Review Article

Deep Learning Approaches and Applications in Toxicologic Histopathology: Current Status and Future Perspectives

Shima Mehrvar[1], Lauren E. Himmel[1], Pradeep Babburi[2], Andrew L. Goldberg[2], Magali Guffroy[1], Kyathanahalli Janardhan[1], Amanda L. Krempley[1], Bhupinder Bawa[1]

[1]Preclinical Safety, AbbVie Inc., North Chicago, IL, USA, [2]Business Technology Solutions, AbbVie Inc., North Chicago, IL, USA

Submitted: 26-May-2021 Accepted: 18-Jul-2021 Published: 01-Nov-2021

Abstract

Whole slide imaging enables the use of a wide array of digital image analysis tools that are revolutionizing pathology. Recent advances in digital pathology and deep convolutional neural networks have created an enormous opportunity to improve workflow efficiency, provide more quantitative, objective, and consistent assessments of pathology datasets, and develop decision support systems. Such innovations are already making their way into clinical practice. However, the progress of machine learning - in particular, deep learning (DL) - has been rather slower in nonclinical toxicology studies. Histopathology data from toxicology studies are critical during the drug development process that is required by regulatory bodies to assess drug-related toxicity in laboratory animals and its impact on human safety in clinical trials. Due to the high volume of slides routinely evaluated, low-throughput, or narrowly performing DL methods that may work well in small-scale diagnostic studies or for the identification of a single abnormality are tedious and impractical for toxicologic pathology. Furthermore, regulatory requirements around good laboratory practice are a major hurdle for the adoption of DL in toxicologic pathology. This paper reviews the major DL concepts, emerging applications, and examples of DL in toxicologic pathology image analysis. We end with a discussion of specific challenges and directions for future research.

Keywords: Deep learning, digital image analysis, histopathology, machine learning, preclinical safety, toxicologic pathology, whole slide imaging

Introduction

In nonclinical drug safety assessment, large numbers of slides are generated from healthy laboratory animals exposed to the test item in order to detect toxicity in various tissues mandated to be part of the animal studies by regulatory agencies. Toxicologic pathologists experience the burden of evaluating all of these slides. The majority of the tissues in a routine toxicity study are expected to be within normal limits, while the abnormal tissues could potentially contain any number of morphological changes (e.g., necrosis, hyperplasia, inflammatory infiltrate). Given this scenario, using an artificial intelligence (AI) system for pathology evaluations could save a substantial amount of time in nonclinical toxicity studies leading to accelerated discovery and development of safer drugs that can make a real difference in patients' lives.[1]

For pathologists, the fundamental building block of histologic assessment is the cell. When viewing tissue on a histology slide, the pathologist, based on the years of rigorous training and experience, perceives and interprets cellular distribution, arrangement, and morphology and rapidly classifies cells into types and tissue architectures into processes to arrive at a working diagnosis. This working diagnosis is often compared against literature expert opinions to arrive at a consensus diagnosis or the so-called "ground truth." Thus, it is natural

Access this article online

Quick Response Code:

Website: www.jpathinformatics.org

DOI: 10.4103/jpi.jpi_36_21

Address for correspondence: Dr. Shima Mehrvar, Preclinical Safety, AbbVie Inc., 1 N Waukegan Rd, North Chicago, IL 60064, USA.
E-mail: shima.mehrvar@abbvie.com
Dr. Bhupinder Bawa
Preclinical Safety, AbbVie Inc., 1 N Waukegan Rd, North Chicago, IL 60064, USA.
E-mail: bhupinder.bawa@abbvie.com

How to cite this article: Mehrvar S, Himmel LE, Babburi P, Goldberg AL, Guffroy M, Janardhan K, *et al*. Deep learning approaches and applications in toxicologic histopathology: Current status and future perspectives. J Pathol Inform 2021;12:42.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2021/12/1/42/329733

for a pathologist to anthropomorphize an AI system and assume that it "sees" cells and tissues in the same way that a human does to arrive at a segmentation or classification result, perhaps even with some level of "memory" of what it has seen in other studies. However, there are some key differences in the way AI systems function. A computer operates in bits (0s and 1s) and fundamentally "sees" an image in pixels. Pixels are the smallest addressable picture elements on a computer screen. Looking at individual pixels does not provide meaningful information. However, with advancements in mathematical algorithms and computational power, machines can be trained to look at the features in an image at a higher level of abstraction like humans do. For example, pixel-wise measurements of shape and texture can be used by AI systems to "learn" features like nuclei. Furthermore, a computer learns from whatever features exist in its training examples. For instance, if a computer is not trained on cytomorphological variants and does not have the opportunity to "see" biological and preanalytical variability, it will misclassify the cells and/or processes present in such images. Despite exponential progress in developing state-of-the-art AI algorithms in recent years, the current paradigm of task-oriented AI is generally considered narrow and is not yet at a level that is on par with human intelligence.[2]

In pathology, deep learning (DL)-based methods to date have been largely developed for tumor identification and biomarker-based characterization in human biopsies. In a meta-review of over 130 articles published between 2013 and 2019 using various DL methods, nearly all were based on human cancer.[3] These applications are highly useful and relevant in the practice of clinical oncology or discovery biology;[4] however, these methods do not address the unique circumstances of toxicologic pathology, and to date, the field is largely bereft of such methodologies. Although there are detailed and thorough reviews, commentaries, opinions, and surveys on different aspects of digital pathology (DP) with most covering clinical diagnostic applications of DL,[2,3,5-23] this is the first in-depth review of DL methods for toxicologic histopathology. This review provides (a) basic information on the discipline of toxicologic pathology, (b) a brief description of the DL methods in DP with the focus on implementation in toxicologic pathology, (c) details on the recent applications of DL-based image analysis in toxicologic pathology, and (d) the challenges facing the adoption of a digital workflow leveraging DL.

Role of toxicologic pathology in drug discovery and development

Drug discovery and development is a complex process [Figure 1] that starts from target identification and validation of a chemical entity or compound, and the process could take an average of 10–12 years from bench to bedside.[24] The purpose of this approach is to bring safe and effective drugs to the market after rigorous testing in nonclinical and clinical areas. A toxicologic pathologist works in a highly matrixed, multidisciplinary environment supporting the drug development pipeline from early discovery until regulatory approvals, and in some instances, to postmarketing research.[24]

Toxicologic pathology is the study of the molecular, cellular, tissue, organ, and organism-level response to novel agents. Toxicologic pathologists play a critical role in the characterization of potential unintended effects of new drugs to support human trials. As a discipline, toxicologic pathology traditionally and most commonly uses hematoxylin and eosin (H & E)-stained tissue sections on glass slides to microscopically examine the effects of treatment (e.g., drugs, devices, or chemicals) on tissues in laboratory animals.

A toxicologic pathology study has unique challenges compared to clinical diagnostic pathology (CDP). The volume of slides created for each study is large, and studies typically involve multiple animals. A standard investigational new drug (IND)-enabling rodent study has approximately 60 tissues for each of 80–100 animals.[25] Microscopic evaluation of thousands of slides per study is labor-intensive and time-consuming. Another unique challenge with toxicology studies is that the majority of tissues are normal, which means they do not display any test item-related findings. Adding time constraints to slide review/interpretation is detrimental to performance in clinical practice,[26] and toxicologic pathologists face similar time pressure under project timelines. Furthermore, years of experience and expertise are needed to differentiate normal background lesions–those developing spontaneously in laboratory animals due to age, sex, diet, or strain–from drug-induced abnormalities.[27]

AI has found its way to many industries, although pharmaceutical research and development is still in an early phase of integrating AI into their workflow.[28,29] AI-based toxicity predictions have been applied in the field of molecular toxicology,[30-33] but they have not yet emerged as a widely used tool in the discipline of toxicologic histopathology.

Digitization of glass slides

To adopt AI in toxicologic pathology, digitization of glass slides is the essential first step. Advances in whole slide images (WSI) and significantly improved DP systems have paved the path toward a fully digitalized pathology workflow.[19] WSI-based diagnoses have proven concordant with traditional glass slide-based diagnoses in both human and veterinary pathology.[34] The digitalized pathology workflow has several advantages: (a) easy and efficient archival of WSI, (b) traceability with effortless and rapid retrieval of cases compared to glass slides, (c) accessibility and the possibility to get opinions from experts across the globe in a timely fashion, for peer review or diagnostic concordance, and (d) comparability of multiple images on the same screen, which can avoid diagnostic drift or resolve subtle differences between animals. It is notable that in the face of the global COVID-19 pandemic in 2020, minimally adequate peripheral instrumentation to conduct remote digital slide reviews has accelerated adoption with surprisingly favorable results.[35]

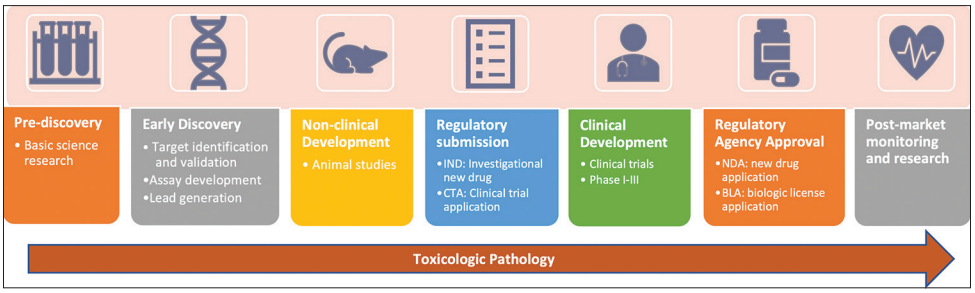

Figure 1: Drug discovery and development process. This flowchart is a simplified version of the pipeline, and there are overlaps and close collaborations between different steps.

These illustrate only the most superficial benefits of a DP workflow even without any computational image analysis.

Overview Of Deep Learning Approaches

For more information on basic terms used in AI, we refer to publications detailing the subject.[6,16,36] Before introducing some key approaches in DL, it is important to understand the terms "architecture," model," and "algorithm." and their differences. An architecture describes a general approach to a DL task and the parameterization of that approach (e.g., the number and size of different layers or the type of each layer). For example, an artificial neural network is a type of architecture [Figure 2a]. A model is one specific instance of a given architecture or modular instances of multiple architectures trained on a given dataset [Figure 2b and c]. An algorithm is a set of rules to follow to implement architectures [Figure 2b and c].

Early applications of AI in histopathology, dating back to 1966, involved extracting quantitative measures from microscopy images.[37] Some early works in cytology and histopathology were before the invention and widespread use of whole slide scanners, and the image analysis was performed on a small field of view captured by conventional microscopes.[38] Prior to the success of DL models, traditional AI/ML models such as decision trees and thresholding had already been used in DP. However, these conventional models usually rely on the data representations and manually selected (hand-crafted) features, which provide high contrast between the desired image classes. Hand-crafted features are preprocessed image representations such as color, texture, or shape that are used as the input of conventional machine learning models. In contrast, DL methods can learn features directly from raw images (i.e., no hand-crafted features required), yielding higher performance and generalizability. The DL models have consistently outperformed earlier ML techniques in various fields such as image classification.[39] In pathology, DL-based approaches usually have four main steps: (1) database creation, (2) preprocessing, (3) model selection and training, and (4) postprocessing and evaluation. The following subsections explain each step of this process and some considerations in the context of DL in toxicologic pathology.

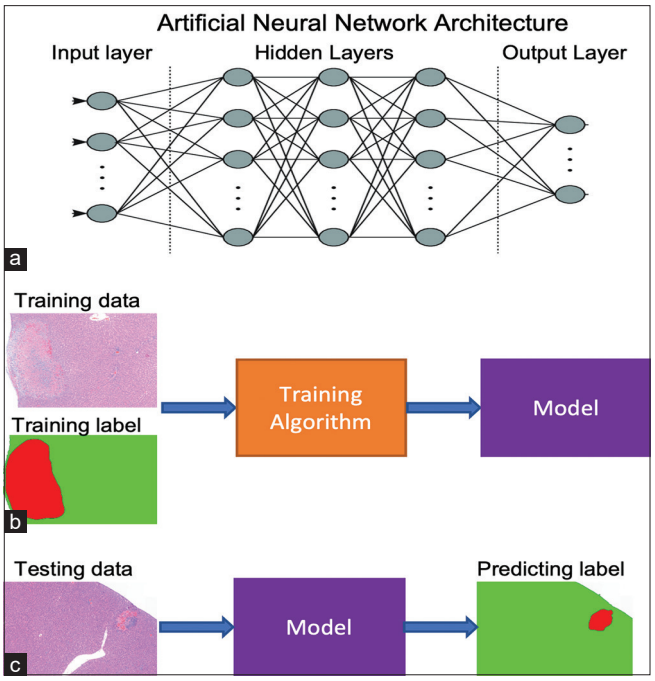

Figure 2: (a) An artificial neural network architecture contains input layer, hidden layers, and output layer of neurons. (b) The architecture can be trained by a training algorithm mapping the data to label. (c) The trained model can then be used for inference.

Database creation

A DL study starts with creating a diverse collection of image data stored electronically in a computer system. Although there are multiple publicly available pathology databases in human medicine, only a few exist for nonclinical studies [Table 1]. Notably, these nonclinical databases are only WSI repositories and do not come with ground truth annotations reviewed by pathologists, and they are not aimed at DP. Although collecting WSI repositories without ground truth annotations is relatively easy, for most AI/DL applications, these databases would need to come with pathologist annotations. Efforts like the Innovative Medicines Initiative (BIGPICTURE project) to develop a central repository of digital slides from humans and nonclinical species to create AI tools will likely accelerate these efforts in toxicologic pathology in the near future.[40]

Table 1: Available whole slide image repositories of nonclinical digital pathology slides and some of the largest H&E-stained repositories for clinical applications

Repository	Clinical/nonclinical	Location (link)
TG-GATE	Nonclinical	http://toxico.nibio.go.jp/english/index.html
CEBS	Nonclinical	https://connect.niehs.nih.gov/cebs3/ui/
VMD	Clinical and nonclinical	http://www.virtualmicroscopydatabase.org/
TCGA	Clinical	https://portal.gdc.cancer.gov/repository
GTEx	Clinical	https://www.gtexportal.org/home/datasets
TMAD	Clinical	https://tma.im/cgi-bin/home.pl
KIMIA	Clinical	https://kimialab.uwaterloo.ca/kimia/index.php/data-and-code-2/
Camelyon	Clinical	https://camelyon17.grand-challenge.org/Data/
TUPAC	Clinical	http://tupac.tue-image.nl/node/3

Preprocessing

The giga-pixel size of WSIs is too large to be processed by present-day computers using any DL model. Therefore, prior to training any model, some preprocessing steps are essential, such as data curation, tile extraction, selection, and splitting the data into training and test sets. A WSI is stored on a computer as a pyramid of tiled images, and these tiles can be extracted into smaller, easily manageable patches via grid sampling [Figure 3]. Typically, patches are extracted at the desired magnification [Figure 3; and more details in Section "image format, size, and magnification"] and can then be utilized either for training or inference (run predictions on the DL model for validation and test).

Splitting database into training and test sets – Considerations

For training a DL model, the data should be split into training, validation, and test sets. The biggest proportion of the data is the training set, which is used to train the models. The validation set is typically preserved for fine-tuning the models or selecting the best model. The performance of the final model is then evaluated using images that are not part of either training or validation, which is referred to as the test set. Ideally, these data sets should not have any overlap, but it must be noted that they should be selected from the same distribution that is representative of prospective data that the model will be used for. As an example, if the initial dataset contains drug-treated and control groups with various normal/abnormal characteristics, the images from all groups should be present in the training, validation, and test sets to avoid suboptimal performance on future data where the model will actually be applied in practice.

Deep learning methods

Supervised learning

Supervised learning, as the name suggests, is a type of machine learning model that requires supervision during training. In the case of image data, the supervision is provided in the form of hand-labeled annotations by subject matter experts (SMEs). The goal of supervised learning is to develop an appropriate mathematical model that can map images to their corresponding labels provided by pathologists. Figure 4 illustrates different DL tasks in supervised learning with an example of normal/abnormal histology of a liver WSI. In supervised learning architectures, the DL task dictates the granularity of the labels in three levels: (a) patch-level labels in classification tasks, (b) object-level in object detection tasks, and (c) pixel-level associations to different morphologies in segmentation tasks. For a classification task, an image is classified as a binary or multi-class outcome, such as predicting a positive/negative class or the presence/absence of a region of interest (ROI). Multiple convolutional neural network (CNN) architectures have been proposed to classify patches in WSIs (e.g. AlexNet for brain tumors[41] and deep convolutional features for colorectal adenocarcinoma classification[42]).Tile-based classification models are also proposed to consider the context of neighboring patches in WSIs (e.g., the use of an architecture with a combination of recurrent neural networks and CNN for bladder cancer diagnosis[43]). However, a simple classification may not always be helpful, and one may require more information, such as the type and location of the ROI in the image. Object detection is a computer vision technique that allows for the identification and localization of objects in an image or video.

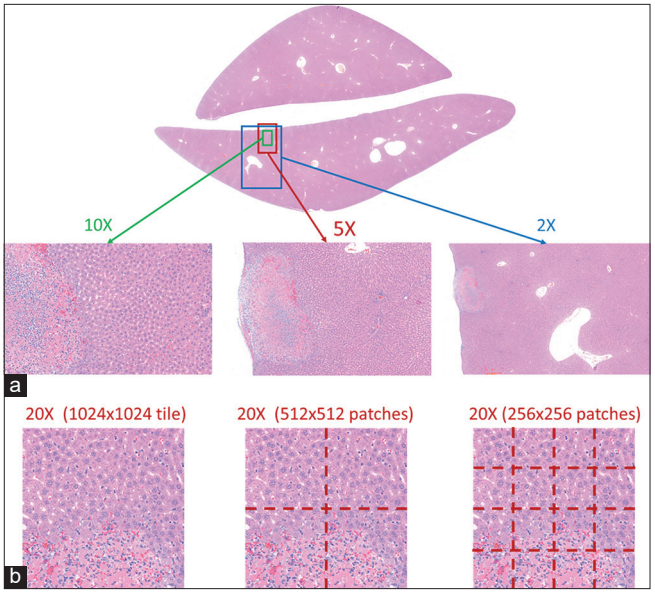

Figure 3: Different (a) magnification level and (b) patch pixel-size in a liver whole slide image

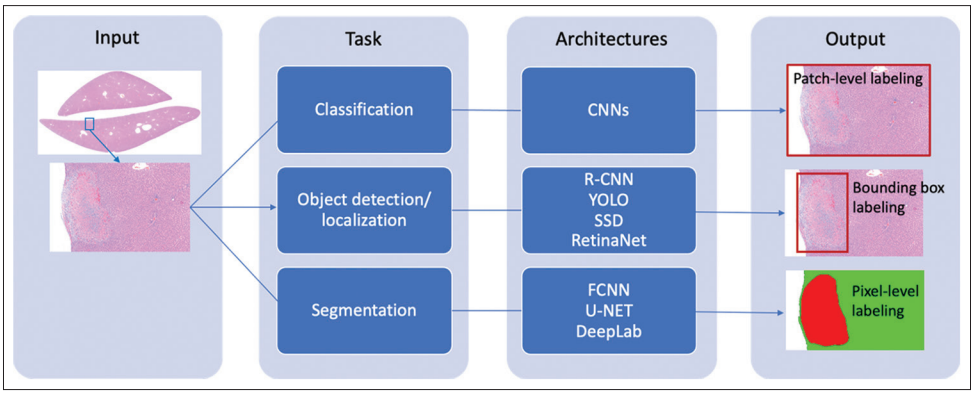

Figure 4: An overview of supervised deep learning tasks that can be applied on whole slide images. A hypothetical example of a liver is used to illustrate the output of different tasks (green: normal; red: abnormal). R-CNN: Region-based CNN, YOLO: You Only Look Once, SSD (Single-Shot Detector), FCNN (Fully Connected Neural Networks)

Object location is usually shown by drawing a bounding box around the target object or ROI. In DP, object detection has been most commonly utilized in cell/nuclei detection tasks.[15] Examples of DL-based object detection architectures include the region-based convolutional neural network (R-CNN) family, you only look once (YOLO)[44] single-shot detectors (SSD),[45] and RetinaNet.[46] Finally, image segmentation takes this even a step further by classifying each pixel in the ROI that essentially recognizes not only the instance and spatial location of an object or ROI but also its precise shape distinct from the background. Segmentation architectures that are commonly used on WSIs include fully CNNs,[47] U-NET,[48] and DeepLab.[49,50]

Unsupervised Learning

Unsupervised learning aims at finding a mathematical model that can describe the patterns of an unlabeled image. Although unsupervised learning in DP is at a very early stage of development, this learning methodology presents an untapped potential ready to be explored by computational scientists.[19] Unsupervised learning approaches do not require mapping an input image to a predefined annotated label by an expert. Instead, the image is transformed into lower-dimensional representations which are also referred to as latent features (a simpler representation of data for analysis), and these features can be analyzed based on various techniques such as clustering to draw a meaningful conclusion (e.g., the use of convolutional adversarial autoencoders, an unsupervised architecture for prostate cancer detection[51]). In addition, the latent features of a DL model can be further reduced to a 2D map, and the neighboring data points can be grouped to provide an informative visualization of heterogeneity in the data. UMAP (e.g., for normal histology)[52] and t-SNE (e.g., for showing stain variation)[53] are examples that use dimensionality reduction techniques to effectively show how a model identifies various morphological features in the data.

Given a slide containing tissue with normal and abnormal regions, using unsupervised learning without any ground truth annotations, the latent features can be extracted and generate clusters that are unique to each tissue region – normal and abnormal. Similar work has been done on brain tissue that contains lesional (e.g., lymphomas) and nonlesional (e.g., normal cortical gray and white matter) categories where the unsupervised model generates multiple clusters representing different areas.[54] Furthermore, these lower-dimensional features in unsupervised learning can be used to artificially generate new images (e.g., stain normalization[55] and stain transfer). Figure 5 shows how these two different tasks use unsupervised architectures. Generative adversarial networks (GANs)[55-57] and autoencoders[54,57-59] are examples of unsupervised architectures. In classifying esophageal cancer WSIs, the classifiers have the best performance using unsupervised extracted features compared to supervised and weakly supervised approaches.[60] Cycle-consistent GAN architecture has been utilized for the conversion of H & E slides to a virtual trichrome stain to aid in the staging of nonalcoholic steatohepatitis.[61] This type of technology could reduce the temporal and financial burdens of block retrieval, re-cutting, and shipping. Since an expert does not have to manually annotate/label the tissues, unsupervised learning might have a faster turn-around time and can be extremely useful in toxicologic pathology.

Weakly supervised learning

Weakly supervised learning is a derivative of supervised and unsupervised learning approaches, which does not require intensive labeling as normally required for supervised learning and yet can infer slide-level labels to predict local detections. The goal is to train a model using minimal (weak) annotations (slide-level label) to predict the finer (pixel/patch-level) labels [Figure 6]. In practice, these slide-level labels are often available as metadata, and labor-intensive pixel-wise labels are not required. Therefore, weakly supervised learning is an ideal approach for tasks such as slide triaging that does not require detailed ground truth information. There are multiple architectures based on weakly supervised learning, such as multiple instance learning (MIL) and weakly supervised attention-based models.[3] In toxicologic pathology, a WSI can be labeled as abnormal if any part of the image contains a lesion and these slide-level labels are available as tabular data within

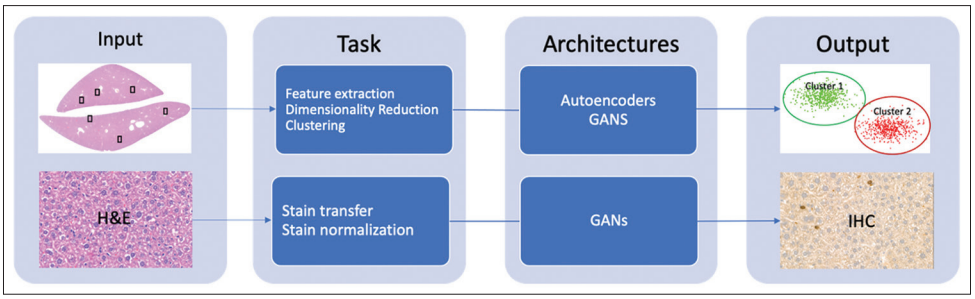

Figure 5: An overview of unsupervised deep learning tasks that can be applied on whole slide images. In the output of clustering, each dot represents the feature of its corresponding tile (green: normal; red: abnormal). GANs: Generative Adversarial Neural Networks.

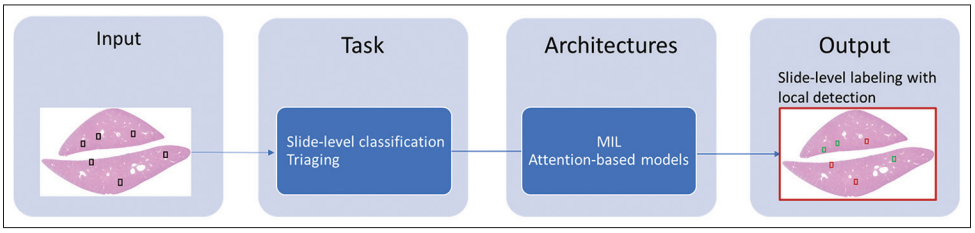

Figure 6: An overview of weakly supervised deep learning tasks that can be applied on whole slide images. In this example, the slide level label assigned with local detection of abnormal patches (green: normal; red: abnormal). MIL: Multiple Instance Learning.

the pathology report. Therefore, weakly supervised learning has a high potential for abnormality triaging in toxicologic pathology (this application is discussed thoroughly in Section "Computer-assisted abnormality detection").

Postprocessing and evaluations

Trained DL models need to undergo a rigorous performance evaluation before putting them to practical use, especially in the case of decision support systems. DL model validation should include a comparison of the results from the DL model with the ground truth (manually annotated test data set and/or diagnosis made by a pathologist).[62,63] Proprietary and inadequately documented methods in published literature create barriers to determining the reproducibility of these results. It is important to be aware that manually annotated ground truth might suffer from user-induced noise and bias due to inter- and intra-observer variance, the experience and familiarity of a pathologist with the annotation software, and visual and cognitive traps that commonly impact the interpretation of histology images.[64] Therefore, the ground truth (consensus diagnosis) is best determined by the input of at least two experienced pathologists or preferably three in case of disagreement.[26,63]

Model performance assessments can also help discover any discrepancies in the data or its distribution among training, validation, and test sets. For example, an increased false-negative rate has been reported in treated prostate cancers,[65] raising the consideration that additional data from treated cancers need to be included in the training set for AI to reliably discern the cytologic features that separate tumor from benign tissue. Another point worth noting is that DL training could extend in perpetuity with diminishing returns. Hence, it is important to evaluate the performance during training at regular intervals on the validation set and stop training when there is no measurable improvement in performance. In a proof-of-concept study of a DL CNN for use in human inflammatory gastric biopsies, the algorithm stopped learning once an error rate of <0.01 was achieved.[66]

Understanding the evaluation methods is important for pathologists and DL developers so that both parties can objectively assess the performance of the models qualitatively and quantitatively. While a subjective visual evaluation of the performance of the algorithm by a pathologist may be acceptable in some applications.[67] A quantitative measure of the model performance may be more appropriate for an objective assessment. For a classification task, metrics such as sensitivity, specificity, and accuracy would be beneficial. For a segmentation task, dice coefficient (F1 score = 2*area of overlap/(area of prediction + area of ground truth) or intersection over union (IoU = area of overlap/area of the union) calculate pixel-level accuracy are frequently used.[68] In the case of object detection (e.g., nuclei detection), where the boundary of the object (pixel-level) may not be as important as the number of true positive objects, the F1 score is calculated per object, which is then considered as true positive when it is higher than a threshold.[69] Aggregated Jaccard Index is proposed as an evaluation metric to take into account both the pixel and object level evaluations.[70] For a multi-class detection model, a confusion matrix table demonstrating a per-class score metric can be deduced to rank the performance of each class.

Machine Learning Applications In Toxicologic Pathology

In this section, various applications of machine/DL in toxicological pathology are presented, as well as a vision for

how DL can elevate pathology by providing computer-assisted evaluations that uncover previously unrecognized features in histopathology images. While DL-based computational pathology is showing great promise to be beneficial in clinical practice, its implementation in toxicologic pathology practice is still in the early stages. Figure 7 illustrates a conceptual AI-integrated digital system within the routine toxicologic pathology workflow. The process starts with the generation of WSIs followed by computer-assisted quality control (QC) (more details in Section "Computer-assisted quality control") that checks each slide and flags those that are of poor quality. The next key step in the pipeline is to adopt a flexible and user-friendly image management system where the slides are stored and can be readily accessed. The system needs to be flexible, so it can integrate well with other downstream and upstream applications, and it needs to be user-friendly, so that it provides a quality viewing experience the pathologists are willing to adopt as their method for primary reading in routine practice. Another vital point to be noted is that with the increase in demand for people wanting to work from virtually anywhere, embracing a system that can use cloud-based solutions will improve productivity and is also beneficial for remote collaboration. With a flexible image management system in place, deploying AI solutions for various use cases [Figure 7] will be faster and more efficient for processing large volumes of data and visualizing the end results.

Recently, many AI-assisted analyses in nonclinical research and toxicologic pathology have been published [Table 2]. Some these studies are the result of collaboration between industry-based researchers and external vendors such as Deciphex, Visiopharm, AIRA matrix, Aiforia, or HALO to provide DL-based solutions for different pathology tasks. The following tabulated information suggests that these nonclinical applications are on supervised DL (with the exception of Freyre *et al*.)[71] in specific species/tissue/abnormality, and attention to DL applications in toxicologic pathology has increased over the past year. The following subsections detail four different categories of DL applications in toxicological pathology: Computer-assisted QC, research-driven computational image analysis, computer-assisted abnormality detection, and content-based image retrieval. In some cases, where clinical medicine is further progressed, select H & E-based DL applications from CDP are also provided to illustrate its successes and draw parallels between DL in CDP and toxicologic pathology.

Computer-assisted quality control

Manual QC of the scanned digital slides typically involves multiple steps: (a) reviewing the slide metadata for accuracy and discrepancies, (b) comparing the physical tissue on the glass slide and digital image for scan quality, and (c) evaluation of the tissue to identify digital artifacts such as stitching or out-of-focus areas. Due to the high quality of slides produced in nonclinical pathology labs, standard practices in laboratories involve reviewing a predetermined proportion of slides, such as 10% of the total slides per animal. Any slides determined to be unacceptable are rescanned and rechecked for quality. Manual QC of WSI is a time-consuming and laborious task that is prone to fatigue, human error, and inefficiencies in a laboratory setting where large volumes of slides are scanned on a daily basis. The use of manual QC is common in identifying slides that need to be re-scanned due to missing tissue on the digitized slide or out-of-focus regions.

The National Society for Histotechnology and College of American Pathologists have initiated a program known as (HistoQIP) Whole Slide Image Quality Improvement Program) to improve the quality of WSI.[97,98] A computer-assisted QC workflow could significantly improve this process by rapidly reviewing the WSIs as they are scanned and flagging those that need human intervention [Figure 8]. To this end, automated tools are being developed for computerized in-line digital QC. HistoQC[99] is an open-source tool that assesses the heterogeneity of WSI datasets and identifies artifacts present on glass or digital slides; it is being evaluated in NEPTUNE, CureGN, and Kidney Precision Medicine Project.[98] Artifacts induce unwanted noise in AI-trained systems that potentially generate suboptimal results. For

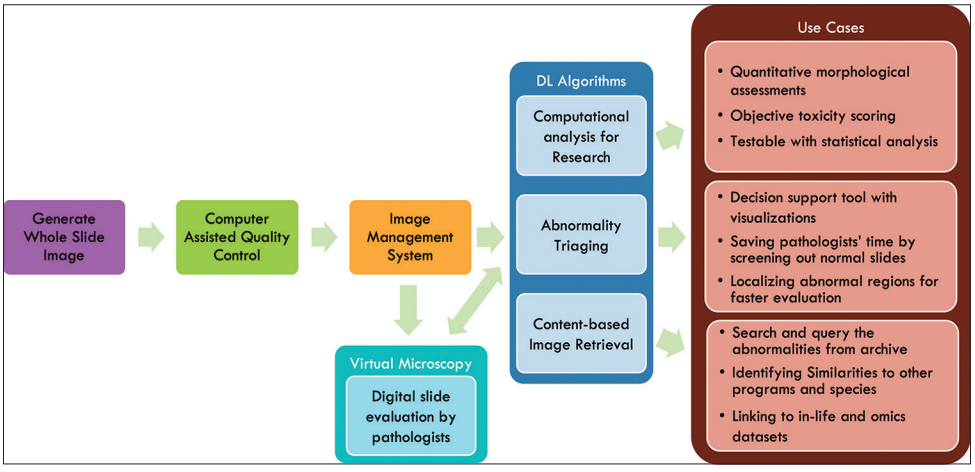

Figure 7: Artificial intelligence-integrated digital workflow in toxicologic pathology

Table 2: Overview of deep learning-based applications in nonclinical histopathology

Reference	Species	Tissue	Application	Method	Dataset
Nonclinical basic science research					
Bukowy *et al.*, 2018[72]	Rat	Kidney	Glomeruli detection	Detection: (R-CNN)	74 kidneys, trichrome-stained
Heinemann *et al.*, 2019[73]	Mouse Rat	Liver	Pathologist-like scoring of NASH models	Classification: (Inception-V3)	258 cases, trichrome-stained
Asay *et al.*, 2020[74]	Mouse	Lung	Tuberculosis pulmonary pathology	Classification: (Modular CNNs)	176 slides, H & E
Yurttakal *et al.*, 2020[75]	Rat	Kidney	Diabetic versus nondiabetic	Classification: (VGG19)	396 slides, H & E
Kumar *et al.*, 2020[76]	Dog Human	Mammary tumor	Tumor detection	Classification: (VGG-16)	352 slides, H & E
Aubreville *et al.*, 2020[77]	Dog	Skin tumor	Counting mitotic figures	Segmentation, detection, and regression: (U-Net, RetinaNet, customized CNN with ResNet50 stem)	32 cases, H & E (public dataset[33])
Zormpas-Petridis *et al.*, 2020[78]	Human Mice	Abdominal tumor in mice	Mapping tumor heterogeneity	Classification: (Super-resolution CNN)	13 specimens, H & E
Nonclinical safety and toxicologic pathology					
Bigley *et al.*, 2016[79]	Rat/human	Xenograft	Counting and classifying mitotic figures into different types (normal, aberrant, and degenerate)	Classification: Image analysis (filters and shape distinction)*	60 slides, H & E
Horai *et al.*, 2017[80]	Not mentioned	Liver Adrenal gland Spleen Kidney Intestine Lung Adipocyte	Quantifying specific histopathological findings such as vacuolation, hypertrophy, inflammatory cell infiltration, and necrosis in liver	Segmentation: Image-pro plus image analysis (filters and shape distinction)*	Not mentioned
Sonigo *et al.*, 2018[81]	Mouse	Ovary	Ovarian follicle counting	Classification: (CNN inspired by VGG19)	194 slides, H & E
Yu *et al.*, 2018[82]	Rat	Liver	Liver fibrosis staging	Classification: (AlexNet)	25 rats, collagen-stained
Horai *et al.*, 2019[67]	Not mentioned	Liver Kidney Thymus Skeletal muscle Spleen Adipocyte Parotid gland Sublingual gland Adrenal gland	Quantifying specific histopathological findings such as vacuolation, hypertrophy, bile duct proliferation, and necrosis in liver	Segmentation: HALO (image analysis, such as filters and shape distinction and random forest*	Not mentioned
Hu *et al.*, 2020[83]	Rat	Ovary	Ovarian toxicity assessment based on corpora lutea count	Detection: (Model based on RetinaNet)	224 slides, H& E
Hoefling et al., 2021[52]	Rat	46 different tissue types	Normal histology	Classification: (VGG-16, Inception-V3, ResNet-50)	1690 slides, H & E
Rudmann *et al.*, 2021[84]	Mouse	Lung Thymus Stomach	Carcinogenicity	Segmentation: Deciphex (inception, resnet-50 efficientnet)	170 slides, H & E
Pischon *et al.*, 2021[86]	Rat	Liver	Hepatocellular hypertrophy quantification	Segmentation: visiopharm (U-Net)	28 slides for training, H & E
Mudry *et al.*, 2021[87]	Rat	Eye	Retinal atrophy evaluation	Segmentation: MATLAB (VGG-16)	112 rats, H & E

Contd...

Table 2: Contd...

Reference	Species	Tissue	Application	Method	Dataset
Hvid *et al.*, 2021[88]	Rat Minipig	Mammary gland Oviduct	Quantification of epithelial proliferation	Segmentation: HALO (DenseNet, VGG)	31 rats, 18 minipigs, H & E
Carboni *et al.*, 2021[89]	Rat	Ovary	Ovarian follicle counting	Detection: (fast R-CNN)	1450 images, H & E
Tokarz *et al.*, 2020[90]	Rat	Heart	Cardiomyopathy scoring with artifact segmentation	Segmentation: AIRA matrix (FCN8s-ResNet50)	300 slides, H & E
Xu *et al.*, 2021[91]	Mouse	Testis	Spermatogenic staging	Segmentation: (U-Net)	12 slides, H & E
Creasy *et al.*, 2021[92]	Rat	Testis	Spermatogenic staging	Segmentation: (U-Net)	33 slides, H & E
Smith *et al.*, 2021[93]	Monkey	Bone	Quantification of bone marrow cellularity	Segmentation: Aiforia (details not mentioned)	6 slides for training, H & E
Bédard *et al.*, 2021[94]	Mouse	Colon	Quantification of DSS-induced colitis	Segmentation: Aiforia (details not mentioned)	65 slides, H & E
Ramot *et al.*, 2021[95]	Mouse	Liver	Quantification of liver fibrosis	Segmentation: AIRA matrix (U-NET)	140 microscopic field images, PSR stained
Freyre *et al.*, 2021[71]	Rat	Kidney	Biomarker level classification with localization of renal lesions	MIL classification: (HistoNet[52] and ImageNet as feature extractor)	349 slides, H & E
Kuklyte *et al.*, 2021[96]	Rat	Liver Kidney Heart Lung Brain	Segmentation of selected abnormalities	Segmentation: Deciphex (multi-magnification CNN architectures)	1342 slides, H & E

*Are not DL-based but are relevant image analysis-based in toxicologic pathology domain. NASH: Nonalcoholic steatohepatitis, R-CNN: Region-based convolutional neural network, CNN: Convolutional neural network, MIL: Multiple instance learning, DL: Deep learning, DSS: Dextran Sulfate Sodium, PSR: Picro-Sirius Red,

instance, in a performance evaluation of an algorithm for the detection of metastatic breast cancer within lymph nodes, artifacts including poor tissue fixation/processing, floaters/contaminants (i.e., non-nodal tissue), and out-of-focus regions led to false positives.[26]

Computational image analysis

Computational image analysis, including AI-assisted objective toxicity scoring and morphological assessments, is a bourgeoning field in toxicologic pathology. Inter-and intra-observer variability and bias have long troubled pathologists' attempts at harmonizing diagnostic thresholds across programs and studies.[36,80,100] Semi-quantitative severity scoring is most commonly applied to discern the magnitude of a change, with variable and subjective definitions for each score between individuals and organizations. Ordinal values are assigned to different severities (e.g., 0 = normal, 1 = minimal, 2 = mild, 3 = moderate, 4 = marked, 5 = severe) where the numerical values themselves are not ratio data and are therefore not appropriate for parametric statistical analysis.[100] In traditional pathology practice, the findings are typically not objectively assessed and measured on a regular or rolling basis, so drift does occur. DL-based quantitative assessments can potentially bring coherence and objectivity to pathology evaluations. ML and DL examples in the literature have demonstrated that AI performs with good concordance to pathologist assessment, often with improved sensitivity and efficiency.[101] In addition, DL models have proven to be successful in quantifying morphological assessments of ROI or lesions from H & E slides in the heart,[90] testis[91,92] ovary,[81,89] eye/retina,[87] and liver.[85,86] In addition, DL-based morphological quantifications are a promising alternative for subjective assessments as well as for alleviating the burden of manual semi-quantitative analysis by pathologists. These outputs may be used to standardize pathologists' assessment of the nature and severity of lesions present and potentially link to biometrics, molecular signatures, drug exposures, or clinical outcomes. Use cases of scoring and quantitative assessments in toxicologic pathology are briefly presented in the following paragraphs.

Discerning background heart findings from potential toxicologically relevant changes can be challenging in rat toxicity studies. There can be considerable overlap between key features of spontaneous rodent progressive cardiomyopathy (PCM) and test item-related findings, including necrosis, inflammatory cell infiltrates, and fibrosis. It is critical to distinguish these spontaneous findings from those which are toxicity-related. Although there is no known translational equivalent of PCM in humans, there are cases in which a test item induces higher severity or incidence of PCM within the study. A DL model has been developed to detect, classify, and score spontaneous cardiomyopathy in the rat and mouse heart.[90] It remains to be shown how the CNN would

perform when applied across a test set containing both PCM and cardiotoxic findings.

In traditional toxicologic pathology, the evaluation of testicular tissues in a stage-aware manner is a regulatory recommendation and is currently done manually. The stage-awareness of testes requires broadly categorizing seminiferous tubules into the early, mid, or late stage of the spermatogenic cycle, a task that is amenable to automation.[92] DL-based staging assessments are being developed for the mouse[91] and rat[92] testes. Beyond eliminating the potentially burdensome and specialized task of stage-aware evaluation of the testis, analyzing stage frequencies and germ cell quantifications of treated groups against a concurrent control group could allow for the detection of slight perturbations in spermatogenesis.[92]

Similarly, for the registration of new chemicals, a female reproductive assessment for potential toxic effects is a regulatory requirement. This assessment entails the quantification of primary and growing follicles of the ovaries of the offspring, a highly tedious and time-consuming task for which DL has been utilized to count primordial follicles in the mouse ovary[81] and follicles in the rat ovary.[89] These solutions alleviate the need for time-consuming manual counting, promise to reduce inter-observer variability, and enable routine ovarian corpora lutea enumeration in rat toxicity studies.[83]

In ocular histopathology, quantification of retinal atrophy is another example where DL-based algorithms have the potential to reduce labor-intensive measurement and at the same time minimize drawbacks of manual pathology evaluation, including diagnostic drift and observational bias. VGG16-based DL models have been applied to detect retinal atrophy in H & E stained slides from rats.[87] Briefly, WSIs were binarized, and retinas converted to 10,746 patches to train a retinal classification model and then a nuclear layer classification model. Pathologist-annotated retinal layer measurements were used as ground truth. Using DL to detect differences in the thickness of the retina and its inner and outer nuclear layers enabled the identification of retinal atrophy, which can occur as a test item-related effect or from other causes (e.g., light exposure and rodent strain).[87]

Hepatocellular hypertrophy and vacuolation are very common drug or chemical-induced lesions seen in toxicity studies. As compared to quantitative liver weight assessment, histopathology is still the best method to diagnose hepatocellular alterations, even though it is subject to interobserver variability and differences in visual perception. The quantitative evaluation of hepatocellular hypertrophy in the rat has recently been enabled by DL approaches.[86] In a stepwise manner, the hepatocytes were first segmented according to lobular regions. These sub-anatomic locations have important implications for drug metabolism and cellular physiology. Next, the mean cytoplasmic area was calculated for each of the three regions (centrilobular, midzonal, and periportal). Finally, a known inducer of hepatocellular hypertrophy, phenobarbital, was administered to rats, and H & E liver slides were analyzed against vehicle-treated controls. This method achieved similar results to gold-standard pathologist grading, as well as liver weights and gene expression. In a second approach, the DL model was trained to detect hypertrophy without any prior zonal segmentation of hepatocytes. This approach achieved similar results as those with zonal segmentation.[86] In a mouse model of hepatic steatosis, in which hepatocytes become distorted by cytoplasmic fatty vacuoles, a DL-based quantitative assessment of hepatocellular vacuolation provided a strong, significant correlation between the quantitative automated measurement of steatosis and the semi-quantitative pathologist-assigned score ($r = 0.89$).[85]

When it comes to the applicability of AI in DP, there are two main areas of focus for quantitative scoring: (a) attempting to duplicate the pathologist panel approach where a ground truth from three or more pathologists is used to develop the training set, or (b) true quantification of the morphologies that cannot be manually assessed. The use of multiple reviewers improves the baseline of the DL models for better reproducibility of scoring and provides objective assessments and quantitation. Utilizing the pathologist's feedback regularly during the test would reduce the drift.

Computer-assisted abnormality detection

A final solution for abnormality triaging has not been developed, yet it would be the single most impactful ML/DL-based application in toxicologic pathology. This solution will screen out the normal slides, allowing pathologists more time to focus on slides with abnormalities. This workflow keeps the human (pathologist/SME) in the loop for critical decision making with regard to the specific nomenclature and interpretation applied to a finding, and DL could further assist by applying standardized grading across groups once an abnormality of toxicologic significance is confirmed. Computer-assisted abnormality detection is not entirely new in CDP, with examples applying proprietary algorithms in cervical cytology screening surfacing in the 1990s. Two FDA-approved automated imaging systems, the FocalPoint GS Imaging System and the ThinPrep Imaging System are widely used today for this purpose.[102-104] Advances in DL have made considerable headway in CDP applications, particularly regarding assisted cancer diagnostics and immunophenotyping in surgical biopsy specimens, which has the potential to accelerate workflow and tangibly improve diagnostic accuracy.[65,105-107] In human and veterinary diagnostic cytopathology, proprietary, neural-network-based pre-classification systems are widely used, particularly for hematology applications.[108,109] Most recently, two AI-based decision support tools in CDP have gained FDA Breakthrough Device designation: Paige. AI in 2019 for cancer histology[5] and 4D Q-plasia OncoReader Breast in 2020 for breast cancer histology.[110]

An abnormality-agnostic method based on normal tissue identification has been advocated as an alternative approach to the detection of specific abnormalities. In this paradigm, the model is trained only on normal tissues, and outlier patches

would be flagged broadly as abnormalitiesor foci potentially containing toxicologically significant findings - for pathologist review. Establishing this workflow is predicated on a foundation in the features of normal tissues. For human tissues, there have been some efforts in this area. A patch library representing up to 57 histological tissue types based on a 3-level hierarchical taxonomy of tissue architecture (ranging from least specific, e.g., epithelial, to most specific, e.g., stratified squamous epithelial) has been used to train VGG-16, ResNet18, and Inception-v3 networks on tissue type classification.[111] A purported benchmark against which AI model performance can be measured may exist in DAPPER,[112] which evaluates the accuracy and feature stability of ML classifiers by comparison against a GTEx-derived dataset of normal histology. Similar efforts in toxicologic pathology towards abnormality detection have largely focused on training neural networks on histologic atlases of normal tissues. A rat tissue slide catalog of 1690 slides comprising 46 different tissue classes was imaged at six magnifications, and patches were used to train VGG-16, ResNet50, and Inception-v3 networks to identify histologically distinct tissues.[52] Some important findings included improved reliability of tissue prediction with decreasing magnification, owing to greater context, except in the case of small tissues (e.g., parathyroid gland), where higher magnification was required, as well as confusion in less histologically distinct tissue regions (e.g., segments of the large intestinal tract).[52] Through UMAP visualizations, the investigators also found that the networks were learning sub-structural elements of organs that had not been given explicit labels in training.[52] Furthermore, these trained models on rats served as the basis for transfer learning in nonhuman primate and minipig tissues.[52] Finally, representing a middle ground between a broadly-trained CNN that recognizes normal histology across multiple studies and one that is trained to recognize a single class of abnormalities, Deciphex's Patholytix AI provides computer-aided diagnosis by developing classifiers based on a concurrent control set within the study, then presenting a color-coded output indicating the location and class or severity of the detected abnormality. An evaluation of this system in carcinogenesis studies using Tg-RasH2 mice utilized three top-performing CNN models based on U-Net architecture with an F1 score (dice coefficient; details for evaluation metrics in Section "Postprocessing and evaluations") threshold of 0.7 to determine acceptable performance,[84] based on ground truth assessment by pathologists. The selected and trained CNN, EfficientNet-b0, automatically produced masks for lung, stomach, and thymus upon scanning, allowing the pathologist to simultaneously review the digitized slide and AI overlay, which indicated the presence, location, and class of proliferative change present.

Recently, a supervised DL model has been developed to detect selected lesions (abnormalities) in selected organs.[96] To achieve their models, the group prepared a private, thousand-slide, pathologist-reviewed, exhaustively annotated WSI database. Due to the high class imbalance (very low percentage of lesions vs. normal tissue), the multi-class classification for detecting multiple lesions concurrently was not successful; however, consolidating the data, i.e., considering all lesion types as one class, helped the models to be able to classify lesion versus no lesion successfully.[96] Another consolidation approach was also examined: A selected lesion versus all other lesions in the dataset plus normal tissue. The performance of the latter classifiers was good on some lesions with higher occurrence in the dataset. For example, the one-versus-all classification approach in kidneys resulted in F1 scores of higher than 0.9 in mineralization, casts, and infiltration, while F1 score in tubule degeneration was only 0.55.[96] The authors noted that despite the robust database, the data did not contain all the abnormalities that might occur during a toxicologic pathology study.

Developing an AI-enabled DP workflow that could rapidly screen out the normal tissues will enable pathologists to prioritize the principal task of reviewing only the abnormalities – "slides that matter"[36] – allowing for higher overall throughput and shorter study timelines. The bulk of a toxicologic pathologist's work is centered around triaging normal and abnormal tissues, and although an AI-based solution is conceptually attractive, it remains just out of reach in toxicologic pathology. Currently, there might still be some challenges in the definitions. First, there is no true definition of normal; it is highly context-dependent because "normal" changes with age, gender, diet, strain, etc. Another problem is that the changes associated with "abnormal" are even more heterogeneous and much broader in their range of severity. These vague and subjective definitions will increase the discrepancies between pathologists and a DL model. In addition, a DL model would be expected to recognize abnormality types that it has never seen before. Currently, the efforts are mostly in a simplified version (in a specific tissue type, species, and abnormality type).[67,80] These will be the first steps toward accomplishing a DL-based system that could significantly accelerate the pathologist's review of normal and abnormal tissues.

Content-based image retrieval (CBIR)

Nowadays, photographs can be indexed and easily mined to identify people, objects, and locations – allowing a user to search through thousands of digital photos. This technology is extremely mature and is now being applied to histopathology.[113] However, unlike photos, a WSI is gigapixels in size with typically about 100,000 × 100,000 pixels, which makes indexing and mining a challenging task. Content-based image retrieval (CBIR) allows to search and retrieve images from a digital slide database based on similarity to a query image or feature or perhaps natural language query text.

A reliable CBIR system is based on image feature extraction, a robust digital image database, and efficient similarity metrics.[114] Digital images are high-dimensional data. Therefore, in order to separate noise from signal and to achieve a meaningful image query, a similarity metric

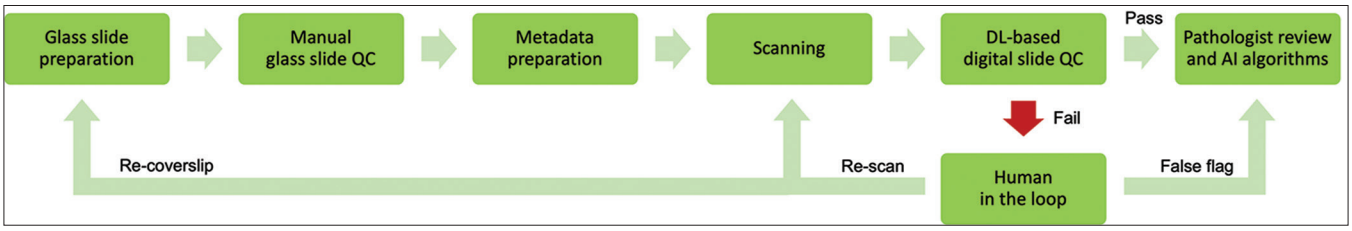

Figure 8: Flowchart for implementation of digital slide quality control

based on a weighted approach should be employed. Based on this approach, the discriminatory features from the query image should carry more weight than background features such that the images that are relevant to the query image have a higher probability to be retrieved (higher precision retrieval). The focus of pathology image analysis services like Google Smily[115,116] is to allow the pathologist to identify an ROI and search across "like images" that have been tagged and validated to help speed up the identification and provide consistent semi-quantitative answers. If designed correctly, the system can provide a "match" score and assist the pathologists in their decisions. CBIR also proves to be useful during annotations to search for visually similar annotations in the database, thus accelerating the annotation process.[117]

In toxicologic pathology, chemical- or drug-induced lesions come in various morphological forms and severity, and sometimes these lesions could overlap with or exacerbate background or spontaneous findings. Since diagnosis is applied to the entire image or set of images, finding the image patch that is representative of the diagnosis could be very challenging. CBIR can be used to fetch images or regions of interest where image metadata alone may not be sufficient. CBIR could also be another tool in the pathologist's toolbox to quickly search and discover images from large repositories instead of having to rely on their memory for historical cases. Therefore, a large-scale and relatively structured image database is the foundation for an efficient and powerful CBIR system. Fortunately, most institutions practicing toxicologic pathology (CROs and large pharmaceutical industries) generate or have archived large number of glass slides which if, strategically scanned, can build massive amounts of structured digital data that can be used for training and implementing a large-scale CBIR system.

Challenges In Implementation Of Deep Learning-Based Applications In Toxicologic Pathology

The analysis of histopathology images poses unique challenges. Despite the specific impressive examples and applications of AI in DP, there are clear obstacles that limit the employment of DL methods in toxicologic pathology.[2] What follows will cover recent efforts in response to these challenges.

Regulatory landscape

A significant portion of toxicologic pathology work is conducted under a highly regulated good laboratory practice (GLP) environment guided by documents issued by FDA[118] and other regulatory bodies (e.g. Environmental Protection Agency and Organization for Economic Co-operation Development). The principles of GLP are meant to assure the public that sound scientific practices were used during the conduct of the nonclinical study and data collection. Like other systems currently used in GLP studies, validation would be a requirement for DL-based software. However, deployment of DL-based software in a regulated workflow can be a daunting task as not only the DP system used to generate WSI needs to be validated but also the AI-based software. Early on and due to lack of substantially equivalent determination, FDA classified WSI systems as class III medical devices – the "highest risk" category and requiring general controls (quality system regulation, good manufacturing procedures) and premarket approval.[61] In 2015, FDA published technical performance guidelines, which created a roadmap for scanner manufacturers to get FDA approval leading to the first approval (under the *de novo* pathway) of a complete WSI system for primary diagnosis in surgical pathology in 2017.[19] Importantly, with this approval FDA reclassified WSI systems as class II medical devices, requiring manufacturers to get approval through 510(k) by demonstrating "substantial equivalence" to a previously approved device.[119] With this precedent in place, many more scanner manufacturers will likely submit their devices for FDA approval.

Recently, FDA has approved multiple DL-based software and medical devices for clinical use to support several medical specialties, including radiology, cardiology, and ophthalmology.[120] Although many DL-based pathology use cases have been published in the literature, so far, there is no FDA-approved software on pathology data. One AI-based pathology company has received FDA's Breakthrough Designation and is promising to bring AI-products across many cancer types to support pathologists.[121] From a regulatory perspective, software intended to be used for medical purposes without any accompanying hardware is considered software as medical device (SaMD), and DL-based SaMD brings a unique challenge owing to its ability to adapt or change after the approval process. In 2021, FDA released a discussion paper and solicited public feedback on a proposed regulatory framework for AI/ML-based software.[122] Based on the feedback received

from a wide array of stakeholders, FDA recently issued an action plan for AI/ML-based SaMD.[123] It is only a matter of time before we will start seeing FDA-approved DL-based software in the pathology domain.

Nonclinical safety and toxicologic pathology are highly regulated practices, making it mandatory to audit and verify any decisions made by a machine. This will increase the importance of model interpretability and the demand for the ability to question, understand, and trust DL systems.[63,124] In addition, suitable interpretation methods must be devised to understand the predicted outcomes and provide evidence for explanatory and regulatory purposes. Trained DL models have generally been considered "black boxes" due to their lack of interpretability while making predictions.[125] Although a pathologist does not have to understand all the technical details of DL models to determine their performance, regulatory agencies might want to know how the models work. Thus, it is essential to draw some form of the reasoning, either by a pathologist or by the data scientist who is part of the team, behind the predictions by exposing the black box as much as possible through visualization of the DL neural network embeddings to demonstrate how the output was reached. One approach is to localize the areas in the WSI that contribute significantly to a prediction, which would help pathologists interpret the results and gain insight into the evidence therein. A few methods have been developed to visualize the neural network embeddings that substantiate the achieved outcome: (a) visual attention maps,[126] (b) heatmaps,[127] (c) saliency maps,[128] and (d) image captioning.[43] For instance, Tellez *et al*.[128] have used saliency maps using Gradient-weighted Class Activation Mapping[129] to interpret how their Neural Image Compression model captures semantic features at the patch level and identifies the areas at the slide level that contributed significantly to the tumor prediction. In an AI-based detection system for nodal metastasis of breast cancer, the 5% most important pixels with 1 μm dilation were used for heatmap visualization.[25]

Infrastructure

To be able to benefit from DL-based applications in toxicologic pathology, a DP workflow must first be implemented. In addition to the lack of guidelines/clarity regarding the nature and extent of DP system validation for use in highly regulated disciplines like toxicologic pathology, infrastructure is one of the biggest hurdles in the implementation of DP workflow in toxicologic pathology.[34] For DP to be accepted and adopted in a routine nonclinical toxicologic pathology evaluation, the process must be equally or more efficient, accurate, and user-friendly than evaluating tissues using a microscope. In addition, the implementation of DP must be cost-effective to gain budgetary approval of DP endeavors. The hardware, data storage systems, internet speed, speed of loading the WSI, the refresh rate of a monitor, and image management software must provide a seamless experience with no lag or downtime for pathologists.[2]

Currently, the cost of DP workflow implementation is more than what it costs to evaluate tissues using a microscope. This is one main reason why complete implementation of DP workflow has not occurred in many organizations. Until the benefit-cost ratio and the user experience of DP workflow are on par with or demonstrate a clear improvement over microscopic evaluation, infrastructure will continue to be a major challenge.

Image format, size, and magnification

WSIs may come with different image formats depending on the scanner. This variation in file format might cause an extra step of unifying file format before a DL model development. While the DICOM (digital imaging and communications in medicine) format offers the benefit of interoperability and remote telepathology, this has not yet been adopted as the standard at this stage in DP. A conversion-based approach is typically employed using open-source or commercial programs to bring legacy, proprietary files into a unified format in either a high-throughput or in-line manner.[130]

Digitization of glass slides rapidly generates massive amounts of data, necessitating large and thoughtfully designed networks of stored and cached data. The choice of scanning magnification depends on finding the right balance/tradeoff to best optimize resolution and image size (data storage considerations). A typical 40X (Please standardize magnification format) scan with a 0.25 μm pixel resolution and 24-bit color depth contains 384 million bits of information within a single 1 mm^2 area of the slide, resulting in a file size of 48 MB if no further steps are taken in data efficiency.[19] Therefore, depending on the tissue size, a WSI can have a file size of up to 6 GB. Due to the large size of WSIs, before any image analysis, each WSI must be broken down into hundreds or thousands of smaller tiles. Magnification in digital image analysis can be lower than scanning magnification and is defined by closely the image needs to be zoomed in for the analysis to be able to resolve the level of detail that will discriminate objects of interest [Figure 3]. The magnification and the patch pixel size will define the number of patches per WSI. Without down-sampling (resizing), the field of view will have a large pixel size, and deep neural networks with larger input sizes would need much deeper topology and a much larger number neural mappings making them even more difficult and perhaps impossible to train. Thanks to advances in accelerated computing leveraging graphical processing units (GPUs), a la larger input size (up to 1024x1024) can be trained. However, most DL models in DP usually use 256 × 256 or 512 × 512 patch size to be able to perform on low GPU systems. Still, extracting patches from a high-resolution WSI requires down-sampling prior to feeding them into a deep network.

The next step is to determine which magnification and patch size are the best choices for DL WSI analysis. The selection of magnification can be variable and dependent upon the "intended use" of the AI-based model. However, there must be a balance of magnification along with down-sampling to get the

appropriate patch pixel size. Down-sampling these tiles may result in the loss of crucial information that would justify the use of the highest magnification to have the least down-sampling effect. However, the use of a high magnification such as 40X is not practical because a deep network that is trained on higher magnification will not only be slower to analyze but will also lack the contextual information at lower magnification. Although the context at low magnification possesses key information, higher magnified tiles would determine the details that would be needed for a specific task. Therefore, a DL model might perform differently at different magnifications. While a specific class (e.g., bone tissue) can be detectable at low magnification, prediction of another class (e.g., parathyroid tissue, germ cells for spermatogenic staging) can be more accurate at higher magnification.[52] To address this challenge, instead of training a separate network for each level of magnification, multi-magnification architectures have been proposed to integrate information from multiple scales.[131] The multi-scale architectures use multiple encoders and/or multiple decoders, offering the dual benefit of context and resolution. Another approach is modeling neighboring patch correlation that takes the contextual information into account.[43] Due to the disjointed/random selection of tiles/patches, loss of visual context is inevitable because the order of patches defines the textures in a WSI. The optimal magnification and tile size depend on the complexity of the task. The best approach is to experiment with the training to find a balance of image fidelity and computational efficiency for the algorithm.

Amount of annotated images

An important statistic for the success of any DL task is the abundance of training data. As described in Section "Deep learning methods", in supervised learning, WSI themselves are insufficient for training and must be appropriately regionalized and labeled to leverage supervised learning techniques successfully. Label information from ordinary pictures can be easily accessible, and it does not need any special expertise (e.g., anyone can identify objects such as cats vs. dogs). However, only an experienced pathologist can label a pathology image accurately. Therefore, even if a large repository of images is available, there might not be enough labeled data (ground truth) for training. Furthermore, public data sets with hand-annotation can only be useful if there is a similar task/purpose, staining, magnification level, and resolution. The quality and extent of the curated data set ensure the robustness and functionality of DL models. Segmentation models are very sensitive to correctly defining the borders or ROI in WSIs,[132] i.e., labeling at pixel/patch-level. Pixel-wise annotating an entire WSI requires a lot of time and labor. Researchers take four approaches to increase the amount of annotated data: (1) using data augmentation, (2) increasing the amount of training data efficiently, (3) utilizing transfer learning, and (4) implementing weakly-supervised learning approaches.

Data augmentation is the first approach that should be taken when there is not enough image data or labeled images available for training. Synthetic images (augmented images) can be generated based on certain rules in order to augment the training data. Image augmentation techniques include arbitrary rotation, patch flipping on a vertical or horizontal axis, and HSV (hue, saturation, value) variable manipulation by random number multipliers.[133] Data augmentation is also useful when the training data set does not contain images that are diverse enough to reassure the generalizability of the DL models. Stain color augmentation has been shown to improve the performance of classifiers on almost all experimental scenarios.[134] Due to the effects of data augmentation on the performance of models, it has become a standard preprocessing step in most DL approaches.

Efficient labeling is the approach that can be taken to enhance the performance of annotating procedure. For example, QuPath is an open-source easy-to-use graphical user interface that can automatically refine ROIs around the targeted objects.[135] Furthermore, labeling can be sped up by implementing an integrated workflow to localize the ROIs during the pathology practice by tracking the pathologist's eye movement[136] or mouse cursor positions.[137] In practice, one of the most reliable approaches to increase the amount of annotated data needed for training is the initiation of the training by a small amount of annotated data, refining the predicted label by a pathologist, and adding the newly refined annotations to the training data set.[138] Refining and reviewing labeled data are much easier than hand annotating a WSI from scratch. Active learning is another approach that uses machine learning to identify the most valuable unlabeled data. In concept, instead of labeling all WSIs available, a classifier chooses the images for labeling based on their expectancy in the improvement of the classification performance which can be defined by the confidence of the model in its prediction (i.e., having a low probability). Then, a pathologist can annotate and add the data to the training data set. In a study investigating DL for assisted detection of prostate cancer, the CNN had an increased likelihood of adding a patch to the training data if the center pixel of the patch was initially classified incorrectly by the network, thus feeding back additional training data containing more challenging samples.[106]

Transfer learning is the most widely used approach to overcome the need for a large, *de novo* training data set. A DL model developed on a source domain with a huge amount of data can be fine-tuned (used as a pretrained model) to learn a target domain with a lesser amount of data. This domain adaptation is called transfer learning, and the goal is to extract the knowledge from one domain and apply it to another. For example, a transfer learning model based on AlexNet has been able to score liver fibrosis stages using only 25 rat liver WSIs.[82] Transfer learning is also overcoming the limitation of DL models to the specific tasks for which they have been designed. For example, a model developed for the detection of metastatic breast cancer in lymph nodes will not be able to correctly identify other pathology that may be present in the slide, such as lymphoma or an infection,[107] much less the presence of two or more concurrent pathologies. Specific to toxicologic pathology, transferring the domain between

species or organs is challenging. Unlike human pathology, multiple species (rodent and nonrodent) are commonly used in toxicologic pathology. For example, a liver in mouse and monkey might have similar basic histology, but a model in a mouse liver may not have the same performance on monkeys without transfer learning. When the direct cross-species prediction was not performing well for normal histology, transfer learning, retraining on a new domain, proved to be effective in cross-species domain adaptation.[52]

Weakly supervised-based models are another approach that incorporates weak labels and does not need exhaustively annotated data. The slide-level label is more likely available in most pathology studies, including toxicologic pathology. Avoiding pixel-wise annotations and focusing on the slide-level label, Campanella *et al*. have been able to scale their database to 44,732 WSIs.[105] Details in weakly supervised learning models can be found in Section "Deep learning methods".

While there are some shared data sets in CDP, and many efforts are underway to facilitate this process,[40] such shared, pathologist-annotated datasets are not available in toxicologic pathology [Table 1]. Each organization trying to implement AI in its workflow can develop its private, time-consuming, and expensive dataset. However, it will benefit the entire toxicologic pathology community and speed up the implementation process if shared datasets of common pathology abnormalities in animal species can be created.

Class imbalance

In biomedical image applications such as DP, imbalanced datasets are a common problem since some classes are rarely occurring. In toxicologic pathology, this issue is even more pronounced specifically for abnormality detection because most of the tissues are normal, and abnormalities can be highly variable and infrequent. Balancing the dataset by expanding under-represented classes may alleviate the imbalance, but this could be challenging and sometimes impossible. Other techniques that can be used to overcome this challenge are loss weighting (the loss computed for different samples will be weighted differently based on whether they belong to the majority or the minority classes). Under-sampling the majority classes or oversampling the minority classes can also be effective. In a systematic review of different scenarios in multi-class classification, oversampling has been shown to be the most effective approach.[139] In DP, oversampling might increase the likelihood of overfitting, which might affect the performance of the models during testing.

Stain variations and artifacts

During tissue processing, microtomy, and staining procedures, various artifacts can inadvertently be induced to an image that can add unintended noise and affect the AI models' performance.[2] These artifacts can also interfere with glass slide microscopy assessments.[140] This challenge is present in both CDP and toxicological pathology, but toxicologic pathology samples have generally higher quality. CDP samples suffer from being very small in size collected using various instruments, or they might be collected after an extended postmortem interval. On the other hand, the toxicologic pathology samples represent full-thickness tissue sections obtained immediately following euthanasia that are batch-processed in the histology laboratory, which ensures higher quality slides and images. Examples of common glass slide preparation artifacts are stain fade, overstaining, tissue tear, folds, debris, mounting bubbles, autolysis, and uneven cut.[140] Having experienced histology staff and high-quality laboratory equipment can help decrease the occurrence of these artifacts. Some other artifacts only occur in WSIs and relate to poor scanning, for example, stitching artifact, out-of-focus areas, and missing tissue (i.e., the scanner misses some parts of the tissue that are actually present on the glass slide). In these cases, rescanning might help to to improve the quality of WSI.

The staining intensity may have variability due to differences in protocol design, reagent quality, and section composition during glass slide preparation.[141] H & E staining variability is a well-recognized preanalytical confounder,[141] and different techniques have been employed to either augment the training dataset to accommodate fluctuations or normalize the test set through color transformations that approximate the training set.[142] One technique investigated in a study of CNN-based detection of nodal metastatic breast cancer was transforming colors into a hue-saturation-density space to account for the nonlinear relationship between the stain and pixel intensity values. This transformation was applied to change color statistics into a reference slide, with the median color statistics across the training set used as a reference to perform normalization.[26] Another insight into the value that color information brings to a deep CNN was demonstrated in a study of computer-aided *Mycobacterium tuberculosis* bacillus detection in WSI, wherein the accuracy in Ziehl–Neelsen acid fast-stained slides was 95.3% but decreased to 73.8% in the same decolorized/binarized images.[143] Decreases in the area under the curve (AUC, an aggregate measure of performance) have been observed when there is a perceived difference in brightness, contrast, and sharpness due to the use of a different scanner.[105] As a result, training on a mixed dataset, including scans from multiple instruments or fine-tuning a DL model using images from a different scanner, may be necessary to ensure consistent performance.[105] Illustrative of this point, when the CAMELYON16-trained model for nodal breast cancer metastasis detection was applied to a test set from Memorial Sloan Kettering, there was a 20% drop in AUC.[105] In fact, it has been shown that without compensation for variation among images sourced from different laboratories, nonmorphological differences such as color variations are more prominent in t-SNE plots feature representations, more details in Section "Deep learning methods" of the feature space than in real morphological differences between specimens of the same diagnostic class.[53] The presence of stain variations and artifacts demonstrates the need to focus on manual or digital QC to adjust the current methodology for slide preparation and scanning. Automated

digital QC methods presented in Section "Computer-assisted quality control") for artifacts detection can not only benefit in delivering high-quality glass slides for pathologists' daily practice but can also assist in creating quality data with a high signal-to-noise ratio for the downstream AI models.

Summary And Outlook

Nonclinical toxicity studies are an essential step in evaluating the safety of INDs or chemicals before they can be approved for human use. While a toxicologic pathologist's evaluation is currently the gold standard in assessing target organ toxicities and microscopic alterations related to drugs and dosages, manually reviewing thousands of tissue slides is extremely tedious and time-consuming. Thus, the field is ripe for technological modernization.

Our outlook is a fully digitized toxicologic pathology workflow. If toxicologic pathologists want to experience and benefit from the renaissance of digitized workflow transformation, substantial investments in training and equipment must be made toward digitalization.[34] The adoption might be slower initially, but after adequate training and acclimatization, overall pathologist performance (speed, accuracy) increases.[98] Owing to the need for a massive investment of time and infrastructure required on the front end, fully integrated DL-based workflows for digital toxicologic pathology are not likely to be implemented at this time outside of industry (pharmaceutical companies, contract research organizations) or large diagnostic and medical research centers. Those well-resourced disciplines have an opportunity to develop and deploy such systems that may in the future become widely accessible in the scientific community.

This paper provided a review and discussion of applications of AI/DL in toxicological pathology, which is transforming the traditional pathology practice into a computer-assisted digitized workflow. The first application of AI/DL in a DP workflow would be the digital slide QC to help prepare high-quality WSIs for both virtual microscopy and image analysis. Next, AI/DL can be used for computational image analysis. Recent publications have demonstrated the potential of computational image analysis in assisting toxicologic pathologists by automating labor-intensive measurements and at the same time minimizing diagnostic drift and observational bias originating from subjective evaluation. Most importantly, AI systems that could rapidly indicate the presence of abnormalities in a tissue potentially related to test-item administration is of utmost value to toxicologic pathologists. Rather than having specific DL models to recognize each possible categorical change (e.g., cancer, as is being applied in the clinic), learning and screening out the normal tissue and flagging the abnormalities will significantly cut down the time spent by toxicologic pathologists reviewing thousands of tissue samples thereby shortening the study timeline.[36] CBIR is also presented as a potential tool to assist toxicologic pathologists for data discovery and rapid retrieval of relevant images from archives.

Recent advances in response to the challenges in developing DL algorithms were also covered in this paper. Even with all the early success in supervised learning in pathology, the response of the pathology community in general to DL methods is mixed. One main criticism for current DL is the narrow nature of these algorithms and their inability to mimic diverse complements of skills and capabilities demonstrated by pathologists. Unsupervised learning may present the opportunity to broaden the scope of DL-based models and free pathologists from annotating the images. In the context of toxicologic pathology, the conceptual idea of unsupervised/weakly supervised learning seems highly enticing as toxicologic pathologists deal with a variety of tissues in multiple species with unlimited lesion representation. We also discussed some of the challenges in validating the DL models and getting regulatory approval of such software before its use in a GLP environment.

While there will be many challenges to overcome in bringing a remarkable change to toxicological pathology both technologically and culturally. culturally, keeping pathologists at the center in collaboration with AI scientists and engineers will be instrumental to the formation of multifaceted approaches that can deliver safer drugs to the patients faster.[2] On this note, we see tangible near-term successes and a bright future for AI/DL in toxicologic pathology, in which interdisciplinary teams have an opportunity to continue to build and strengthen explainable and adaptable AI to advance drug development.

Acknowledgments

The authors would like to thank Robert Dunstan, Rebecca Kohnken, and Erik Hagendorn, of AbbVie, for providing insightful comments and feedback on the manuscript.

Financial support and sponsorship

AbbVie sponsored and funded the study; contributed to the design; participated in the collection, analysis, and interpretation of data, and in writing, reviewing, and approval of the final publication. .

Conflicts of interest

All authors are employees of AbbVie and may own AbbVie stock. The mention of the vendors in this paper does not imply an endorsement or recommendation.

References

1. Marble HD, Huang R, Dudgeon SN, Lowe A, Herrmann MD, Blakely S, *et al*. A regulatory science initiative to harmonize and standardize digital pathology and machine learning processes to speed up clinical innovation to patients. J Pathol Inform 2020;11:22.
2. Tizhoosh HR, Pantanowitz L. Artificial intelligence and digital pathology: Challenges and opportunities. J Pathol Inform 2018;9:38.
3. Srinidhi CL, Ciga O, Martel AL. Deep neural network models for computational histopathology: A survey. Med Image Anal 2021;67:101813.
4. Salto-Tellez M, Maxwell P, Hamilton P. Artificial intelligence-the third revolution in pathology. Histopathology 2019;74:372-6.

5. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. Nat Rev Clin Oncol 2019;16:703-15.
6. Rashidi HH, Tran NK, Betts EV, Howell LP, Green R. Artificial intelligence and machine learning in pathology: The present landscape of supervised methods. Acad Pathol 2019;6:1-16.
7. Zhou X, Li C, Rahaman MM, Yao Y, Ai S, Sun C, *et al.* A comprehensive review for breast histopathology image analysis using classical and deep neural networks. IEEE Access 2020;8:90931-56.
8. Madabhushi A, Lee G. Image analysis and machine learning in digital pathology: Challenges and opportunities. Med Image Anal 2016;33:170-5.
9. Aeffner F, Zarella MD, Buchbinder N, Bui MM, Goodman MR, Hartman DJ, *et al.* Introduction to digital image analysis in whole-slide imaging: A white paper from the digital pathology association. J Pathol Inform 2019;10:9.
10. BenTaieb A, Hamarneh G. Deep Learning Models for Digital Pathology. arXive 2019;1910:12329.
11. Browning L, Colling R, Rakha E, Rajpoot N, Rittscher J, James JA, *et al.* Digital pathology and artificial intelligence will be key to supporting clinical and academic cellular pathology through COVID-19 and future crises: The PathLAKE consortium perspective. J Clin Pathol 2021;74:443-7.
12. Gauthier BE, Gervais F, Hamm G, O'Shea D, Piton A, Schumacher VL. Toxicologic Pathology Forum*: Opinion on integrating innovative digital pathology tools in the regulatory framework. Toxicol Pathol 2019;47:436-43.
13. Niazi MK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. Lancet Oncol 2019;20:e253-61.
14. Komura D, Ishikawa S. Machine learning methods for histopathological image analysis. Comput Struct Biotechnol J 2018;16:34-42.
15. Hayakawa T, Prasath VB, Kawanaka H, Aronow BJ, Tsuruoka S. Computational nuclei segmentation methods in digital pathology: A survey. Arch Comput Methods Eng 2021;28:1-13.
16. Wang S, Yang DM, Rong R, Zhan X, Xiao G. Pathology image analysis using segmentation deep learning algorithms. Am J Pathol 2019;189:1686-98.
17. Aeffner F, Adissu HA, Boyle MC, Cardiff RD, Hagendorn E, Hoenerhoff MJ, *et al.* Digital microscopy, image analysis, and virtual slide repository. ILAR J 2018;59:66-79.
18. Zarella MD, Bowman D, Aeffner F, Farahani N, Xthona A, Absar SF, *et al*. A practical guide to whole slide imaging: A white paper from the digital pathology association. Arch Pathol Lab Med 2019;143:222-34.
19. Roohi A, Faust K, Djuric U, Diamandis P. Unsupervised machine learning in pathology: The next frontier. Surg Pathol Clin 2020;13:349-58.
20. Acs B, Hartman J. Next generation pathology: Artificial intelligence enhances histopathology practice. J Pathol 2020;250:7-8.
21. Djuric U, Zadeh G, Aldape K, Diamandis P. Precision histology: How deep learning is poised to revitalize histomorphology for personalized cancer care. NPJ Precis Oncol 2017;1:22.
22. Elazab N, Soliman H, El-Sappagh S, Islam SM, Elmogy M. Objective diagnosis for histopathological images based on machine learning techniques: Classical approaches and new trends. Mathematics 2020;8:1863.
23. Parwani AV. Next generation diagnostic pathology: Use of digital pathology and artificial intelligence tools to augment a pathological diagnosis. Diagn Pathol 2019;14:138.
24. van Tongeren S, Fagerland JA, Conner MW, Diegel K, Donnelly K, Grubor B, *et al.* The role of the toxicologic pathologist in the biopharmaceutical industry. Int J Toxicol 2011;30:568-82.
25. Center for Food Safety and Applied Nutrition, Office of Food Additive Safety, Toxicological Principles for the Safety Assessment of Food Ingredients, Redbook 2000: Chapter IV.C.4.a. Subchronic Toxicity Studies with Rodents; November, 2003. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/redbook-2000-ivc4a-subchronic-toxicity-studies-rodents#test. [Last accessed 2021 May 25]
26. Liu Y, Kohlberger T, Norouzi M, Dahl GE, Smith JL, Mohtashamian A, *et al.* Artificial intelligence-based breast cancer nodal metastasis detection: Insights into the black box for pathologists. Arch Pathol Lab Med 2019;143:859-68.
27. Society of Toxicologic Pathology, International Harmonization of Nomenclature and Diagnostic Criteria (INHAND). URL: https://www.toxpath.org/inhand.asp#pubg. [Last Accessed 2021 April 01] .
28. Schuhmacher A, Gatto A, Hinder M, Kuss M, Gassmann O. The upside of being a digital pharma player. Drug Discov Today 2020;25:1569-74.
29. Yoshikawa T, Horai Y, Asaoka Y, Sakurai T, Kikuchi S, Yamaoka M, *et al.* Current status of pathological image analysis technology in pharmaceutical companies: A questionnaire survey of the Japan Pharmaceutical Manufacturers Association. J Toxicol Pathol 2020;33:131-9.
30. Zhang L, Zhang H, Ai H, Hu H, Li S, Zhao J, *et al.* Applications of machine learning methods in drug toxicity prediction. Curr Top Med Chem 2018;18:987-97.
31. Vo AH, Van Vleet TR, Gupta RR, Liguori MJ, Rao MS. An overview of machine learning and big data for drug toxicity evaluation. Chem Res Toxicol 2020;33:20-37.
32. Innovative Medicine Initiative, Translational Quantitative Systems Toxicology to Improve the Understanding of the Safety of Medicines; Available from: https://www.imi.europa.eu/projects-results/project-factsheets/transqst. [Accessed Date 22 Feb 2021]
33. Bertram CA, Aubreville M, Marzahl C, Maier A, Klopfleisch R. A large-scale dataset for mitotic figure assessment on whole slide images of canine cutaneous mast cell tumor. Sci Data 2019;6:274.
34. Schumacher VL, Aeffner F, Barale-Thomas E, Botteron C, Carter J, Elies L, *et al.* The application, challenges, and advancement toward regulatory acceptance of digital toxicologic pathology: Results of the 7th ESTP International Expert Workshop (September 20-21, 2019). Toxicol Pathol 2021;49:720-37.
35. Hanna MG, Reuter VE, Ardon O, Kim D, Sirintrapun SJ, Schüffler PJ, *et al.* Validation of a digital pathology system including remote review during the COVID-19 pandemic. Mod Pathol 2020;33:2115-27.
36. Turner OC, Aeffner F, Bangari DS, High W, Knight B, Forest T, *et al.* Society of Toxicologic Pathology Digital Pathology and Image Analysis Special Interest Group Article*: Opinion on the application of artificial intelligence and machine learning to digital toxicologic pathology. Toxicol Pathol 2020;48:277-94.
37. Prewitt JM, Mendelsohn ML. The analysis of cell images. Ann N Y Acad Sci 1966;128:1035-53.
38. Mukhopadhyay S, Feldman MD, Abels E, Ashfaq R, Beltaifa S, Cacciabeve NG, *et al.* Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: A multicenter blinded randomized noninferiority study of 1992 cases (pivotal study). Am J Surg Pathol 2018;42:39-52.
39. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44.
40. Moulin P, Grünberg K, Barale-Thomas E, der Laak JV. IMI-Bigpicture: A central repository for digital pathology. Toxicol Pathol 2021;49:711-3.
41. Xu Y, Jia Z, Wang LB, Ai Y, Zhang F, Lai M, *et al.* Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. BMC Bioinformatics 2017;18:281.
42. Qaiser T, Tsang YW, Taniyama D, Sakamoto N, Nakane K, Epstein D, *et al.* Fast and accurate tumor segmentation of histology images using persistent homology and deep convolutional features. Med Image Anal 2019;55:1-14.
43. Zhang Z, Chen P, McGough M, Xing F, Wang C, Bui M, *et al*, Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. Nat Mach Intell 2019;1:236-45.
44. Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016.
45. Girshick R, Donahue J, Darrell T, Malik J. Region-based convolutional networks for accurate object detection and segmentation. IEEE Trans Pattern Anal Mach Intell 2016;38:142-58.
46. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision; 2017.
47. Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell 2017;39:640-51.

48. Ronneberger O, Fischer P, Brox T. U-net: Convolutional Networks for Biomedical Image Segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention; 2015.
49. Chen LC, Lin TY, Goyal P, Girshick R, He K, Dollár P Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV); 2018.
50. Pal A, Garain U, Chandra A, Chatterjee R, Senapati S. Psoriasis skin biopsy image segmentation using Deep Convolutional Neural Network. Comput Methods Programs Biomed 2018;159:59-69.
51. Bulten W, Litjens G. Unsupervised Prostate Cancer Detection on H & E Using Convolutional Adversarial Autoencoders. arXive 2018;1804:07098.
52. Hoefling H, Sing T, Hossain I, Boisclair J, Doelemeyer A, Flandre T, Piaia A, Romanet V, Santarossa G, Saravanan C, Sutter E, Turner O, Wuersch K, Moulin P. HistoNet: A Deep Learning-Based Model of Normal Histology. Toxicol Pathol. 2021 Jun;49(4):784-797. doi: 10.1177/0192623321993425. Epub 2021 Mar 3. PMID: 33653171.
53. Ianni JD, Soans RE, Sankarapandian S, Chamarthi RV, Ayyagari D, Olsen TG, *et al.* Tailored for real-world: A whole slide image classification system validated on uncurated multi-site data emulating the prospective pathology workload. Sci Rep 2020;10:3217.
54. Naud L, Lavin A. Manifolds for Unsupervised Visual Anomaly Detection. arXiv preprint arXiv: arXive 2020;2006:11364.
55. Swiderska-Chadaj Z, de Bel T, Blanchet L, Baidoshvili A, Vossen D, van der Laak J, *et al.* Impact of rescanning and normalization on convolutional neural network performance in multi-center, whole-slide classification of prostate cancer. Sci Rep 2020;10:14398.
56. Bug D, Gräbel P, Feuerhake F, Oswald E, Schüler J, Merhof D. Supervised and Unsupervised Cell-Nuclei Detection in Immunohistology. in Proceedings of the 2nd MICCAI Workshop on Computational Pathology (COMPAY), 2019.
57. Zanjani FG, Zinger S, Bejnordi BE, van der Laak JA. Histopathology Stain-Color Normalization Using Deep Generative Models. 1st Conference on Medical Imaging with Deep Learning (MIDL), Amsterdam, The Netherlands (2018).
58. Janowczyk A, Basavanhally A, Madabhushi A. Stain Normalization using Sparse AutoEncoders (StaNoSA): Application to digital pathology. Comput Med Imaging Graph 2017;57:50-61.
59. Hou L, Nguyen V, Kanevsky AB, Samaras D, Kurc TM, Zhao T, *et al.* Sparse autoencoder for unsupervised nucleus detection and representation in histopathology images. Pattern Recognit 2019;86:188-200.
60. Sali R, Moradinasab N, Guleria S, Ehsan L, Fernandes P, Shah TU, *et al.* Deep learning for whole-slide tissue histopathology classification: A comparative study in the identification of dysplastic and non-dysplastic Barrett's esophagus. J Pers Med 2020;10:141.
61. Levy JJ, Azizgolshani N, Andersen MJ Jr., Suriawinata A, Liu X, Lisovsky M, *et al*. A large-scale internal validation study of unsupervised virtual trichrome staining technologies on nonalcoholic steatohepatitis liver biopsies. Mod Pathol 2021;34:808-22.
62. Gauthier BE, Gervais F, Hamm G, O'Shea D, Piton A, Schumacher VL. Toxicologic Pathology Forum*: Opinion on integrating innovative digital pathology tools in the regulatory framework. Toxicol Pathol 2019;47:436-43.
63. Zuraw A, Staup M, Klopfleisch R, Aeffner F, Brown D, Westerling-Bui T, *et al.* Developing a qualification and verification strategy for digital tissue image analysis in toxicological pathology. Toxicol Pathol 2021;49:773-83.
64. Aeffner F, Wilson K, Martin NT, Black JC, Hendriks CL, Bolon B, *et al.* The gold standard paradox in digital image analysis: Manual versus automated scoring as ground truth. Arch Pathol Lab Med 2017;141:1267-75.
65. Raciti P, Sue J, Ceballos R, Godrich R, Kunz JD, Kapur S, *et al.* Novel artificial intelligence system increases the detection of prostate cancer in whole slide images of core needle biopsies. Mod Pathol 2020;33:2058-66.
66. Martin DR, Hanson JA, Gullapalli RR, Schultz FA, Sethi A, Clark DP. A deep learning convolutional neural network can recognize common patterns of injury in gastric pathology. Arch Pathol Lab Med 2020;144:370-8.
67. Horai Y, Mizukawa M, Nishina H, Nishikawa S, Ono Y, Takemoto K, *et al.* Quantification of histopathological findings using a novel image analysis platform. J Toxicol Pathol 2019;32:319-27.
68. Eelbode T, Bertels J, Berman M, Vandermeulen D, Maes F, Bisschops R, *et al.* Optimization for medical image segmentation: Theory and practice when evaluating with dice score or Jaccard Index. IEEE Trans Med Imaging 2020;39:3679-90.
69. 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.
70. Kumar N, Verma R, Sharma S, Bhargava S, Vahadane A, Sethi A. A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE Trans Med Imaging 2017;36:1550-60.
71. Freyre CA, Spiegel S, Gubser Keller C, Vandemeulebroecke M, Hoefling H, Dubost V, *et al.* Biomarker-based classification and localization of renal lesions using learned representations of histology – A machine learning approach to histopathology. Toxicol Pathol 2021;49:798-814.
72. Bukowy JD, Dayton A, Cloutier D, Manis AD, Staruschenko A, Lombard JH, *et al.* Region-based convolutional neural nets for localization of glomeruli in trichrome-stained whole kidney sections. J Am Soc Nephrol 2018;29:2081-8.
73. Heinemann F, Birk G, Stierstorfer B. Deep learning enables pathologist-like scoring of NASH models. Sci Rep 2019;9:18454.
74. Asay BC, Edwards BB, Andrews J, Ramey ME, Richard JD, Podell BK, *et al.* Digital image analysis of heterogeneous tuberculosis pulmonary pathology in non-clinical animal models using deep convolutional neural networks. Sci Rep 2020;10:6047.
75. Yurttakal AH, Erbay H, Çinarer G, Baş H Classification of diabetic rat histopathology images using convolutional neural networks. Int J Comput Intell Syst 2020;14:715-22.
76. Kumar A, Singh SK, Saxena S, Lakshmanan K, Sangaiah AK, Chauhan H, *et al*, Deep feature learning for histopathological image classification of canine mammary tumors and human breast cancer. Inf Sci 2020;508:405-21.
77. Aubreville M, Bertram CA, Marzahl C, Gurtner C, Dettwiler M, Schmidt A, *et al.* Deep learning algorithms out-perform veterinary pathologists in detecting the mitotically most active tumor region. Sci Rep 2020;10:16447.
78. Zormpas-Petridis K, Noguera R, Ivankovic DK, Roxanis I, Jamin Y, Yuan Y. SuperHistopath: A deep learning pipeline for mapping tumor heterogeneity on low-resolution whole-slide digital histopathology images. Front Oncol 2020;10:586292.
79. Bigley AL, Klein SK, Davies B, Williams L, Rudmann DG. Using automated image analysis algorithms to distinguish normal, aberrant, and degenerate mitotic figures induced by Eg5 inhibition. Toxicol Pathol 2016;44:663-72.
80. Horai Y, Kakimoto T, Takemoto K, Tanaka M. Quantitative analysis of histopathological findings using image processing software. J Toxicol Pathol 2017;30:351-8.
81. Sonigo C, Jankowski S, Yoo O, Trassard O, Bousquet N, Grynberg M, *et al.* High-throughput ovarian follicle counting by an innovative deep learning approach. Sci Rep 2018;8:13499.
82. Yu Y, Wang J, Ng CW, Ma Y, Mo S, Fong EL, *et al.* Deep learning enables automated scoring of liver fibrosis stages. Sci Rep 2018;8:16016.
83. Hu F, Schutt L, Kozlowski C, Regan K, Dybdal N, Schutten MM. Ovarian toxicity assessment in histopathological images using deep learning. Toxicol Pathol 2020;48:350-61.
84. Rudmann D, Albretsen J, Doolan C, Gregson M, Dray B, Sargeant A, *et al.* Using deep learning artificial intelligence algorithms to verify N-nitroso-N-methylurea and urethane positive control proliferative changes in Tg-RasH2 mouse carcinogenicity studies. Toxicol Pathol 2021;49:938-49.
85. Ramot Y, Zandani G, Madar Z, Deshmukh S, Nyska A. Utilization of a deep learning algorithm for microscope-based fatty vacuole quantification in a fatty liver model in mice. Toxicol Pathol 2020;48:702-7.
86. Pischon H, Mason D, Lawrenz B, Blanck O, Frisk AL, Schorsch F, *et al.* Artificial intelligence in toxicologic pathology: Quantitative evaluation of compound-induced hepatocellular hypertrophy in rats. Toxicol Pathol

2021;49:928-37.
87. De Vera Mudry MC, Martin J, Schumacher V, Venugopal R. Deep learning in toxicologic pathology: A new approach to evaluate rodent retinal atrophy. Toxicol Pathol 2021;49:851-61.
88. Hvid H, Skydsgaard M, Jensen NK, Viuff BM, Jensen HE, Oleksiewicz MB, *et al.* Artificial intelligence-based quantification of epithelial proliferation in mammary glands of rats and oviducts of göttingen minipigs. Toxicol Pathol 2021;49:912-27.
89. Carboni E, Marxfeld H, Tuoken H, Klukas C, Eggers T, Gröters S, *et al*. A workflow for the performance of the differential ovarian follicle count using deep neuronal networks. Toxicol Pathol 2021;49:843-50.
90. Tokarz DA, Steinbach TJ, Lokhande A, Srivastava G, Ugalmugle R, Co CA, *et al.* Using artificial intelligence to detect, classify, and objectively score severity of rodent cardiomyopathy. Toxicol Pathol 2021;49:888-96.
91. Xu J, Lu H, Li H, Yan C, Wang X, Zang M, *et al.* Computerized spermatogenesis staging (CSS) of mouse testis sections via quantitative histomorphological analysis. Med Image Anal 2021;70:101835.
92. Creasy DM, Panchal ST, Garg R, Samanta P. Deep learning-based spermatogenic staging assessment for hematoxylin and eosin-stained sections of rat testes. Toxicol Pathol 2021;49:872-87.
93. Smith MA, Westerling-Bui T, Wilcox A, Schwartz J. Screening for bone marrow cellularity changes in cynomolgus macaques in toxicology safety studies using artificial intelligence models. Toxicol Pathol 2021;49:905-11.
94. Bédard A, Westerling-Bui T, Zuraw A. Proof of concept for a deep learning algorithm for identification and quantification of key microscopic features in the murine model of DSS-induced colitis. Toxicol Pathol 2021;49:897-904.
95. Ramot Y, Deshpande A, Morello V, Michieli P, Shlomov T, Nyska A. Microscope-based automated quantification of liver fibrosis in mice using a deep learning algorithm. Toxicol Pathol 2021;49:1126-33.
96. Kuklyte J, Fitzgerald J, Nelissen S, Wei H, Whelan A, Power A, *et al.* Evaluation of the use of single- and multi-magnification convolutional neural networks for the determination and quantitation of lesions in nonclinical pathology studies. Toxicol Pathol 2021;49:815-42.
97. National Society for Histotechnology, HistoQIP Whole Slide Image Quality Improvement Program (HQWSI). 2019 Available from: https://www.nsh.org/learn/histoqip/histoqip-specialty650. [Last Accessed 01 Feb 2021]
98. Barisoni L, Lafata KJ, Hewitt SM, Madabhushi A, Balis UG. Digital pathology and computational image analysis in nephropathology. Nat Rev Nephrol 2020;16:669-85.
99. Janowczyk A, Zuo R, Gilmore H, Feldman M, Madabhushi A. HistoQC: An open-source quality control tool for digital pathology slides. JCO Clin Cancer Inform 2019;3:1-7.
100. Gibson-Corley KN, Olivier AK, Meyerholz DK. Principles for valid histopathologic scoring in research. Vet Pathol 2013;50:1007-15.
101. Azam AS, Miligy IM, Kimani PK, Maqbool H, Hewitt K, Rajpoot NM, *et al.* Diagnostic concordance and discordance in digital pathology: A systematic review and meta-analysis. J Clin Pathol 2021;74:448-55.
102. Pantanowitz L, Bui MM. Computer-assisted pap test screening. Monogr Clin Cytol 2019;25:67-74.
103. Food and Drug Administration ThinPrep Integrated Imager - P950039/S036. May 17, 2018 Available from: https://www.fda.gov/medical-devices/recently-approved-devices/thinprep-integrated-imager-p950039s036. [Last Accessed 2021 Feb 24]
104. Food and Drug Administration Automated Cervical Cytology Screening and Imaging System Dec 3, 2008 [Last Accessed 2021 Feb 24].
105. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019;25:1301-9.
106. Litjens G, Sánchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, *et al.* Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep 2016;6:26286.
107. Steiner DF, MacDonald R, Liu Y, Truszkowski P, Hipp JD, Gammage C, *et al.* Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. Am J Surg Pathol 2018;42:1636-46.
108. Rollins-Raval MA, Raval JS, Contis L. Experience with CellaVision DM96 for peripheral blood differentials in a large multi-center academic hospital system. J Pathol Inform 2012;3:29.
109. Tvedten HW, Lilliehöök IE. Canine differential leukocyte counting with the CellaVision DM96Vision, Sysmex XT-2000iV, and Advia 2120 hematology analyzers and a manual method. Vet Clin Pathol 2011;40:324-39.
110. FirstWord MedTech New Kind of Artificial Intelligence Provides Breakthrough in Breast Cancer Diagnosis. November 20, 2020; URL: https://www.eurekalert.org/news-releases/90165. [Last Accessed 2021 Feb 24]
111. Hosseini MS, Chan L, Tse G, Tang M, Deng J, Norouzi S, *et al*, Atlas of Digital Pathology: A Generalized Hierarchical Histological Tissue Type-Annotated Database for Deep Learning. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019.
112. Bizzego A, Bussola N, Chierici M, Maggio V, Francescatto M, Cima L, *et al.* Evaluating reproducibility of AI algorithms in digital pathology with DAPPER. PLoS Comput Biol 2019;15:e1006269.
113. Schaumberg AJ, Juarez-Nicanor WC, Choudhury SJ, Pastrián LG, Pritt BS, Prieto Pozuelo M, *et al.* Interpretable multimodal deep learning for real-time pan-tissue pan-disease pathology search on social media. Mod Pathol 2020;33:2169-85.
114. Sridhar A, Doyle S, Madabhushi A. Content-based image retrieval of digitized histopathology in boosted spectrally embedded spaces. J Pathol Inform 2015;6:41.
115. Hegde N, Cai CJ. Building SMILY, a Human-Centric, Similar-Image Search Tool for Pathology. July 19, 2019. Available from: https://ai.googleblog.com/2019/07/building-smily-human-centric-similar.html. [Last Accessed 2021 Jan 10]
116. Hegde N, Hipp JD, Liu Y, Emmert-Buck M, Reif E, Smilkov D, *et al.* Similar image search for histopathology: SMILY. NPJ Digit Med 2019;2:56.
117. Marée R, Rollus L, Stévens B, Hoyoux R, Louppe G, Vandaele R, *et al.* Collaborative analysis of multi-gigapixel imaging data using cytomine. Bioinformatics 2016;32:1395-401.
118. Food and Drug Administration CFR - Code of Federal Regulations Title 21 (Parts 11 and 58). April 1, 2020. Available from: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm. [Last Accessed 2021 Feb 10]
119. Abels E, Pantanowitz L. Current state of the regulatory trajectory for whole slide imaging devices in the USA. J Pathol Inform 2017;8:23.
120. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: An online database. NPJ Digit Med 2020;3:118.
121. Businesswire. FDA Grants Breakthrough Designation to Paige. AI. March 7, 2019. Available from: https://www.businesswire.com/news/home/20190307005205/en/FDA-Grants-Breakthrough-Designation-Paige.AI. [Last Accessed 2021 Feb 25].
122. Food and Drug Administration, Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD). Discussion Paper and Request for Feedback; 2019.
123. Food and Drug Administration FDA releases Artificial Intelligence/Machine Learning Action Plan. January 12; 2021. Available from: https://www.fda.gov/news-events/press-announcements/fda-releases-artificial-intelligencemachine-learning-action-plan. [last accessed 2021 Jan 14]
124. Carvalho DV, Pereira EM, Cardoso JS. Machine learning interpretability: A survey on methods and metrics. Electronics 2019;8:832.
125. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2019;9:e1312.
126. Huang Y, Chung AC. Evidence Localization for Pathology Images Using Weakly Supervised Learning. In International conference on medical image computing and computer-assisted intervention; 2019.
127. Paschali M, Naeem MF, Simson W, Steiger K, Mollenhauer M, Navab N. Deep Learning Under the Microscope: Improving the Interpretability

of Medical Imaging Neural Networks. arXive 2019;1904:03127.
128. Tellez D, Litjens G, van der Laak J, Ciompi F. Neural image compression for gigapixel histopathology image analysis. IEEE Trans Pattern Anal Mach Intell 2021;43:567-78.
129. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-Cam: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the IEEE International Conference on Computer Vision; 2017.
130. Clunie DA. DICOM format and protocol standardization – A core requirement for digital pathology success. Toxicol Pathol 2021;49:738-49.
131. Ho DJ, Yarlagadda DV, D'Alfonso TM, Hanna MG, Grabenstetter A, Ntiamoah P, *et al.* Deep Multi-Magnification Networks for multi-class breast cancer image segmentation. Comput Med Imaging Graph 2021;88:101866.
132. Foucart A, Debeir O, Decaestecker C. SNOW: Semi-Supervised, NOisy and/or Weak Data for Deep Learning in Digital Pathology. In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE; 2019.
133. Sirinukunwattana K, Raza SE, Tsang YW, Snead DR, Cree IA, Rajpoot NM. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE Trans Med Imaging 2016;35:1196-206.
134. Tellez D, Litjens G, Bándi P, Bulten W, Bokhorst JM, Ciompi F, *et al.* Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Med Image Anal 2019;58:101544.
135. Bankhead P, Loughrey MB, Fernández JA, Dombrowski Y, McArt DG, Dunne PD, *et al.* QuPath: Open source software for digital pathology image analysis. Sci Rep 2017;7:16878.
136. Brunyé TT, Carney PA, Allison KH, Shapiro LG, Weaver DL, Elmore JG. Eye movements as an index of pathologist visual expertise: A pilot study. PLoS One 2014;9:e103447.
137. Raghunath V, Braxton MO, Gagnon SA, Brunyé TT, Allison KH, Reisch LM, *et al*. Mouse cursor movement and eye tracking data as an indicator of pathologists' attention when viewing digital whole slide images. J Pathol Inform 2012;3:43.
138. He Y, Wei J, Che S, Liu S, Luo P. Computer-Aided Pathological Annotation Framework: A Deep Learning-Based Diagnostic Algorithm of Lung Cancer. In 2019 International Conference on Information Technology and Computer Application (ITCA). IEEE; 2019.
139. Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw 2018;106:249-59.
140. Taqi SA, Sami SA, Sami LB, Zaki SA. A review of artifacts in histopathology. J Oral Maxillofac Pathol 2018;22:279.
141. Chlipala EA, Butters M, Brous M, Fortin JS, Archuletta R, Copeland K, *et al.* Impact of preanalytical factors during histology processing on section suitability for digital image analysis. Toxicol Pathol 2021;49:755-72.
142. Zarella MD, Breen DE, Plagov A, Garcia FU. An optimized color transformation for the analysis of digital images of hematoxylin and eosin stained slides. J Pathol Inform 2015;6:33.
143. Lo CM, Wu YH, Li YC, Lee CC. Computer-aided bacillus detection in whole-slide pathological images using a deep convolutional neural network. NATO Adv Sci Inst Ser E Appl Sci 2020;10:4059.