checked the evaluation criteria for code, data, and statistical variance detailed in Section 3 and detailed them in Supplementary Table T2.

Model weights and frameworks. Pretrained model weights were only available for half of the publications that also provided code (Fig. 3b); this renders a reproduction of experimental results difficult, in particular, because the pre-processing steps are rarely available. Also, without model weights a direct application without retraining the model is not possible, hence hampering its use by pathologists that are not specialized in deep learning. Almost 75% of the methods were implemented in Python using the open and publicly available machine learning frameworks TensorFlow[3] (38%) and PyTorch[4] (36%; Fig. 4c).

Datasets. More than half of all methods (57%) were evaluated on publicly available datasets (Fig. 3d). Most studies (e.g., all 13 studies reviewed for generic alteration prediction) developed their methods based on TCGA (Gutman et al., 2013) it contains data from multiple institutions; thus, it can be split into training and test sets by cohort level. It was also a common practice to complement TCGA with external, mostly private data as an

[3] https://www.tensorflow.org/
[4] https://pytorch.org/

independent evaluation cohort. Nevertheless, TCGA, with a few hundred slides for each cancer type, is not sufficient to represent all cancer heterogeneity. The reliance of computational pathology on relatively few publicly available datasets renders their selection strategy and processing critical. Batch effects can be detected by deep learning models and lead to overestimation of the model's performance (Howard et al., 2020). Therefore, we strongly encourage the development of more publicly available multi-institutional datasets.

Statistical variance. We believe that a thorough evaluation of sources and magnitude of variability, both on an algorithmic and a data level, is an important step towards 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. 3f). Many different sources of variability can be relevant to the performance of computational pathology algorithms, and it may therefore be difficult to devise a single strategy for quantifying all sources.

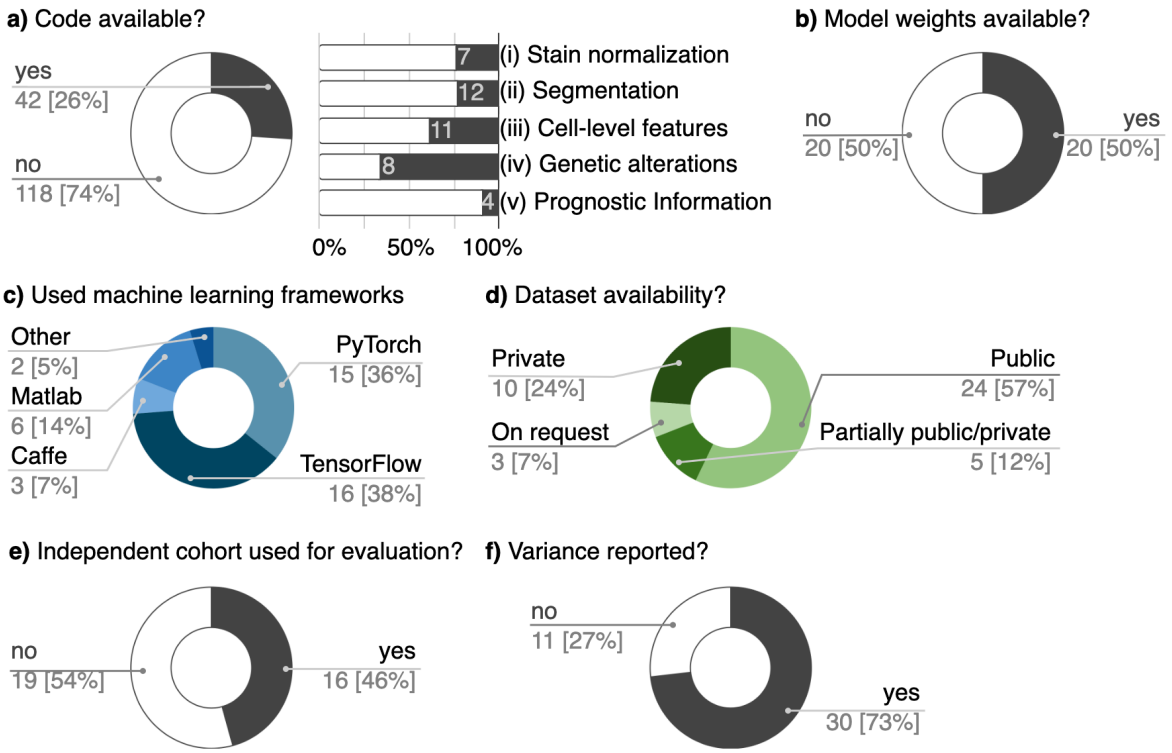


Figure 4: Analysis of our systematic literature search on 161 computational pathology papers. a) The proportion of methods with publicly available code (26%) differs across the use cases. b) Half of the publications with code release their final model weights. c) Most works (74%) used PyTorch or Tensorflow as machine learning frameworks. d) Mainly, large public datasets are used and sometimes complemented by private cohorts. e) Almost half of the publications used an independent cohort for evaluation. f) The largest part analyzes their results statistically.

5 Conclusion

It has been increasingly recognized that computational reproducibility and reusability are an essential part of good scientific practice for deep learning applications (Haibe-Kains et al., 2020; Hutson, 2018; Stodden et al., 2016). Especially for the interdisciplinary field of

computational pathology, both are key requirements for enabling a wider use of algorithms and eventually a clinical application.

In our survey of recently published computational pathology deep learning 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 histological 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 (Section 2.1), their use in downstream applications is still limited, probably because pre-trained stain normalization models are rarely available and in most cases code is not shared. Instead, we observe that easy-to-use algorithms without further model training are typically applied. The lack of reusability hinders the practical application of innovative network-based methods.

Even if the code is shared, supporting documentation or convincing experiments on external cohorts are often missing, hence lowering the chances of a successful reuse and translation. Most state-of-the-art methods in computational pathology are based on deep learning algorithms, and typically require large amounts of labeled training data. Making this data available is as crucial as providing well documented code. We acknowledge that in some cases, data and appropriate annotations cannot be publicly shared, e.g. due to legal or ethical constraints. Here, reasonable compromises like partial data sharing or evaluation using public datasets (Haibe-Kains et al., 2020) should be considered.

Despite the lack of reproducibility and reusability in many computational pathology approches, we hope that the field will profit from the surging discussions, e.g. computer vision (Pineau et al., 2020). As a step in this direction, large conferences, such as MICCAI since 2021, started to employ reproducibility checklists for authors in their submission form that will be publicly available upon acceptance of the paper. 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., Ch.M., T.P., Ca.M. designed the concept of the review. M.B., Ch.M. supported with domain-related advice. S.J.W., L.L., D.W. compiled the evaluation criteria for reproducibility and reusability. S.J.W., Ch.M. assessed current data handling tools. For the use cases, S.J.W., D.W., S.S.B. reviewed the publications for stain normalization, L.L., S.J.W. for tissue type segmentation, D.W., T.P. for evaluation of cell-level features, T.P. for genetic alteration prediction, A.S., Ch.M. for grading, staging, and prognostic information extraction. S.J.W. analyzed the results. All authors contributed to constructive discussions. S.J.W. wrote the manuscript with Ch.M, Ca.M, T.P. and the input from all other authors.

Acknowledgment

We thank Peter Schüffler (Munich) for inspiring feedback. 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”. C.M. has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (Grant agreement No. 866411)

References

Abbet, C., Zlobec, I., Bozorgtabar, B., and Thiran, J.-P. (2020). Divide-and-Rule: Self-Supervised Learning for Survival Analysis in Colorectal Cancer. Medical Image Computing and Computer Assisted Intervention – MICCAI 2020 480–489.

Allan, C., Burel, J.-M., Moore, J., Blackburn, C., Linkert, M., Loynton, S., Macdonald, D., Moore, W.J., Neves, C., Patterson, A., et al. (2012). OMERO: flexible, model-driven data management for experimental biology. Nat. Methods *9*, 245–253.

Alsubaie, N.M., Snead, D., and Rajpoot, N.M. (2021). Tumour Nuclear Morphometrics Predict Survival in Lung Adenocarcinoma. IEEE Access *9*, 12322–12331.

Anand, D., Ramakrishnan, G., and Sethi, A. (2019). Fast GPU-Enabled Color Normalization for Digital Pathology. In 2019 International Conference on Systems, Signals and Image Processing (IWSSIP), pp. 219–224.

Anand, D., Kurian, N.C., Dhage, S., Kumar, N., Rane, S., Gann, P.H., and Sethi, A. (2020). Deep Learning to Estimate Human Epidermal Growth Factor Receptor 2 Status from Hematoxylin and Eosin-Stained Breast Tissue Images. J. Pathol. Inform. *11*, 19.

Artner, R., Verliefde, T., Steegen, S., Gomes, S., Traets, F., Tuerlinckx, F., and Vanpaemel, W. (2020). The reproducibility of statistical results in psychological research: An investigation using unpublished raw data. Psychol. Methods.

Ashley, E.A. (2016). Towards precision medicine. Nat. Rev. Genet. *17*, 507–522.

Aubreville, M., Bertram, C., Klopfleisch, R., and Maier, A. (2018). SlideRunner. In Bildverarbeitung Für Die Medizin 2018, (Springer Berlin Heidelberg), pp. 309–314.

Balkenhol, M.C., Ciompi, F., Świderska-Chadaj, Ż., van de Loo, R., Intezar, M., Otte-Höller, I., Geijs, D., Lotz, J., Weiss, N., de Bel, T., et al. (2021). Optimized tumour infiltrating lymphocyte assessment for triple negative breast cancer prognostics. Breast *56*, 78–87.

Bankhead, P., Loughrey, M.B., Fernández, J.A., Dombrowski, Y., McArt, D.G., Dunne, P.D., McQuaid, S., Gray, R.T., Murray, L.J., Coleman, H.G., et al. (2017). QuPath: Open source software for digital pathology image analysis. Sci. Rep. *7*, 16878.

Bulten, W., Balkenhol, M., Belinga, J.-J.A., Brilhante, A., Çakır, A., Egevad, L., Eklund, M., Farré, X., Geronatsiou, K., Molinié, V., et al. (2021). Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. Mod. Pathol. *34*, 660–671.

Bychkov, D., Linder, N., Tiulpin, A., Kücükel, H., Lundin, M., Nordling, S., Sihto, H., Isola, J., Lehtimäki, T., Kellokumpu-Lehtinen, P.-L., et al. (2021). Deep learning identifies morphological features in breast cancer predictive of cancer ERBB2 status and trastuzumab treatment efficacy. Scientific Reports *11*.

Cao, R., Yang, F., Ma, S.-C., Liu, L., Zhao, Y., Li, Y., Wu, D.-H., Wang, T., Lu, W.-J., Cai, W.-J., et al. (2020). Development and interpretation of a pathomics-based model for the prediction of microsatellite instability in Colorectal Cancer. Theranostics *10*, 11080–11091.

Chan, J.K.C. (2014). The wonderful colors of the hematoxylin-eosin stain in diagnostic surgical pathology. Int. J. Surg. Pathol. *22*, 12–32.

Chang, M.C., and Mrkonjic, M. (2020). Review of the current state of digital image analysis in breast pathology. Breast J. *26*, 1208–1212.

Chen, C.-L., Chen, C.-C., Yu, W.-H., Chen, S.-H., Chang, Y.-C., Hsu, T.-I., Hsiao, M., Yeh, C.-Y., and Chen, C.-Y. (2021). An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning. Nat. Commun. *12*, 1193.

Chen, J.-M., Li, Y., Xu, J., Gong, L., Wang, L.-W., Liu, W.-L., and Liu, J. (2017). Computer-aided prognosis on breast cancer with hematoxylin and eosin histopathology images: A review. Tumour Biol. *39*, 1010428317694550.

Cireşan, D.C., Giusti, A., Gambardella, L.M., and Schmidhuber, J. (2013). Mitosis detection in breast cancer histology images with deep neural networks. Med. Image Comput. Comput. Assist. Interv. *16*, 411–418.

Cohen, J.P., Luck, M., and Honari, S. (2018). Distribution Matching Losses Can Hallucinate Features in Medical Image Translation. In Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, (Springer International Publishing), pp. 529–536.

Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., and Tsirigos, A. (2018). Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat. Med. *24*, 1559–1567.

Cui, Y., Zhang, G., Liu, Z., Xiong, Z., and Hu, J. (2019). A deep learning algorithm for one-step contour aware nuclei segmentation of histopathology images. Med. Biol. Eng. Comput. *57*, 2027–2043.

Echle, A., Rindtorff, N.T., Brinker, T.J., Luedde, T., Pearson, A.T., and Kather, J.N. (2020a). Deep learning in cancer pathology: a new generation of clinical biomarkers. Br. J. Cancer.

Echle, A., Grabsch, H.I., Quirke, P., van den Brandt, P.A., West, N.P., Hutchins, G.G.A., Heij, L.R., Tan, X., Richman, S.D., Krause, J., et al. (2020b). Clinical-Grade Detection of Microsatellite Instability in Colorectal Tumors by Deep Learning. Gastroenterology *159*, 1406–1416.e11.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature.

Failmezger, H., Muralidhar, S., Rullan, A., de Andrea, C.E., Sahai, E., and Yuan, Y. (2020). Topological Tumor Graphs: A Graph-Based Spatial Model to Infer Stromal Recruitment for Immunosuppression in Melanoma Histology. Cancer Res. *80*, 1199–1209.

Fu, Y., Jung, A.W., Torne, R.V., Gonzalez, S., Vöhringer, H., Shmatko, A., Yates, L.R., Jimenez-Linan, M., Moore, L., and Gerstung, M. (2020). Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. Nature Cancer.

Fuchs, T.J., and Buhmann, J.M. (2011). Computational pathology: challenges and promises for tissue analysis. Comput. Med. Imaging Graph. *35*, 515–530.

Gal, Y., and Ghahramani, Z. (2015). Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference.

Geessink, O.G.F., Baidoshvili, A., Klaase, J.M., Ehteshami Bejnordi, B., Litjens, G.J.S., van Pelt, G.W., Mesker, W.E., Nagtegaal, I.D., Ciompi, F., and van der Laak, J.A.W.M. (2019). Computer aided quantification of intratumoral stroma yields an independent prognosticator in rectal cancer. Cell. Oncol. *42*, 331–341.

Ghosh, A., Sirinukunwattana, K., Khalid Alham, N., Browning, L., Colling, R., Protheroe, A., Protheroe, E., Jones, S., Aberdeen, A., Rittscher, J., et al. (2021). The Potential of Artificial Intelligence to Detect Lymphovascular Invasion in Testicular Cancer. Cancers *13*.

Goode, A., Gilbert, B., Harkes, J., Jukic, D., and Satyanarayanan, M. (2013). OpenSlide: A vendor-neutral

software foundation for digital pathology. J. Pathol. Inform. *4*, 27.

Goodman, S.N., Fanelli, D., and Ioannidis, J.P.A. (2016). What does research reproducibility mean? Sci. Transl. Med. *8*, 341ps12.

Graham, S., Epstein, D., and Rajpoot, N. (2020). Dense Steerable Filter CNNs for Exploiting Rotational Symmetry in Histology Images. IEEE Trans. Med. Imaging *39*, 4124–4136.

Gutman, D.A., Cobb, J., Somanna, D., Park, Y., Wang, F., Kurc, T., Saltz, J.H., Brat, D.J., and Cooper, L.A.D. (2013). Cancer Digital Slide Archive: an informatics resource to support integrated in silico analysis of TCGA pathology data. J. Am. Med. Inform. Assoc. *20*, 1091–1098.

Haibe-Kains, B., Adam, G.A., Hosny, A., Khodakarami, F., Massive Analysis Quality Control (MAQC) Society Board of Directors, Waldron, L., Wang, B., McIntosh, C., Goldenberg, A., Kundaje, A., et al. (2020). Transparency and reproducibility in artificial intelligence. Nature *586*, E14–E16.

Hao, J., Kosaraju, S.C., Tsaku, N.Z., Song, D.H., and Kang, M. (2020). PAGE-Net: Interpretable and Integrative Deep Learning for Survival Analysis Using Histopathological Images and Genomic Data. Pac. Symp. Biocomput. *25*, 355–366.

Harmon, S.A., Sanford, T.H., Brown, G.T., Yang, C., Mehralivand, S., Jacob, J.M., Valera, V.A., Shih, J.H., Agarwal, P.K., Choyke, P.L., et al. (2020). Multiresolution Application of Artificial Intelligence in Digital Pathology for Prediction of Positive Lymph Nodes From Primary Tumors in Bladder Cancer. JCO Clin Cancer Inform *4*, 367–382.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969.

Hill, D.R.C. (2019). Repeatability, Reproducibility, Computer Science and High Performance Computing : Stochastic simulations can be reproducible too…. In 2019 International Conference on High Performance Computing Simulation (HPCS), pp. 322–323.

Howard, F.M., Dolezal, J., Kochanny, S., Schulte, J., Chen, H., Heij, L., Huo, D., Nanda, R., Olopade, O.I., Kather, J.N., et al. (2020). The Impact of Digital Histopathology Batch Effect on Deep Learning Model Accuracy and Bias.

Hutson, M. (2018). Artificial intelligence faces reproducibility crisis. Science *359*, 725–726.

Jang, H.-J., Lee, A., Kang, J., Song, I.H., and Lee, S.H. (2020). Prediction of clinically actionable genetic alterations from colorectal cancer histopathology images using deep learning. World J. Gastroenterol. *26*, 6207–6223.

Jayapandian, C.P., Chen, Y., Janowczyk, A.R., Palmer, M.B., Cassol, C.A., Sekulic, M., Hodgin, J.B., Zee, J., Hewitt, S.M., O'Toole, J., et al. (2021). Development and evaluation of deep learning-based segmentation of histologic structures in the kidney cortex with multiple histologic stains. Kidney Int. *99*, 86–101.

Jiménez, G., and Racoceanu, D. (2019). Deep Learning for Semantic Segmentation vs. Classification in Computational Pathology: Application to Mitosis Analysis in Breast Cancer Grading. Front Bioeng Biotechnol *7*, 145.

Jin, Y.W., Jia, S., Ashraf, A.B., and Hu, P. (2020). Integrative Data Augmentation with U-Net Segmentation Masks Improves Detection of Lymph Node Metastases in Breast Cancer Patients. Cancers *12*.

Jung, H., Lodhi, B., and Kang, J. (2019). An automatic nuclei segmentation method based on deep convolutional neural networks for histopathology images. BMC Biomed Eng *1*, 24.

Kang, H., Luo, D., Feng, W., Hu, J., Zeng, S., Quan, T., and Liu, X. (2020). StainNet: a fast and robust stain normalization network.

Kather, J.N., Pearson, A.T., Halama, N., Jäger, D., Krause, J., Loosen, S.H., Marx, A., Boor, P., Tacke, F., Neumann, U.P., et al. (2019a). Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat. Med. *25*, 1054–1056.

Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.-A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D., et al. (2019b). Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS Med. *16*, e1002730.

Kather, J.N., Heij, L.R., Grabsch, H.I., Loeffler, C., Echle, A., Muti, H.S., Krause, J., Niehues, J.M., Sommer, K.A.J., Bankhead, P., et al. (2020). Pan-cancer image-based detection of clinically actionable genetic alterations. Nature Cancer *1*, 789–799.

Kim, R.H., Nomikou, S., Dawood, Z., Jour, G., Donnelly, D., Moran, U., Weber, J.S., Razavian, N., Snuderl, M., Shapiro, R., et al. (2019). A Deep Learning Approach for Rapid Mutational Screening in Melanoma.

Kumar, N., Verma, R., Anand, D., Zhou, Y., Onder, O.F., Tsougenis, E., Chen, H., Heng, P.-A., Li, J., Hu, Z., et al. (2020). A Multi-Organ Nucleus Segmentation Challenge. IEEE Trans. Med. Imaging *39*, 1380–1391.

van der Laak, J., Litjens, G., and Ciompi, F. (2021). Deep learning in histopathology: the path to the clinic. Nat. Med. *27*, 775–784.

Lafarge, M.W., Bekkers, E.J., Pluim, J.P.W., Duits, R., and Veta, M. (2021). Roto-translation equivariant convolutional networks: Application to histopathology image analysis. Med. Image Anal. *68*, 101849.

Lagree, A., Mohebpour, M., Meti, N., Saednia, K., Lu, F.-I., Slodkowska, E., Gandhi, S., Rakovitch, E., Shenfield, A., Sadeghi-Naini, A., et al. (2021). A review and comparison of breast tumor cell nuclei segmentation performances using deep convolutional neural networks. Sci. Rep. *11*, 8025.

Lei, H., Liu, S., Xie, H., Kuo, J.Y., and Lei, B. (2019). An Improved Object Detection Method for Mitosis Detection. Conf. Proc. IEEE Eng. Med. Biol. Soc. *2019*, 130–133.

Lei, H., Liu, S., Elazab, A., Gong, X., and Lei, B. (2021). Attention-Guided Multi-Branch Convolutional Neural Network for Mitosis Detection From Histopathological Images. IEEE J Biomed Health Inform *25*, 358–370.

Li, J.Z. (2018). Principled approaches to robust machine learning and beyond. Massachusetts Institute of Technology.

Liang, H., Plataniotis, K.N., and Li, X. (2020). Stain Style Transfer of Histopathology Images via Structure-Preserved Generative Learning. In Machine Learning for Medical Image Reconstruction, (Springer International Publishing), pp. 153–162.

Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Chen, R.J., Barbieri, M., and Mahmood, F. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng.

Lu, Z., Xu, S., Shao, W., Wu, Y., Zhang, J., Han, Z., Feng, Q., and Huang, K. (2020). Deep-Learning-Based Characterization of Tumor-Infiltrating Lymphocytes in Breast Cancers From Histopathology Images and Multiomics Data. JCO Clin Cancer Inform *4*, 480–490.

Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., and Thomas, N.E. (2009). A method for normalizing histology slides for quantitative analysis. In 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, (IEEE),.

Mahmood, F., Borders, D., Chen, R.J., Mckay, G.N., Salimian, K.J., Baras, A., and Durr, N.J. (2020a). Deep Adversarial Training for Multi-Organ Nuclei Segmentation in Histopathology Images. IEEE Trans. Med. Imaging *39*, 3257–3267.

Mahmood, T., Arsalan, M., Owais, M., Lee, M.B., and Park, K.R. (2020b). Artificial Intelligence-Based Mitosis Detection in Breast Cancer Histopathology Images Using Faster R-CNN and Deep CNNs. J. Clin. Med. Res. *9*.

Marée, R., Rollus, L., Stévens, B., Hoyoux, R., Louppe, G., Vandaele, R., Begon, J.-M., Kainz, P., Geurts, P., and Wehenkel, L. (2016). Collaborative analysis of multi-gigapixel imaging data using Cytomine. Bioinformatics *32*, 1395–1401.

Marzahl, C., Aubreville, M., Bertram, C.A., Maier, J., Bergler, C., Kröger, C., Voigt, J., Breininger, K., Klopfleisch, R., and Maier, A. (2021). EXACT: a collaboration toolset for algorithm-aided annotation of images with annotation version control. Sci. Rep. *11*, 4343.

Matek, C., Schwarz, S., Spiekermann, K., and Marr, C. (2019). Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. Nat Mach Intell *1*, 538–544.

Nagpal, K., Foote, D., Liu, Y., Chen, P.-H.C., Wulczyn, E., Tan, F., Olson, N., Smith, J.L., Mohtashamian, A., Wren, J.H., et al. (2019). Erratum: Publisher Correction: Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. NPJ Digit Med *2*, 113.

Noorbakhsh, J., Farahmand, S., Foroughi Pour, A., Namburi, S., Caruana, D., Rimm, D., Soltanieh-Ha, M., Zarringhalam, K., and Chuang, J.H. (2020). Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images. Nat. Commun. *11*, 6367.

Oala, L., Fehr, J., Gilli, L., Balachandran, P., Leite, A.W., Calderon-Ramirez, S., Li, D.X., Nobis, G., Alvarado, E.A.M., Jaramillo-Gutierrez, G., et al. (2020). ML4H Auditing: From Paper to Practice. In Proceedings of the Machine Learning for Health NeurIPS Workshop, E. Alsentzer, M.B.A. McDermott, F. Falck, S.K. Sarkar, S. Roy, and S.L. Hyland, eds. (PMLR), pp. 280–317.

Pantanowitz, L., Hartman, D., Qi, Y., Cho, E.Y., Suh, B., Paeng, K., Dhir, R., Michelow, P., Hazelhurst, S., Song, S.Y., et al. (2020). Accuracy and efficiency of an artificial intelligence tool when counting breast mitoses. Diagn. Pathol. *15*, 80.

Pati, P., Foncubierta-Rodríguez, A., Goksel, O., and Gabrani, M. (2021). Reducing annotation effort in digital pathology: A Co-Representation learning framework for classification tasks. Med. Image Anal. *67*, 101859.

Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Larochelle, H. (2020). Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program).

Qu, H., Wu, P., Huang, Q., Yi, J., Yan, Z., Li, K., Riedlinger, G.M., De, S., Zhang, S., and Metaxas, D.N. (2020). Weakly Supervised Deep Nuclei Segmentation Using Partial Points Annotation in Histopathology Images. IEEE Trans. Med. Imaging *39*, 3655–3666.

Rączkowski, Ł., Możejko, M., Zambonelli, J., and Szczurek, E. (2019). ARA: accurate, reliable and active histopathological image classification framework with Bayesian deep learning. Sci. Rep. *9*, 14347.

Rathore, S., Niazi, T., Iftikhar, M.A., and Chaddad, A. (2020). Glioma Grading via Analysis of Digital Pathology Images Using Machine Learning. Cancers *12*.

Redmon, J., and Farhadi, A. (2017). YOLO9000: better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271.

van Rijthoven, M., Balkenhol, M., Siliņa, K., van der Laak, J., and Ciompi, F. (2021). HookNet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images. Med. Image Anal. *68*, 101890.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, (Springer International Publishing), pp. 234–241.

Rosai, J. (2007). Why microscopy will remain a cornerstone of surgical pathology. Lab. Invest. *87*, 403–408.

Roux, L., Racoceanu, D., Loménie, N., Kulikova, M., Irshad, H., Klossa, J., Capron, F., Genestie, C., Naour, G., and Gurcan, M. (2013). Mitosis detection in breast cancer histological images An ICPR 2012 contest. Journal of Pathology Informatics *4*, 8.

Roux, L., Racoceanu, D., Capron, F., Calvo, J., Attieh, E., Le Naour, G., and Gloaguen, A. (2014). MITOS & ATYPIA-Detection of mitosis and evaluation of nuclear atypia score in breast cancer histological images. IPAL, Agency Sci, Technol Res Inst Infocom Res. Technol. Res. Inst. Infocom Res. , Singapore, Tech. Rep.

Saha, M., Chakraborty, C., and Racoceanu, D. (2018). Efficient deep learning model for mitosis detection using breast histopathology images. Comput. Med. Imaging Graph. *64*, 29–40.

Schmitz, R., Madesta, F., Nielsen, M., Krause, J., Steurer, S., Werner, R., and Rösch, T. (2021). Multi-scale fully convolutional neural networks for histopathology image segmentation: From nuclear aberrations to the global tissue architecture. Med. Image Anal. *70*, 101996.

Shaban, M., Khurram, S.A., Fraz, M.M., Alsubaie, N., Masood, I., Mushtaq, S., Hassan, M., Loya, A., and Rajpoot, N.M. (2019a). A Novel Digital Score for Abundance of Tumour Infiltrating Lymphocytes Predicts Disease Free Survival in Oral Squamous Cell Carcinoma. Sci. Rep. *9*, 13341.

Shaban, M.T., Baur, C., Navab, N., and Albarqouni, S. (2019b). Staingan: Stain Style Transfer for Digital Histological Images. In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 953–956.

Silva-Rodríguez, J., Colomer, A., and Naranjo, V. (2021). WeGleNet: A weakly-supervised convolutional neural network for the semantic segmentation of Gleason grades in prostate histology images. Comput. Med. Imaging Graph. *88*, 101846.

Simonyan, K., and Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition.

Singh, M.P., Rai, S., Pandey, A., Singh, N.K., and Srivastava, S. (2021). Molecular subtypes of colorectal cancer:

An emerging therapeutic opportunity for personalized medicine. Genes Dis *8*, 133–145.

Sohail, A., Khan, A., Wahab, N., Zameer, A., and Khan, S. (2021). A multi-phase deep CNN based mitosis detection framework for breast cancer histopathological images. Sci. Rep. *11*, 6215.

Steiner, D.F., Nagpal, K., Sayres, R., Foote, D.J., Wedin, B.D., Pearce, A., Cai, C.J., Winter, S.R., Symonds, M., Yatziv, L., et al. (2020). Evaluation of the Use of Combined Artificial Intelligence and Pathologist Assessment to Review and Grade Prostate Biopsies. JAMA Netw Open *3*, e2023267.

Stodden, V., McNutt, M., Bailey, D.H., Deelman, E., Gil, Y., Hanson, B., Heroux, M.A., Ioannidis, J.P.A., and Taufer, M. (2016). Enhancing reproducibility for computational methods. Science *354*, 1240–1241.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826.

Tosun, A.B., Pullara, F., Becich, M.J., Taylor, D.L., Fine, J.L., and Chennubhotla, S.C. (2020). Explainable AI (xAI) for Anatomic Pathology. Adv. Anat. Pathol. *27*, 241–250.

Truong, A.H., Sharmanska, V., Limbäck-Stanic, C., and Grech-Sollars, M. (2020). Optimization of deep learning methods for visualization of tumor heterogeneity and brain tumor grading through digital pathology. Neurooncol Adv *2*, vdaa110.

Tschuchnig, M.E., Oostingh, G.J., and Gadermayr, M. (2020). Generative Adversarial Networks in Digital Pathology: A Survey on Trends and Future Potential. Patterns (N Y) *1*, 100089.

Tsou, P., and Wu, C.-J. (2019). Mapping Driver Mutations to Histopathological Subtypes in Papillary Thyroid Carcinoma: Applying a Deep Convolutional Neural Network. J. Clin. Med. Res. *8*.

Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., Baust, M., Steiger, K., Schlitter, A.M., Esposito, I., and Navab, N. (2016). Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images. IEEE Trans. Med. Imaging *35*, 1962–1971.

Vellal, A.D., Sirinukunwattan, K., Kensler, K.H., Baker, G.M., Stancu, A.L., Pyle, M.E., Collins, L.C., Schnitt, S.J., Connolly, J.L., Veta, M., et al. (2021). Deep Learning Image Analysis of Benign Breast Disease to Identify Subsequent Risk of Breast Cancer. JNCI Cancer Spectr *5*, kaa119.

Veta, M., Heng, Y.J., Stathonikos, N., Bejnordi, B.E., Beca, F., Wollmann, T., Rohr, K., Shah, M.A., Wang, D., Rousson, M., et al. (2019). Predicting breast tumor proliferation from whole-slide images: The TUPAC16 challenge. Med. Image Anal. *54*, 111–121.

Wagner, S.J., Khalili, N., Sharma, R., Boxberg, M., Marr, C., de Back, W., and Peng, T. (2021). Structure-Preserving Multi-Domain Stain Color Augmentation using Style-Transfer with Disentangled Representations.

Wang, S., Wang, T., Yang, L., Yang, D.M., Fujimoto, J., Yi, F., Luo, X., Yang, Y., Yao, B., Lin, S., et al. (2019). ConvPath: A software tool for lung adenocarcinoma digital pathological image analysis aided by a convolutional neural network. EBioMedicine *50*, 103–110.

Wang, S., Rong, R., Yang, D.M., Fujimoto, J., Yan, S., Cai, L., Yang, L., Luo, D., Behrens, C., Parra, E.R., et al. (2020). Computational Staining of Pathology Images to Study the Tumor Microenvironment in Lung Cancer. Cancer Res. *80*, 2056–2066.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data *3*, 160018.

Woerl, A.-C., Eckstein, M., Geiger, J., Wagner, D.C., Daher, T., Stenzel, P., Fernandez, A., Hartmann, A., Wand, M., Roth, W., et al. (2020). Deep Learning Predicts Molecular Subtype of Muscle-invasive Bladder Cancer from Conventional Histopathological Slides. Eur. Urol. *78*, 256–264.

Wollmann, T., and Rohr, K. (2021). Deep Consensus Network: Aggregating predictions to improve object detection in microscopy images. Med. Image Anal. *70*, 102019.

Yamashita, R., Long, J., Banda, S., Shen, J., and Rubin, D.L. (2021). Learning domain-agnostic visual representation for computational pathology using medically-irrelevant style transfer augmentation. IEEE Trans. Med. Imaging *PP*.

Yu, K.-H., Wang, F., Berry, G.J., Ré, C., Altman, R.B., Snyder, M., and Kohane, I.S. (2020). Classifying non-small

cell lung cancer types and transcriptomic subtypes using convolutional neural networks. J. Am. Med. Inform. Assoc. *27*, 757–769.

Zhang, F., Zhong, L.-Z., Zhao, X., Dong, D., Yao, J.-J., Wang, S.-Y., Liu, Y., Zhu, D., Wang, Y., Wang, G.-J., et al. (2020a). 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. *12*, 1758835920971416.

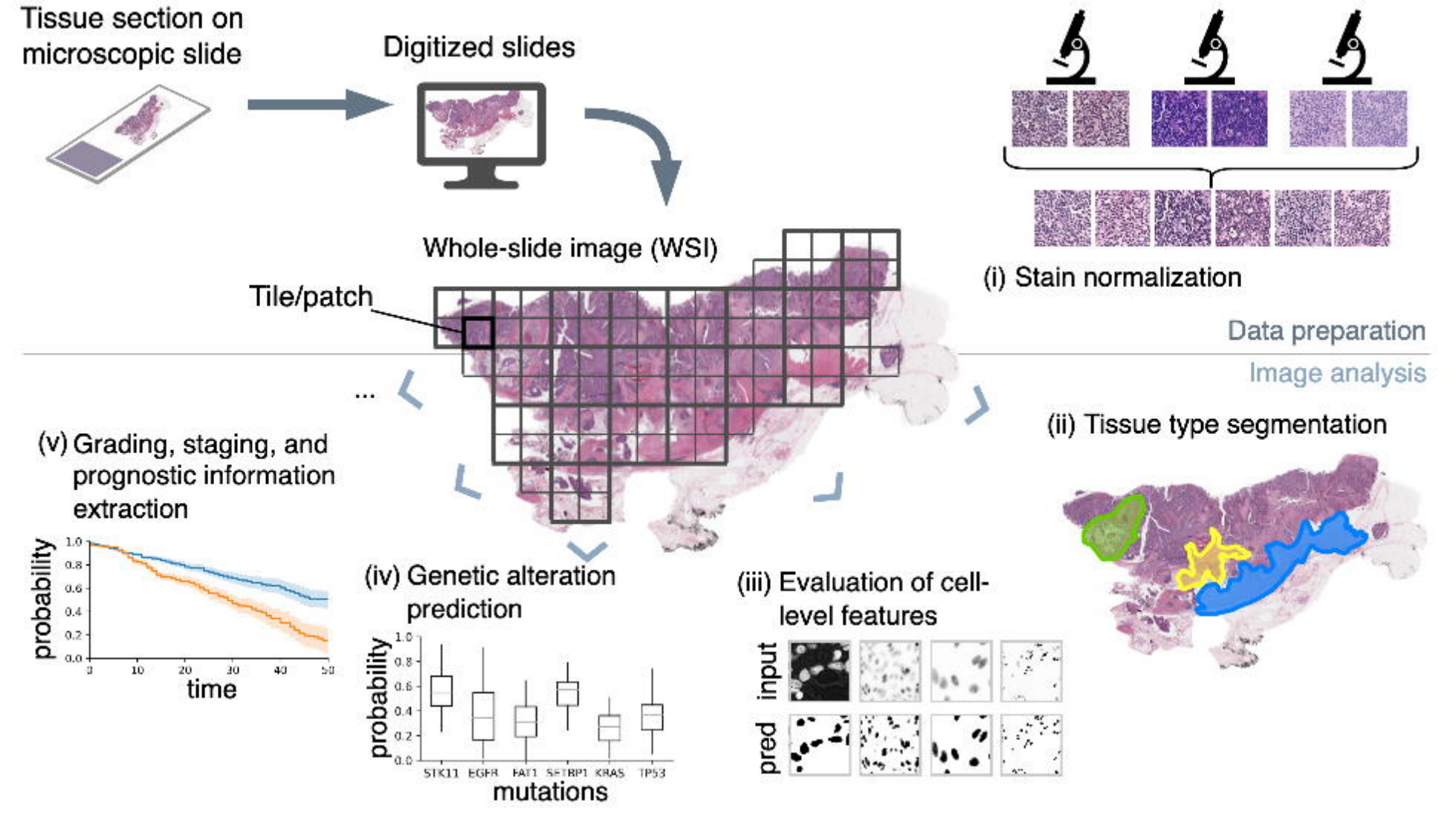
Zhang, H., Kalirai, H., Acha-Sagredo, A., Yang, X., Zheng, Y., and Coupland, S.E. (2020b). Piloting a Deep Learning Model for Predicting Nuclear BAP1 Immunohistochemical Expression of Uveal Melanoma from Hematoxylin-and-Eosin Sections. Transl. Vis. Sci. Technol. *9*, 50.

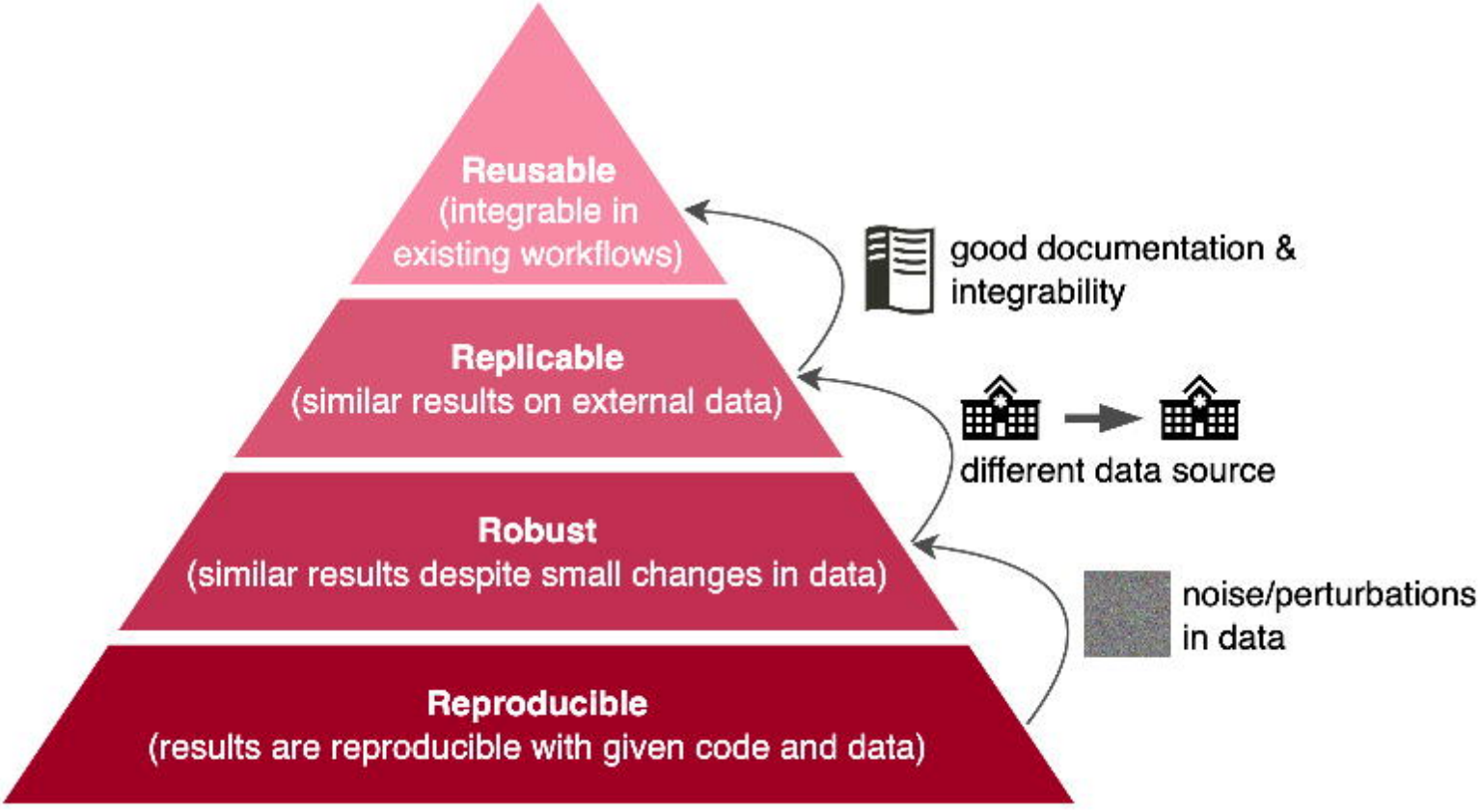
Zhao, B., Chen, X., Li, Z., Yu, Z., Yao, S., Yan, L., Wang, Y., Liu, Z., Liang, C., and Han, C. (2020a). Triple U-net: Hematoxylin-aware nuclei segmentation with progressive dense feature aggregation. Med. Image Anal. *65*, 101786.

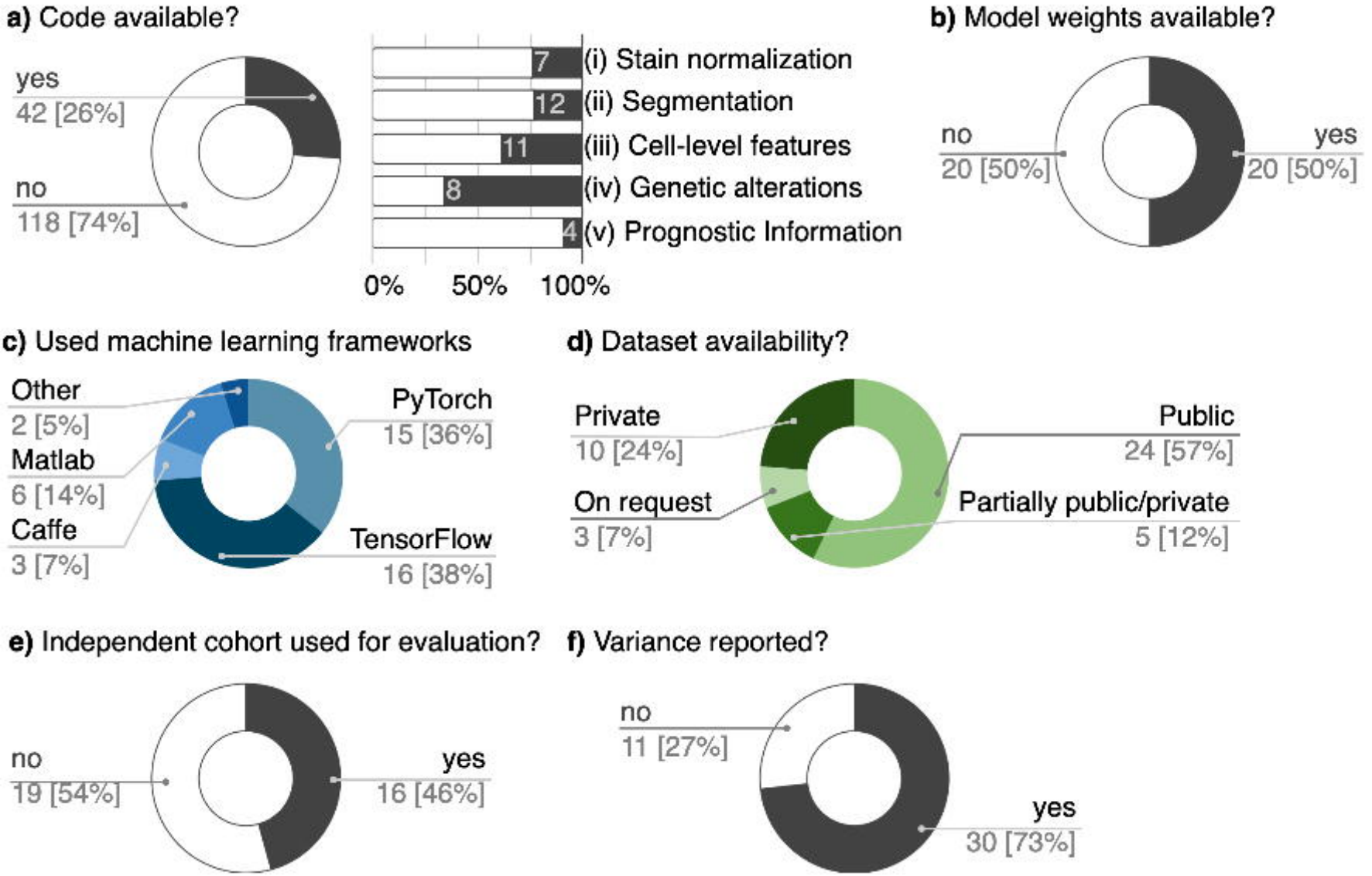
Zhao, K., Li, Z., Yao, S., Wang, Y., Wu, X., Xu, Z., Wu, L., Huang, Y., Liang, C., and Liu, Z. (2020b). Artificial intelligence quantified tumour-stroma ratio is an independent predictor for overall survival in resectable colorectal cancer. EBioMedicine *61*, 103054.

Zheng, Y., Jiang, Z., Zhang, H., Xie, F., Shi, J., and Xue, C. (2019). Adaptive color deconvolution for histological WSI normalization. Comput. Methods Programs Biomed. *170*, 107–120.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929.


Tissue section on microscopic slide
Digitized slides
Whole-slide image (WSI)
Tile/patch
...
(i) Stain normalization
Data preparation
Image analysis
(ii) Tissue type segmentation
(iii) Evaluation of cell-level features
input
pred
(iv) Genetic alteration prediction
probability
STK11
EGFR
FAT1
SFTBP1
KRAS
TP53
mutations
(v) Grading, staging, and prognostic information extraction
probability
time


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


a) Code available?
yes
42 [26%]
no
118 [74%]
7
12
11
8
4
(i) Stain normalization
(ii) Segmentation
(iii) Cell-level features
(iv) Genetic alterations
(v) Prognostic Information
0%
50%
100%
b) Model weights available?
no
20 [50%]
yes
20 [50%]
c) Used machine learning frameworks
Other
2 [5%]
Matlab
6 [14%]
Caffe
3 [7%]
PyTorch
15 [36%]
TensorFlow
16 [38%]
d) Dataset availability?
Private
10 [24%]
On request
3 [7%]
Public
24 [57%]
Partially public/private
5 [12%]
e) Independent cohort used for evaluation?
no
19 [54%]
yes
16 [46%]
f) Variance reported?
no
11 [27%]
yes
30 [73%]