Deep neural network models for computational histopathology: A survey

Chetan L. Srinidhi[a,b,*], Ozan Ciga[b], Anne L. Martel[a,b]

[a] *Physical Sciences, Sunnybrook Research Institute, Toronto, Canada*
[b] *Department of Medical Biophysics, University of Toronto, Canada*

Abstract

Histopathological images contain rich phenotypic information that can be used to monitor underlying mechanisms contributing to disease progression and patient survival outcomes. Recently, deep learning has become the mainstream methodological choice for analyzing and interpreting histology images. In this paper, we present a comprehensive review of state-of-the-art deep learning approaches that have been used in the context of histopathological image analysis. From the survey of over 130 papers, we review the field's progress based on the methodological aspect of different machine learning strategies such as supervised, weakly supervised, unsupervised, transfer learning and various other sub-variants of these methods. We also provide an overview of deep learning based survival models that are applicable for disease-specific prognosis tasks. Finally, we summarize several existing open datasets and highlight critical challenges and limitations with current deep learning approaches, along with possible avenues for future research.

Keywords: Deep Learning, Convolutional Neural Networks, Computational Histopathology, Digital Pathology, Histology Image Analysis, Survey, Review.

1. Introduction

The examination and interpretation of tissue sections stained with haematoxylin and eosin (H&E) by anatomic pathologists is an essential component in the assessment of disease. In addition to providing diagnostic information, the phenotypic information contained in histology slides can be used for prognosis. Features such as nuclear atypia, degree of gland formation, presence of mitosis and inflammation can all be indicative of how aggressive a tumour is, and may also allow predictions to be made about the likelihood of recurrence after surgery. Over the last 50 years, several scoring systems have been proposed that allow pathologists to grade tumours based on their appearance, for example, the Gleason score for prostate cancer (Epstein et al., 2005) and the Nottingham score for breast cancer (Rakha et al., 2008). These systems provide important information to guide decisions about treatment and are valuable in assessing heterogeneous disease. There is, however, considerable inter-pathologist variability, and some systems that require quantitative analysis, for example the residual cancer burden index (Symmans et al., 2007), are too time-consuming to use in a routine clinical setting.

The first efforts to extract quantitative measures from microscopy images were in cytology. Prewitt and Mendelsohn (1966) laid out the steps required for the "effective and efficient discrimination and interpretation of images" which described the basic paradigm of object detection, feature extraction and finally the training of a classification function that is still in use more than 50 years later. Early work in cytology and histopathology was usually limited to the analysis of the small fields of view that could be captured using conventional mi-

[*] Corresponding author:
E-mail address: chetan.srinidhi@utoronto.ca (Chetan L. Srinidhi)
Published in Medical Image Analysis, Vol. 67, Jan 2021, Elsevier (https://doi.org/10.1016/j.media.2020.101813)

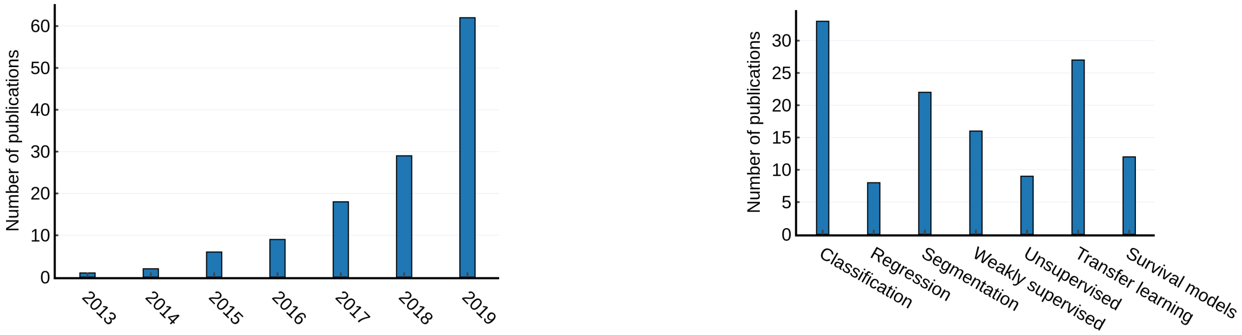
croscopy, and image acquisition was a time-consuming process (Mukhopadhyay et al., 2018). The introduction of whole slide scanners in the 1990s made it much easier to produce digitized images of whole tissue slides at microscopic resolution, and this led to renewed interest in the application of image analysis and machine learning techniques to histopathology. Many of the algorithms developed originally for computer-aided diagnosis in radiology have been successfully adapted for use in digital pathology, and Gurcan et al. (2009); Madabhushi and Lee (2016) provide comprehensive reviews of work carried out prior to the widespread adoption of deep learning methods.

In 2011, Beck et al. (2011) demonstrated that features extracted from histology images could aid in the discovery of new biological aspects of cancer tissue, and Yuan et al. (2012) showed that features extracted from digital pathology images are complementary to genomic data. These advancements have led to a growing interest in the use of biomarkers extracted from digital pathology images for precision medicine (Bera et al., 2019), particularly in oncology. Later in 2012, Krizhevsky et al. (2012) showed that convolutional neural networks (CNNs) could outperform previous machine learning approaches by classifying 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into 1000 different classes. At the same time, Cireşan et al. (2012) showed that CNNs could outperform competing methods in segmenting nerves in electron microscopy images and detecting mitotic cells in histopathology images (Cireşan et al., 2013). Since then, methods based on CNNs have consistently outperformed other handcrafted methods in a variety of deep learning (DL) tasks in digital pathology. The ability of CNNs to learn features directly from the raw data without the need for specialist input from pathologists and the availability of annotated histopathology datasets has also fueled the explosion of interest in deep learning applied to histopathology.

The analysis of whole-slide digital pathology images (WSIs) poses some unique challenges. The images are very large and have to be broken down into hundreds or thousands of smaller tiles before they can be processed. Both the context at low magnification, and the detail at high magnification, may be important for a task, therefore information from multiple scales needs to be integrated. In the case of survival prediction, salient regions of the image are not known *a priori* and we may only have weak slide level labels. The variability within each disease subtype can be high and it usually requires a highly trained pathologist to make annotations. For cell based methods, many thousands of objects need to be detected and characterized. These challenges have made it necessary to adapt existing deep learning architectures and to design novel approaches specific to the digital pathology domain. In this work, we surveyed more than 130 papers, where deep learning has been applied to a wide variety of detection, diagnosis, prediction and prognosis tasks. We carried out this extensive review by searching Google Scholar, PubMed and arXiv for papers containing keywords such as ("convolutional" or "deep learning") and ("digital pathology" or "histopathology" or "computational pathology"). Additionally, we also included conference proceedings from MICCAI, ISBI, MIDL, SPIE and EMBC based on title/abstract of the papers. We also iterated over the selected papers to include any additional cross-referenced works that were missing from our initial search criteria. The body of research in this area is growing rapidly and this survey covers the period up to and including December 2019. A descriptive statistics of published papers according to their category and year is illustrated in Fig. 1. The remainder of this paper is organised as follows. Section 2 presents an overview of various learning schemes in DL literature in the context of computational histopathology. Section 3 discusses in detail different categories of DL schemes commonly used in this field. We categorize these learning mechanisms into supervised (Section 3.1), weakly supervised (Section 3.2), unsupervised (Section 3.3), transfer learning (Section 3.4). Section 4 discusses survival models related to disease prognosis task. In Section 5, we discuss various open challenges including prospective applications and future trends in computational pathology, and finally, conclusions are presented in Section 6.

2. Overview of learning schemas

In this section, we provide a formal introduction to various learning schemes in the context of DL applied to computational pathology. These learning schemes are illustrated with an example of classifying a histology WSI as cancerous or normal. Based on these formulations, various DL models have been proposed in the literature,



(a) (b)

Figure 1: (a) An overview of numbers of papers published from January 2013 to December 2019 in deep learning based computation histopathology surveyed in this paper. (b) A categorical breakdown of the number of papers published in each learning schemas.

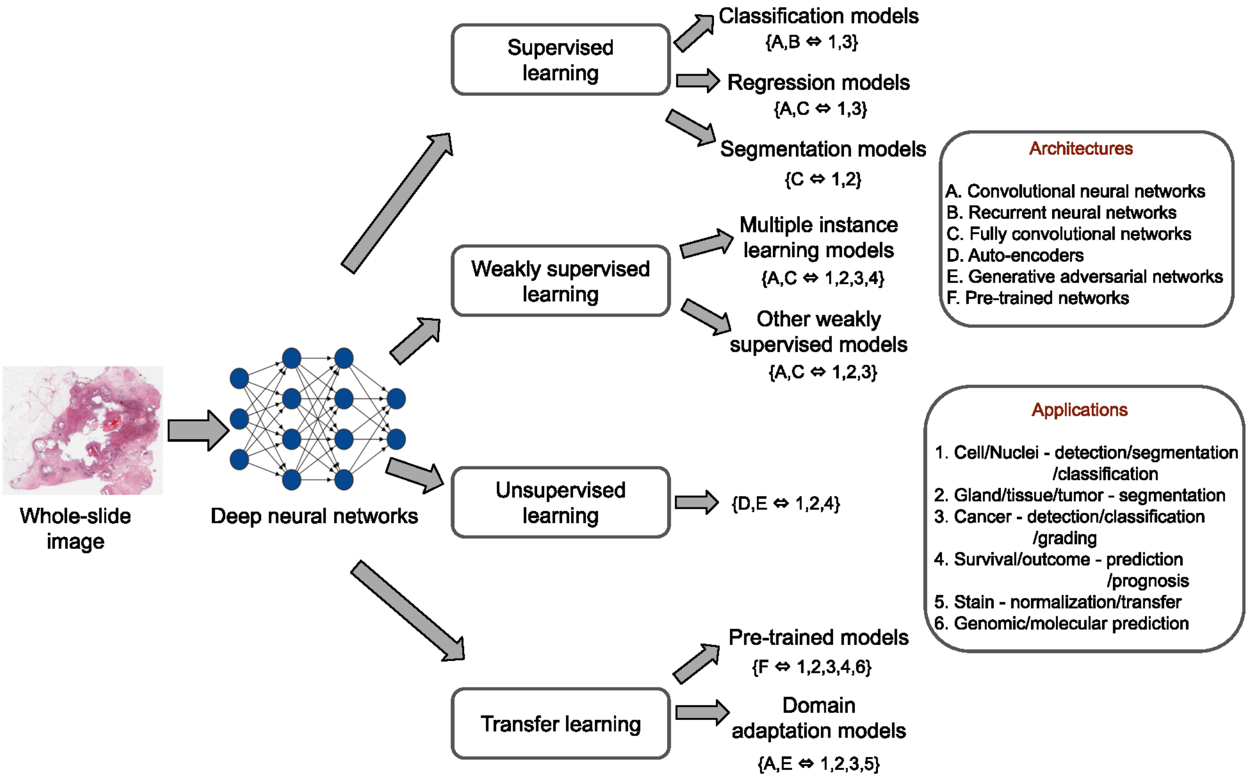

Whole-slide image
Deep neural networks
Supervised learning
Classification models
{A,B ⇔ 1,3}
Regression models
{A,C ⇔ 1,3}
Segmentation models
{C ⇔ 1,2}
Weakly supervised learning
Multiple instance learning models
{A,C ⇔ 1,2,3,4}
Other weakly supervised models
{A,C ⇔ 1,2,3}
Unsupervised learning
{D,E ⇔ 1,2,4}
Transfer learning
Pre-trained models
{F ⇔ 1,2,3,4,6}
Domain adaptation models
{A,E ⇔ 1,2,3,5}
Architectures
A. Convolutional neural networks
B. Recurrent neural networks
C. Fully convolutional networks
D. Auto-encoders
E. Generative adversarial networks
F. Pre-trained networks
Applications
1. Cell/Nuclei - detection/segmentation /classification
2. Gland/tissue/tumor - segmentation
3. Cancer - detection/classification /grading
4. Survival/outcome - prediction /prognosis
5. Stain - normalization/transfer
6. Genomic/molecular prediction

Figure 2: An overview of deep neural network models in computational histopathology. These models have been constructed using various deep learning architectures (shown in *alphabetical* order) and applied to various histopathological image analysis tasks (depicted in *numerical* order).

which are traditionally based on convolutional neural network (CNNs), recurrent neural networks (RNNs), generative adversarial networks (GANs), auto-encoders (AEs) and various other variants. For a detailed and thorough background of DL fundamentals and its existing architectures, we refer readers to LeCun et al. (2015); Goodfellow et al. (2016), and with specific application of DL in medical image analysis to Litjens et al. (2017); Shen et al. (2017); Yi et al. (2019).

In *supervised learning*, we have a set of N training examples $\{(x_i, y_i)\}_{i=1}^{N}$, where, each sample $x_i \in \mathbb{R}^{C_h \times H \times W}$ is an input image (a WSI of dimension $H \times W$ pixels, with C_h channels. For example, $C_h = 3$ channels for an RGB image) associated with a class label $y_i = \mathbb{R}^C$, with C possible classes. For example, in binary classification, C takes the scalar form $\{0, 1\}$, and the set $\mathbb{R}$ for a regression task. The goal is to train a model $f_\theta : x \to y$ that best predicts the label for an unknown test image based on a loss function $\mathcal{L}$. For instance, x's are the patches in WSIs and y's are the labels annotated by the pathologist either as cancerous or normal. During the inference time, the model predicts the label of a patch from a previously unseen test set. This scheme is detailed in Section 3.1, with an example illustrated in Fig. 3.

In *weakly supervised learning* (WSL), the goal is to train a model f_θ using the readily available coarse-grained (image-level) annotations C_i, to automatically infer the fine-grained (pixel/patch)-level labels c_i. In histopathology, a pathologist labels a WSI as cancer, as long as a small part of this image contains cancerous region, without indicating its exact location. Such image-level annotations (often called *"weak labels"*) are relatively easier to obtain in practice compared to expensive pixel-wise labels for supervised methods. An illustrative example for WSL scheme is shown in Fig. 4, and this scheme is covered in-depth in Section 3.2.

The *unsupervised learning* aims at identifying patterns on the image, without mapping an input image sample into a predefined set of output (i.e. label). This type of models includes fully unsupervised methods, where the raw data comes in the form of images without any expert-annotated labels. A common technique in unsupervised learning is to transform the input data into a lower-dimensional subspace, and then group these lower-dimension representations (i.e. the latent vector) into mutually exclusive or hierarchical groups, based on a clustering technique. An example of unsupervised learning scheme is illustrated in Fig. 5, with existing methods in Section 3.3.

In *transfer learning* (TL), the goal is to transfer knowledge from one domain (i.e., source) to another domain (i.e., target), by relaxing the assumption that the train and test set must be independent and identically distributed. Formally, given a domain $\mathcal{D} = \{\mathcal{X}, P(X)\}$, which is defined by the feature space $\mathcal{X}$, a marginal probability distribution $P(X)$ (where $X = \{x_1, \ldots, x_n\} \in \mathcal{X}$), and a task $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$ - consisting of label space $\mathcal{Y}$ and a prediction function $f(\cdot)$. The aim of transfer learning is to improve the predictive function $f^{\mathcal{T}}(\cdot)$ in target domain ($\mathcal{D}^t$) by using the knowledge in source domain ($\mathcal{D}^s$) and source task ($\mathcal{T}^s$). For example, in histology, this scenario can occur while training a classifier on the source task $\mathcal{T}^s$ and possibly fine-tuning on a target task $\mathcal{T}^t$, with limited or no annotations. This scheme is explained in-detail in Section 3.4. Note that, the *domain adaptation*, which is a sub-field of transfer learning, is discussed thoroughly in Section 3.4.1.

Next, we discuss various deep neural network (DNN) models in each of these learning schemes published in histopathology domain, along with the existing challenges and gaps in current research, and possible future directions in this perspective.

3. Methodological approaches

The aim of this section is to provide a general reference guide to various deep learning models applied in computational histopathology from a methodological perspective. The DL models discussed in the following sections were originally developed for specific applications, but are applicable to a wide variety of histopathological tasks (Fig. 2). Based on the learning schemes, the following sections are divided into supervised, weakly supervised, unsupervised and transfer learning approaches. The details are presented next.

3.1. Supervised learning

Among the supervised learning techniques, we identify three major canonical deep learning models based on the nature of tasks that are solved in digital histopathology: classification, regression and segmentation based models,

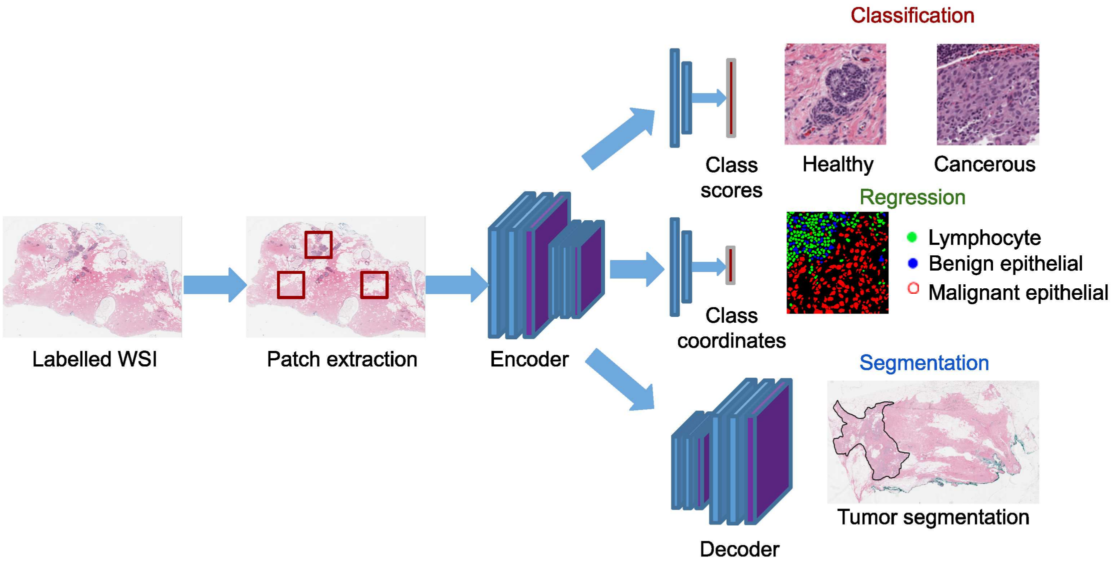


Figure 3: An overview of supervised learning models.

as illustrated in Fig. 3. The first category of models contains methods related to pixel-wise/sliding-window classification based approaches, which are traditionally formulated as object detection (Girshick, 2015) or image classification tasks (He et al., 2016) in the computer vision literature. The second category of models focuses on predicting the position of objects (e.g., cells or nuclei (Sirinukunwattana et al., 2016)) or sometimes predicting a cancer severity score (e.g., H-score of breast cancer images (Liu et al., 2019)) by enforcing topological/spatial constraints in DNN models. Finally, the last category of models is related to fully convolutional network (FCN) based approaches (Long et al., 2015; Ronneberger et al., 2015), which are widely adopted to solve semantic or instance segmentation problems in computer vision and medical imaging scenarios. The overview of papers in supervised learning is summarized in Table 1.

3.1.1. Classification models

This category of methods uses a sliding window approach (i.e., patch centred on a pixel of interest) to identify objects (such as cells, glands, nuclei) or make image-level predictions (such as disease diagnosis and prognosis). Within this category, we further identify two subcategories: (i) *local-level tasks*, and (ii) *global-level tasks*. The former stream of methods is based on a region (i.e., cell, nuclei) represented by a spatially pooled feature representations or scores, aiming at identifying or localizing objects. While the latter consists of methods related to image-level prediction tasks such as whole-slide level disease grading.

A. Local-level task: Image classification such as detection of cells or nuclei is notably one of the most successful tasks, where deep learning techniques have made a tremendous contribution in the field of digital pathology. Methods based on CNNs have been extensively used for pixel-wise prediction task by a sliding window approach, to train the networks on small image patches rather than the entire WSI. Due to giga-resolution of WSIs (e.g., $100,000 \times 100,000$ pixels), applying a CNN directly to WSI is impractical, and hence, the entire WSI is divided into segments of small patches for analysis. In practice, these image patches are often annotated by the pathologist as a region containing an object of interest (e.g., cells/nuclei) or a background. A large corpus of deep learning methods applied to digital pathology is akin to computer vision models applied to visual object recognition task (Russakovsky et al., 2015).

The earliest seminal work proposed in Cireşan et al. (2013) revolutionised the entire field of digital histopathology, by applying CNN based pixel prediction to detect mitosis in routinely stained H&E breast cancer histology images. Subsequently, Rao (2018) proposed a variant of the Faster-RCNN, which significantly outperformed all other competing techniques in both ICPR 2012 and the AMIDA 2013 mitosis detection challenge. The next set of methods were developed based on CNNs or a combination of CNN and handcrafted features. Since training of CNN models is often complex and requires a more extensive training set, the earliest works (Wang et al., 2014; Kashif et al., 2016; Xing et al., 2016; Romo-Bucheli et al., 2016; Wang et al., 2016b) focused on integrating CNN with biologically interpretable handcrafted features and these models showed excellent performance results in addressing the touching nuclei segmentation problem. A hybrid method, based on persistent homology, Qaiser et al. (2019b) was able to capture the degree of spatial connectivity among touching nuclei, which is quite difficult to achieve using CNN models (Sabour et al., 2017).

Training a deep CNN from scratch requires large amounts of annotated data, which is very expensive and cumbersome to obtain in practice. A promising alternative is to use a pre-trained network (trained on a vast set of natural images, such as ImageNet) to fine-tune on a problem in different domain with limited number of annotations. Along these lines, Gao et al. (2017); Valkonen et al. (2019) proposed a fine-tuning based transfer learning approach, which consistently performed better than full training on a single dataset alone. In particular, Gao et al. (2017) made several interesting observations about improving CNN performance by optimising the hyper-parameters of the network, augmenting the training data and fine-tuning rather than full training of the model. For more details and methods that are based on transfer learning are discussed thoroughly in Section 3.4.

Recent studies (Albarqouni et al., 2016; Irshad et al., 2017; Amgad et al., 2019; Marzahl et al., 2019; Hou et al., 2020) have investigated the use of crowdsourcing approaches to alleviate the annotation burden on expert pathologists. Outsourcing labelling to non-experts can, however, lead to subjective and inconsistent labels and conventional DL models may find it challenging to train with noisy annotations. One way to improve the annotation quality is to collect multiple redundant labels per example and aggregate them via various voting techniques before training the model. For instance, Albarqouni et al. (2016) proposed to incorporate data aggregation as a part of the CNN learning process through an additional crowd-sourcing layer for improving model performance. An alternative approach is to make use of expert advice (such as an experienced pathologist) by providing feedback for annotating rare and challenging cases (Amgad et al., 2019). It is evident from the above studies that it is possible to train models using non-expert annotations successfully, but that care has to be taken to ensure quality. An easy and reliable way to obtain crowd labels for a large-scale database is to first obtain a set of precomputed annotations from an automated system, and correct only those labels with inconsistent markings under expert supervision (Marzahl et al., 2019; Hou et al., 2020). For example, Marzahl et al. (2019) showed the use of precomputed labels lead to an increase in model performance, which was independent of the annotator's expertise and that this reduced the interaction time by more than 30% compared to other crowdsourcing approaches. A thorough and in-depth treatment of crowdsourcing methods applicable to medical imaging (including histopathology) is provided in Ørting et al. (2019).

The addition of multi-scale and contextual knowledge into CNN plays an essential role in identifying overlapping cell structures in histopathology images. Conventional single scale models often suffer from two main limitations: 1) the raw-pixel intensity information around a small window does not have enough information about the degree of overlap between cells, and 2) use of a large window leads to an increase in the number of model parameters and training time. To alleviate these issues, several authors (Song et al., 2015, 2017) proposed a multi-scale CNN model to accurately solve the overlapping cell segmentation problem, with the addition of domain-specific shape priors during training. Despite several modifications to CNN architectures, however, traditional deep learning methods often lack generalisation ability due to stain variations across datasets and this is addressed in Section 3.4.2

In summary, among the bottom-up approaches, CNN is the current gold standard technique applied to a wide variety of low-level histopathology tasks such as cell or nuclei detection. Methods based on multi-scale CNN and transfer learning approaches are becoming increasingly popular due to their excellent generalization adaptability across a wide range of datasets and scanning protocols.

B. Global-level task: Most of the published deep learning methods in this category focus on patch-based classification approach for whole-slide level disease prediction task. These techniques range from the use of simple CNN architectures (Cruz-Roa et al., 2014; Ertosun and Rubin, 2015) to more sophisticated models (Qaiser and Rajpoot, 2019; Zhang et al., 2019) for accurate tissue-level cancer localization and WSI-level disease grading. For instance, Cruz-Roa et al. (2014, 2017) proposed a simple 3-layer CNN for identifying invasive ductal carcinoma in breast cancer images which outperformed all the previous handcrafted methods by a margin of 5%, in terms of average sensitivity and specificity. The main disadvantage of these methods is the relatively long computational time required to carry out a dense patch-wise prediction over an entire WSI. To address this issue, Cruz-Roa et al. (2018) proposed a combination of CNN and adaptive sampling based on quasi-Monte Carlo sampling and a gradient-based adaptive strategy, to precisely focus only on those regions with high-uncertainty. Subsequently, a few authors (Litjens et al., 2016; Vandenberghe et al., 2017) employed a simpler patch-based CNN model for the identification of breast and prostate cancer in WSI, achieving an AUC of 0.99 for the breast cancer experiment. In more recent years, some authors (Bejnordi et al., 2018; Wei et al., 2019; Nagpal et al., 2019; Shaban et al., 2019a; Halicek et al., 2019) have trained networks from scratch (i.e., full training) on huge set of WSIs. These networks include the most popular deep learning models traditionally used for natural image classification task such as *VGGNet* (Simonyan and Zisserman, 2014), *InceptionNet* (Szegedy et al., 2015), *ResNet* (He et al., 2016) and *MobileNet* (Howard et al., 2017) architectures. There is no generic rule about the choice of architectures, with the type of disease prediction task. However, the main success of these CNN models depends on the number of images available for training, choice of network hyper-parameters and various other boosting techniques (Cireşan et al., 2013; Nagpal et al., 2019) (Refer to Section 5.1 for more details).

A few authors try to encode both local and global contextual information into CNN learning process for more accurate disease prediction in WSIs. Typically, contextual knowledge is incorporated into a CNN framework by modelling the spatial correlations between neighbouring patches, using the strengths of CNNs and conditional random fields (CRF) (Zheng et al., 2015; Chen et al., 2017b). These techniques have been extensively used in computer vision tasks for sequence labeling (Artieres et al., 2010; Peng et al., 2009) and semantic image segmentation problems (Chen et al., 2017b). While in digital pathology, for instance, Kong et al. (2017) introduced a spatially structured network (Spatio-Net) combining CNN with 2D Long-short Term Memory (LSTM) to jointly learn the image appearance and spatial dependency features for breast cancer metastasis detection. A similar approach has also been adopted in Agarwalla et al. (2017) to aggregate features from neighbouring patches using 2D-LSTM's on WSIs. In contrast, Li and Ping (2018) proposed an alternative technique based on CRF for modelling spatial correlations through a fully connected CRF component. The advantages of such models are that the whole DNN can be trained in an end-to-end manner with the standard backpropagation algorithm, with a slight overhead in complexity. Alternative methods have also been proposed to encode global contextual knowledge by adopting different patch level aggregation strategies. For example, Bejnordi et al. (2017) employed a cascaded CNN model to aggregate patch-level pyramid representations to simultaneously encode multi-scale and contextual information for breast cancer multi-classification. Similarly, Awan et al. (2018) adopted a ResNet based patch classification model to output a high dimensional feature space. These features are then combined using a support vector machine (SVM) classifier to learn the context of a large patch, for discriminating different classes in breast cancer.

Although the above methods include contextual information in the form of patch-based approaches, they still suffer from loss of visual context due to disjoint/random selection of small image patches. Furthermore, applying a CNN based classification model directly to WSI is computationally expensive, and it scales linearly with an increasing number of input image patches (Qaiser and Rajpoot, 2019). Some recent studies (Qaiser and Rajpoot, 2019; BenTaieb and Hamarneh, 2018; Xu et al., 2019) explored task-driven *visual attention* models (Mnih et al., 2014; Ranzato, 2014) for histopathology WSI analysis.

Such models selectively focus on the most diagnostically useful areas (such as tissue components) while ignoring the irrelevant regions (such as the background) for further analysis. These kinds of visual attention models have been extensively explored in computer vision applications including object detection (Liu et al., 2016), image classification (Mnih et al., 2014), image captioning (Sharma et al., 2015), and action recognition (Xu et al., 2015b) tasks.

In routine clinical diagnosis, typically, a pathologist first examines different locations within a WSI to identify diagnostically indicative areas, and then combines this information over time across different eye fixations, to predict the presence or absence of cancer. This human visual attention mechanism can be modelled as a *sequential learning* task in deep learning using RNNs. For instance, Qaiser and Rajpoot (2019) modelled the prediction of immunohistochemical (IHC) scoring of HER2 (Qaiser et al., 2018) as a sequential learning problem, where the whole DNN is optimized via policy gradients trained under a deep reinforcement learning (DRL) framework. Furthermore, the authors also incorporated an additional task-specific mechanism to inhibit the model from revisiting the previously attended locations for further diagnosis. Similarly, BenTaieb and Hamarneh (2018); Xu et al. (2019) proposed recurrent attention mechanisms to selectively attend and classify the most discriminate regions in WSI for breast cancer prediction. Inspired by recent works (Xu et al., 2015b; Krause et al., 2017) in image captioning for natural scenes, Zhang et al. (2019) proposed an attention-based multi-modal DL framework to automatically generate clinical diagnostic descriptions and tissue localization attention maps, mimicking the pathologist. An attractive feature of their system is the ability to create natural language descriptions of the histopathology findings, whose structure closely resembles that of a standard clinical pathology report.

In essence, attention-based models are gaining popularity in recent years and have several intriguing properties over traditional sliding-window (patch-based) approaches: i) by enforcing a region selection mechanism (i.e., attention), the model tries to learn only the most relevant diagnostically useful areas for disease prediction; ii) the number of model parameters is drastically reduced leading to faster inference time; and iii) the model complexity is independent of the size of WSI.

3.1.2. Regression models

This category of methods focuses on detection or localization of objects by directly regressing the likelihood of a pixel being the centre of an object (e.g., cell or nucleus centre). Detection of cells or nuclei in histopathology images is challenging due to their highly irregular appearance and their tendency to occur as overlapping clumps, which results in difficulty in separating them as a single cell or a nucleus (Naylor et al., 2018; Xie et al., 2018b; Graham et al., 2019b). The use of pixel-based classification approaches for this task may result in suboptimal performance, as they do not necessarily consider the topological relationship between pixels that lie in the object centre with those in their neighbourhood (Sirinukunwattana et al., 2016). To tackle this issue, many authors cast the object detection task as a *regression* problem, by enforcing topological constraints, such that the pixels near object centres have higher probability values than those further away. This formulation has shown to achieve better detection or localization of objects, even with significant variability in both the object appearance and their locations in images.

Deep regression models proposed in the literature are mainly based on either CNN or FCN architectures (Long et al., 2015). In the context of FCN, the earlier methods by Chen et al. (2016a); Xie et al. (2018a) proposed a simple FCN based regression model for detecting cells in histopathology images. The most recent methods attempt to improve the detection task by modifying the loss function (Xie et al., 2018b) or incorporating additional features into popular deep learning architectures (Graham et al., 2019b). Xie et al. (2015a, 2018b) proposed a structured regression model based on fully residual convolutional networks for detecting cells. The authors adopted a weighted MSE loss by assigning higher weights to misclassified pixels that are closer to cell centres. A similar approach by Xing et al. (2019), adopted a residual learning based FCN architecture for simultaneous nucleus detection and classification in pancreatic neuroendocrine tumour Ki-67 images. In their model, an additional auxiliary task (i.e., ROI extraction) is also introduced to assist and boost the nucleus classification task using weak annotations. To solve the challenging touching nuclei segmentation problem, Naylor et al. (2018) proposed a model to identify superior markers for the watershed algorithm by regressing the intra-nuclear distance map. Graham et al. (2019b) went one step further, proposing a unified FCN model for simultaneous nuclear instance segmentation and classification which effectively encodes both the horizontal and vertical distance information of nuclei pixels to their centre of mass for accurate nuclei separation in multi-tissue histology images.

Other authors adopted alternative methods by modifying the output layer of CNN, to include distance constraints or a voting mechanism into the network learning process. For instance, Sirinukunwattana et al. (2016) introduced a new layer modifying the output of a CNN to predict a probability map which is topologically constrained, such that the high confidence scores are likely to be assigned to the pixels closer to nuclei centre in colon histology images. This method was later extended in Swiderska-Chadaj et al. (2019) to detect lymphocytes in immunohistochemistry images. Xie et al. (2015a) proposed an alternative method based on the voting mechanism for nuclei localization. This can be viewed as an implicit Hough-voting codebook, which learns to map an image patch to a set of voting offsets (i.e., nuclei positions) and the corresponding confidence scores to weight each vote. This set of weighted votes is then aggregated to estimate the final density map used to localize the nuclei positions in neuroendocrine tumour images.

3.1.3. Segmentation models

Segmentation of histological primitives such as cells, glands, nuclei and other tissue components is an essential pre-requisite for obtaining reliable morphological measurements to assess the malignancy of several carcinomas (Chen et al., 2017a; Sirinukunwattana et al., 2017; Bulten et al., 2020). Accurate segmentation of structures from histology images often requires the pixel-level delineation of object contour or the whole interior of the object of interest. CNNs trained to classify each patch centred on a pixel of interest as either foreground or background, can be used for segmentation tasks by employing a sliding-window approach. However, given the large size of gigapixel WSIs, patch-based approaches lead to a large number of redundant computations in overlapping regions, in turn resulting in a drastic increase in computational complexity and loss of contextual information (Chen et al., 2017a; Lin et al., 2019). The other alternative is to employ fully convolutional networks (FCN) (Long et al., 2015; Ronneberger et al., 2015), which take as input an arbitrary sized image (or a patch) and output a similar-sized image in a single forward pass. The whole FCN model can be trained via end-to-end backpropagation and directly outputs a dense per-pixel prediction score map. Hence, segmentation models in histopathology are mainly built on the representative power of FCN and its variants, which are generally formulated as a *semantic segmentation* task, with applications ranging from nucleus/gland/duct segmentation (Kumar et al., 2019; Sirinukunwattana et al., 2017; Seth et al., 2019) to the prediction of cancer (Liu et al., 2019; Bulten et al., 2019) in WSIs.

In order to determine an optimal model suitable for a given task, Swiderska-Chadaj et al. (2019); de Bel et al. (2018) compared FCN with UNet architecture (Ronneberger et al., 2015) and found that better generalization ability and robustness was achieved using a UNet model. The key feature of the UNet is the upsampling path of the network, which learns to propagate the contextual information to high-resolution layers, along with additional skip connections to yield more biologically plausible segmentation maps, compared to the standard FCN model. The traditional FCN model also lacks smoothness constraints, which can result in poor delineation of object contours and formation of spurious regions while segmenting touching/overlapping objects (Zheng et al., 2015). To circumvent this problem, BenTaieb and Hamarneh (2016) formulated a new loss function to incorporate boundary smoothness and topological priors into FCN learning, for discriminating epithelial glands with other tissue structures in histology images.

The appearance of histological objects such as glands and nuclei vary significantly in their size, shape and often occur as overlapping clumped instances, which makes them difficult to distinguish with the other surrounding structures. A few methods attempted to address this issue by leveraging the representation power of FCN with multi-scale feature learning strategies (Chen et al., 2017b; Lin et al., 2017); to effectively delineate varying size objects in histology images. For instance, Chen et al. (2017a) proposed a multi-level contextual FCN with auxiliary supervision mechanism (Xie and Tu, 2015) to segment both glands and nuclei in histology images. They also devised an elegant multi-task framework to integrate object appearance with contour information, for precise identification of touching glands. This work was later extended in Van Eycke et al. (2018) by combining the efficient techniques of DCAN (Chen et al., 2017a), UNet, and identity mapping in ResNet to build an FCN model for segmenting epithelial glands in double-stained images.

Some authors have proposed variants of FCN to enhance segmentation - in particular at glandular boundaries, by compensating for the loss occurring in max-pooling layers of FCNs. For example, Graham et al. (2019a) introduced minimum information loss dilated units in residual FCNs, to help retain the maximal spatial resolution critical for segmenting glandular structures at boundary locations. Later, Ding et al. (2019) employed a similar technique to circumvent the loss of global information by introducing a high-resolution auxiliary branch in the multi-scale FCN model, to locate and shape the glandular objects. Zhao et al. (2019) proposed a feature pyramid based model (Lin et al., 2017) to aggregate local-to-global features in FCN, to enhance the discriminative capability of the model in identifying breast cancer metastasis. Moreover, they also devised a synergistic learning approach to collaboratively train both the primary detector and an extra decoder with semantic guidance, to help improve the model's ability to retrieve metastasis.

Conventional FCN based models are fundamentally designed to predict the class label for each pixel as either foreground or background, but are unable to predict the individual object instances (i.e., recognizing the categorical label of foreground pixels). In computer vision, such problems can be formulated as an *"instance-aware semantic segmentation"* task (Hariharan et al., 2014; Li et al., 2017), where segmentation and classification of object instances are performed simultaneously in a joint end-to-end manner. In histology, Xu et al. (2017) formulated the gland instance segmentation as two sub-tasks - gland segmentation and instance recognition task, using a multi-channel deep network model (Dai et al., 2016). The gland segmentation is performed using FCN, while, the gland instance boundaries are recognized using the location (Girshick, 2015) and boundary cues (Xie and Tu, 2015). A similar formulation has been adopted in Qu et al. (2019) to solve the joint segmentation and classification of nuclei using an FCN trained with perceptual loss (Johnson et al., 2016).

Most deep learning methods in digital pathology are applied on small-sized image patches rather than the entire WSI, restricting the prediction ability of the model to a narrow field-of-view. The conventional patch-based approaches often suffer from three main limitations: i) the extracted individual patches from WSI have a narrow field-of-view, with limited contextual knowledge about the surrounding structures; ii) patch-based models are not consistent with the way a pathologist analyzes a slide under a microscope; and iii) a large number of redundant computations are carried out in overlapping regions, resulting in increased computational complexity and slower inference speed. In order to alleviate the first two issues, attempts have been made to mimic the way in which a pathologist usually analyzes a slide at various magnification levels before arriving at the final decision. Such mechanisms are integrated into the FCN model by designing multi-magnification networks (Ho et al., 2019; Tokunaga et al., 2019), each trained on different field-of-view image patches to obtain a better discriminative feature representation compared to a single-magnification model. For instance, Ho et al. (2019) proposed a multi-encoder and multi-decoder FCN model utilizing multiple input patches at various magnification levels (e.g., 20x, 10x and 5x) to obtain intermediate feature representations that are shared among each FCN model for accurate breast cancer image segmentation. A similar approach has been adopted in Tokunaga et al. (2019); Gecer et al. (2018) by training multiple FCN's on different field-of-view images, which are aggregated to obtain a final segmentation map. In contrast, Gu et al. (2018) designed a multiple encoder model to aggregate information across different magnification levels, but utilized only one decoder to generate a final prediction map.

Nevertheless, the above patch-based models still suffer from significant computational overhead at higher magnification levels, and hence, do not scale well to WSIs. Therefore, some authors (Lin et al., 2019, 2018) have proposed a variant of FCN which consists of a dense scanning mechanism, that shares computations in overlapping regions during image scanning. To further improve the prediction accuracy of the FCN model, a new pooling layer named as 'anchor layer' is also introduced in Lin et al. (2019) by reconstructing the loss occurred in max-pooling layers. Such models have been shown to have inference speeds a hundred times faster than traditional patch-based approaches, while still ensuring a higher prediction accuracy in WSI analysis. On the other hand, Guo et al. (2019) presented an alternative method for fast breast tumour segmentation, in which, a network first pre-selects the possible tumour area via CNN based classification, and later refines this initial segmentation using an FCN based model. Their proposed framework obtains dense predictions with 1/8 size of original WSI in 11.5 minutes (on CAMELYON16 dataset), compared to the model trained using FCN alone.

Table 1: Overview of supervised learning models. The acronyms for the staining stands for: H&E (haematoxylin and eosin); DAB-H (Diaminobenzidine-Hematoxylin); IFL (Immunofluorescent); ER (Estrogen receptor), PR (Progesterone receptor); PC (Phase contrast); HPF (High power field); Pap (Papanicolaou stain); PHH3 (Phosphohistone-H3); IHC (Immunohistochemistry staining); PAS (Periodic acid–Schiff). Note: (✓) indicates the code is publicly available and the link is provided in their respective paper.

Reference	Cancer types	Staining	Application	Method	Dataset
Classification models					
A. Local-level task					
Cireşan et al. (2013)	Breast	H&E	Mitosis detection	Pixel based CNN classifier	ICPR2012 (50 images)
Wang et al. (2014)	Breast	H&E	Mitosis detection	Cascaded ensemble of CNN + hand-crafted features	ICPR2012 (50 images)
Song et al. (2015)	Cervix	H&E	Segmentation of cervical cytoplasm and nuclei	Multi-scale CNN + graph-partitioning approach	Private set containing 53 cervical cancer images
Kashif et al. (2016)	Colon	H&E	Cell detection	Spatially constrained CNN + hand-crafted features	15 images of colorectal cancer tissue images
Xing et al. (2016)	Multi-Cancers	H&E, IHC	Nuclei segmentation	CNN + selection-based sparse shape model	Private set containing brain tumour (31), pancreatic NET (22), breast cancer (35) images
Romo-Bucheli et al. (2016)	Breast	H&E	Tubule nuclei detection and classification	CNN based classification of pre-detected candidate nuclei	174 images with ER(+) breast cancer cases
Wang et al. (2016b)	Lung	H&E	Cell detection	Two shared-weighted CNNs for joint cell detection and classification	TCGA (300 images)
Albarqouni et al. (2016)	Breast	H&E	Mitosis detection	Multi-scale CNN via crowdsourcing layer	AMIDA2013 (666 - HPF images)
Song et al. (2017)	Cervix	Pap, H&E	Segmentation of cervical cells	Multi-scale CNN model	Overlapping cervical cytology image segmentation challenge (ISBI 2015) - 8 images, private set - 21 images
Gao et al. (2017)	Multi-Cancers	IFL	Cell classification	CNN (LeNet-5) based classification of HEp2-cells	ICPR2012 (28 images), ICPR2014 (83 images)
Rao (2018)	Breast	H&E	Mitosis detection	Faster RCNN based multi-scale region proposal model	ICPR2012 (50 images), AMIDA2013 (23 images), ICPR2014 (2112 images)
Tellez et al. (2018)	Breast	H&E, PHH3	Mitosis detection	Ensemble of CNNs using H&E registered to PHH3 tissue slides as reference standard	TNBC (36 images), TUPAC (814 images)
Qaiser et al. (2019b)	Colon	H&E	tumour segmentation	Combination of CNN and persistent homology feature based patch classifier	Two private sets containing 75 and 50 colorectal adenocarcinoma WSIs
B. Global-level task					
Cruz-Roa et al. (2014)	Breast	H&E	Detection of invasive ductal carcinoma	CNN based patch classifier	Private set - 162 cases
Ertosun and Rubin (2015)	Brain	H&E	Glioma grading	Ensemble of CNN models	TCGA (54 WSIs)
Litjens et al. (2016)	Multi-Cancers	H&E	Detection of prostate and breast cancer	CNN based pixel classifier	Two private sets (225 + 173 WSIs)
Bejnordi et al. (2017)	Breast	H&E	Breast cancer classification	Stacked CNN incorporating contextual information	Private set - 221 images
Agarwalla et al. (2017)	Breast	H&E	tumour segmentation	CNN + 2D-LSTM for representation learning and context aggregation	Camelyon16 (400 WSIs)
Kong et al. (2017)	Breast	H&E	Detection of breast cancer metastases	CNN with the 2D-LSTM to learn spatial dependencies between neighboring patches	Camelyon16 (400 WSIs)
Vandenberghe et al. (2017)	Breast	IHC	IHC scoring of HER2 status in breast cancer	CNN based patch classifier	71 WSIs of invasive breast carcinoma (Private set)
Cruz-Roa et al. (2017)	Breast	H&E	Detection of invasive breast cancer	CNN based patch classifier	TCGA + four other private sets (584 cases)

Sharma et al. (2017)	Stomach	H&E, IHC	Gastric cancer classification and necrosis detection	Patch-based CNN classifier	Private set - 454 WSIs
BenTaieb and Hamarneh (2018) (✓)	Breast	H&E	Detection of breast cancer metastases	CNN based recurrent visual attention model	Camelyon16 (400 WSIs)
Awan et al. (2018)	Breast	H&E	Breast cancer classification	CNN based patch classification model incorporating contextual information	BACH 2018 challenge (400 WSIs)
Li and Ping (2018) (✓)	Breast	H&E	Detection of breast cancer metastases	CNN + CRF to model spatial correlations between neighboring patches	Camelyon16 (400 WSIs)
Bejnordi et al. (2018)	Breast	H&E	Detection of invasive breast cancer	Multi-stage CNN that first identifies tumour-associated stromal alterations and further classify into normal/benign vs invasive breast cancer	Private set - 2387 WSIs
Cruz-Roa et al. (2018)	Breast	H&E	Detection of invasive breast cancer	Patch based CNN model with adaptive sampling method to focus only on high uncertainty regions	TCGA + 3 other public datasets (596 cases)
Qaiser et al. (2019b)	Breast	IHC	Immunohistochemical scoring of HER2	Deep reinforcement learning model that treats IHC scoring as a sequential learning task using CNN + RNN	HER2 scoring contest (172 images), private set - 82 gastroenteropancreatic NET images
Wei et al. (2019) (✓)	Lung	H&E	Classifcation of histologic subtypes on lung adenocarcinoma	ResNet-18 based patch classifier	Private set - 143 WSIs
Nagpal et al. (2019)	Prostate	H&E	Predicting Gleason score	CNN based regional Gleason pattern classification + k-nearest-neighbor based Gleason grade prediction	TCGA (397 cases) + two private sets (361 + 11 cases)
Shaban et al. (2019a) (✓)	Mouth	H&E	tumour infiltrating lymphocytes abundance score prediction for disease free survival	CNN (MobileNet) based patch classifier, followed by statistical analysis	70 cases of oral squamous cell carcinoma WSIs (Private set)
Halicek et al. (2019)	Head & Neck	H&E	Detection of squamous cell carcinoma and thyroid carcinoma	CNN (Inception-v4) based patch classifier	Private set - 381 images
Xu et al. (2019)	Breast	H&E	Detection of breast cancer	Deep hybrid attention (CNN + LSTM) network	BreakHis (7,909 images)
Zhang et al. (2019) (✓)	Bladder	H&E	Bladder cancer diagnosis	CNN + RNN to generate clinical diagnostic descriptions and network visual attention maps	913 images of urothelial carcinoma from TCGA and private set
Regression models					
Xie et al. (2015a)	Multi-Cancers	Ki-67	Nuclei detection	CNN based hough voting approach	Neuroendocrine tumour set (private - 44 images)
Xie et al. (2015b)	Multi-Cancers	H&E, Ki-67	Cell detection	CNN based structured regression model	TCGA (Breast-32 images), HeLa cervical cancer (22 images), Neuroendocrine tumour images (60 images)
Chen et al. (2016a)	Breast	H&E	Mitosis detection	FCN based deep regression network	ICPR2012 (50 images)
Sirinukunwattana et al. (2016)	Colon	H&E	Nuclei detection and classification	CNN with spatially constrained regression	CRCHisto (100 images)
Naylor et al. (2018) (✓)	Multi-Cancers	H&E	Nuclei segmentation	CNN based regression model for touching nuclei segmentation	TNBC (50 images), MoNuSeg (30 images)
Xie et al. (2018b)	Multi-Cancers	H&E, Ki-67	Cell detection	Structured regression model based on fully residual CNN	TCGA (Breast-70 image patches), Bone marrow (11 image patches), HeLa cervical cancer (22 images), Neuroendocrine tumour set (59 image patches)
Graham et al. (2019b) (✓)	Multi-Cancers	H&E	Nuclei segmentation and classification	CNN based instance segmentation and classification framework	CoNSeP (41 images), MoNuSeg (30 images), TNBC (50 images), CRCHisto (100 images), CPM-15 (15 images), CPM-17 (32 images)
Xing et al. (2019)	Pancreas	Ki-67	Nuclei detection and classification	FCN based structured regression model	Pancreatic neuroendocrine tumour set (private - 38 images)

Segmentation models					
BenTaieb and Hamarneh (2016)	Colon	H&E	Segmentation of colon glands	A loss function accounting for boundary smoothness and topological priors in FCN learning	GLAS challenge (165 images)
Chen et al. (2017a)	Multi-Cancers	H&E	Segmentation of glands and nuclei	Multi-task learning framework with contour-aware FCN model for instance segmentation	GLAS challenge (165 images), MICCAI 2015 nucleus segmentation challenge (33 images)
Xu et al. (2017)	Colon	H&E	Segmentation of colon glands	Multi-channel deep network model for gland segmentation and instance recognition	GLAS challenge (165 images)
de Bel et al. (2018)	Kidney	PAS	Segmentation of renal tissue structures	Evaluated three different architectures: FCN, Multi-scale FCN and UNet	15 WSIs of renal allograft resections (private set)
Van Eycke et al. (2018)	Colon	H&E, IHC	Segmentation of glandular epithelium in H&E and IHC staining images	CNN model based on integration of DCAN, UNet and ResNet models	GLAS challenge (165 images) and a private set containing colorectal tissue microarray images
Gecer et al. (2018)	Breast	H&E	Detection and classification of breast cancer	Ensemble of multi-scale FCN's followed by CNN based patch classifier	240 breast histopathology WSIs (private set)
Gu et al. (2018)	Breast	H&E	Detection of breast cancer metastasis	UNet based multi-resolution network with multi-encoder and single decoder model	Camelyon16 (400 images)
Guo et al. (2019)	Breast	H&E	Detection of breast cancer metastasis	Classification (Inception-V3) based semantic segmentation model (DCNN)	Camelyon16 (400 images)
Bulten et al. (2020)	Prostate	H&E	Grading of prostate cancer	UNet based segmentation of Gleason growth patterns, followed by subsequent cancer grading	1243 WSIs of prostate biopsies (private set)
Lin et al. (2019)	Breast	H&E	Detection of breast cancer metastasis	FCN based model for fast inference of WSI analysis	Camelyon16 (400 WSIs)
Liu et al. (2019)	Breast	DAB-H	Immunohistochemical scoring for breast cancer	Multi-stage FCN framework that directly predicts H-Scores of breast cancer TMA images	105 TMA images of breast adenocarcinomas (private set)
Bulten et al. (2019)	Prostate	IHC, H&E	Segmentation of epithelial tissue	Pre-trained UNet on IHC is used as a reference standard to segment epithelial structures in H&E WSIs	102 prostatectomy WSIs
Swiderska-Chadaj et al. (2019)	Multi-Cancers	IHC	Lymphocyte detection	Investigated the effectiveness of four DL methods - FCN, UNet, YOLO and LSM	LYON19 (test set containing 441 region-of-interests (ROIs))
Graham et al. (2019a)	Colon	H&E	Segmentation of colon glands	FCN with minimum information loss units and atrous spatial pyramid pooling	GLAS challenge (165 images), CRAG dataset (213 images)
Ding et al. (2019)	Colon	H&E	Segmentation of colon glands	Multi-scale FCN model with a high-resolution branch to circumvent the loss in max-pooling layers	GLAS challenge (165 images), CRAG dataset (213 images)
Zhao et al. (2019)	Breast	H&E	Detection and classification of breast cancer metastasis	Feature pyramid aggregation based FCN network with synergistic learning approach	Camelyon16 (400 WSIs), Camelyon17 (1000 WSIs)
Qu et al. (2019) (✓)	Lung	H&E	Nuclei segmentation and classification	FCN trained with perceptual loss	40 tissue images of lung adenocarcinoma (private set)
Ho et al. (2019)	Breast	H&E	Breast cancer multi-class tissue segmentation	Deep multi-magnification model with multi-encoder, multi-decoder and multi-concatenation network	Private set containing TNBC (38 images) and breast margin dataset (10 images)
Tokunaga et al. (2019)	Lung	H&E	Segmentation of multiple cancer subtype regions	Multiple UNets trained with different FOV images + an adaptive weighting CNN for output aggregation	29 WSIs of lung adenocarcinoma (private set)
Lin et al. (2019)	Breast	H&E	Detection of breast cancer metastasis	FCN based model with anchor layers for fast and accurate prediction of cancer metastasis	Camelyon16 (400 images)
Pinckaers and Litjens (2019) (✓)	Colon	H&E	Segmentation of colon glands	Incorporating neural ordinary differential equations in UNet to allow an adaptive receptive field	GLAS challenge (165 images)
Seth et al. (2019)	Breast	H&E	Segmentation of DCIS	Compared UNets trained at multiple resolutions	training:183 WSIs, testing:19 WSIs (private set)

3.2. Weakly supervised learning

The idea of weakly supervised learning (WSL) is to exploit coarse-grained (image-level) annotations to automatically infer fine-grained (pixel/patch-level) information. This paradigm is particularly well suited to the histopathology domain, where the coarse-grained information is often readily available in the form of image-level labels, e.g., cancer or non-cancer, but where pixel-level annotations are more difficult to obtain. Weakly supervised learning dramatically reduces the annotation burden on a pathologist (Xu et al., 2014), and an overview of these models is provided in Table 2.

In this survey, we explore one particular form of WSL, namely *multiple-instance learning* (MIL), which aims to train a model using a set of weakly labeled data (Dietterich et al., 1997; Quellec et al., 2017). In MIL, a training set consists of bags, labeled as positive or negative; and each bag includes many instances, whose label is to be predicted or unknown. For instance, each histology image with cancer/non-cancer label forms a *'bag'* and each pixel/patch extracted from the corresponding image is referred to as an *'instance'* (e.g., pixels containing cancerous cells). Here, the main goal is to train a classifier to predict both bag-level and instance-level labels, while only bag-level labels are given in the training set. We further categorize MIL approaches into three categories similar to Cheplygina et al. (2019): i) *global detection* - identifying a target pattern in a histology image (i.e., at bag level) such as the presence or absence of cancer; ii) *local detection* - identifying a target pattern in an image patch or a pixel (i.e., at instance level) such as highlighting the cancerous tissues or cells; iii) *global and local detection* - detecting whether an image has cancer and also identifying the location where it occurs within an image. These categories are illustrated in Fig. 4. There is also a significant interest in histopathology to include various kinds of weak annotations such as image-level tags (Campanella et al., 2019), points (Qu et al., 2019), bounding boxes (Yang et al., 2018), polygons (Wang et al., 2019) and percentage of the cancerous region within each image (Jia et al., 2017), to obtain clinically satisfactory performance with minimal annotation effort. For an in-depth review of MIL approaches in medical image analysis, refer to Quellec et al. (2017); Cheplygina et al. (2019); Rony et al. (2019); Kandemir and Hamprecht (2015).

Due to the variable nature of histopathology image appearances, the standard instance-level aggregation methods, such as voting or pooling, do not guarantee accurate image-level predictions, due to misclassifications of instance-level labels (Campanella et al., 2019; Rony et al., 2019). Hence, several papers on global detection based MIL method rely on alternative instance-level aggregation strategies to obtain reliable bag-level predictions suitable for a given histology task. For instance, Hou et al. (2015) integrated an expectation-maximization based MIL method with a CNN to output patch-level predictions. These instances are later aggregated by training a logistic regression model to classify glioma subtypes in WSIs. Dov et al. (2019) proposed an alternative approach based on ordinal regression framework for aggregating instances containing follicular (thyroid) cells to simultaneously predict both thyroid malignancy and TBS score in whole-slide cytopathology images. Recently, a remarkable work in Campanella et al. (2019) adopted an RNN model to integrate semantically rich feature representations across patch-level instances to obtain a final slide-level diagnosis. In their method, the author's managed to obtain an AUC greater than 0.98 in detecting four types of cancers on an extensive multi-centre dataset of 44,732 WSIs, without expensive pixel-wise manual annotations.

The local detection based MIL approaches are based on an image-centric paradigm, where image-to-image prediction is performed using an FCN model - by computing features for all instances (pixels) together. These approaches are generally applied to image segmentation task for precisely delineating cancerous region in histology images. In the local detection approach, the bag labels are propagated to all instances to train a classifier in a supervised manner. However, sometimes even the best bag-level classifier seems to underperform on instance-level predictions due to lack of supervision (Cheplygina et al., 2019). To tackle this issue, additional weak constraints have been incorporated into FCN models to improve segmentation accuracy. For example, Jia et al. (2017) included an area constraint in the MIL formulation by calculating the rough estimate of the relative size of the cancerous region. However, calculating such area constraints is tedious and can only be performed by an expert pathologist. Consequently, Xu et al. (2019) proposed an alternative MIL framework to generate instance-level labels from image-level annotations. These pre-

Table 2: Overview of weakly supervised learning models. Note: (✓) indicates the code is publicly available and the link is provided in their respective paper.

Reference	Cancer types	Staining	Application	Method	Dataset
Multiple instance learning (MIL)					
Hou et al. (2015)	Brain	H&E	Glioma subtype classification	Expectation-maximization based MIL with CNN + logistic regression	TCGA (1,064 slides)
Jia et al. (2017)	Colon	H&E	Segmentation of cancerous regions	FCN based MIL + deep supervision and area constraints	Two private sets containing colon cancer images (910+60 images)
Liang et al. (2018)	Stomach	H&E	Gastric tumour segmentation	Patch-based FCN + iterative learning approach	China Big Data and AI challenge (1,900 images)
Ilse et al. (2018) (✓)	Multi-Cancers	H&E	Cancer image classification	MIL pooling based on gated-attention mechanism	CRCHisto (100 images)
Shujun Wang et al. (2019)	Stomach	H&E	Gastric cancer detection	Two-stage CNN framework for localization and classification	Private set (608 images)
Wang et al. (2019)	Lung	H&E	Lung cancer image classification	Patch based FCN + context-aware block selection and feature aggregation strategy	Private (939 WSIs), TCGA (500 WSIs)
Campanella et al. (2019) (✓)	Multi-Cancers	H&E	Multiple cancer diagnosis in WSIs	CNN (ResNet) + RNNs	Prostate (24,859 slides), skin (9,962 slides), breast cancer metastasis (9,894 slides)
Dov et al. (2019)	Thyroid	—	Thyroid malignancy prediction	CNN + ordinal regression for prediction of thyroid malignancy score	Private set (cytopathology 908 WSIs)
Xu et al. (2019) (✓)	Multi-Cancers	H&E	Segmentation of breast cancer metastasis and colon glands	FCN trained on instance-level labels, which are obtained from image-level annotations	Camelyon16 (400 WSIs), Colorectal adenoma private dataset (177 WSIs)
Huang and Chung (2019)	Breast	H&E	Localization of cancerous evidence in histopathology images	CNN + multi-branch attention modules and deep supervision mechanism	PCam (327,680 patches extracted from Camelyon16) and Camelyon16 (400 WSIs)
Other approaches					
Campanella et al. (2018)	Prostate	H&E	Prostate cancer detection	CNN trained under MIL formulation with top-1 ranked instance aggregation approach	Prostate biopsies (12,160 slides)
Akbar and Martel (2018) (✓)	Breast	H&E	Detection of breast cancer metastasis	Clustering (VAE + K-means) based MIL framework	Camelyon16 (400 WSIs)
Tellez et al. (2019b) (✓)	Multi-Cancers	H&E	Compression of gigapixel histopathology WSIs	Unsupervised feature encoding method (VAE, Bi-GAN, contrastive training) that maps high-resolution image patches to low-dimensional embedding vectors	Camelyon16 (400 WSIs), TUPAC16 (492 WSIs), Rectum (74 WSIs)
Qu et al. (2019) (✓)	Multi-Cancers	H&E	Nuclei segmentation	Modified UNet trained using coarse level-labels + dense CRF loss for model refinement	MoNuSeg (30 images), lung cancer private set (40 images)
Bokhorst et al. (2019)	Colon	H&E	Segmentation of tissue types in colorectal cancer	UNet with modified loss functions to circumvent sparse manual annotations	Colorectal cancer WSIs (private set - 70 images)
Li et al. (2019a) (✓)	Breast	H&E	Mitosis detection	FCN trained with concentric loss on weakly annotated centriod label	ICPR12 (50 images), ICPR14 (1,696 images), AMIDA13 (606 images), TUPAC16 (107 images)

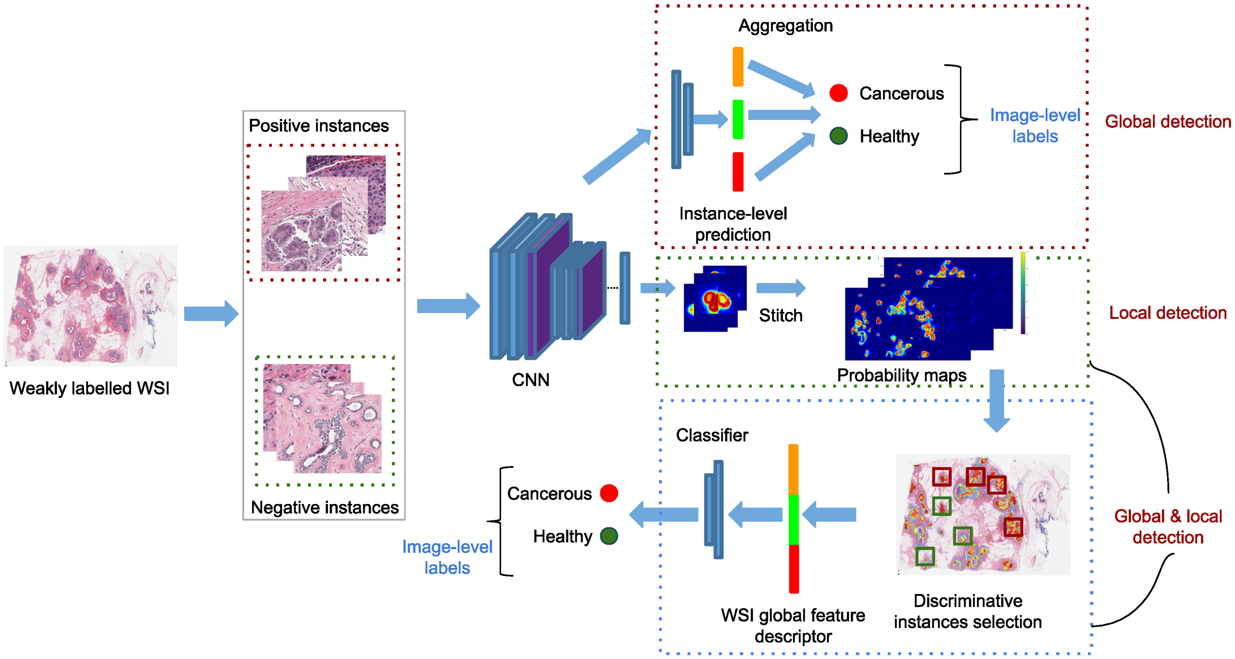


Figure 4: An overview of weakly supervised learning models.

dicted instance-level labels are later assigned to their corresponding image pixels to train an FCN in an end-to-end manner, while achieving comparable performance with supervised counterparts. Finally, in some cases, both a large number of bag labels and a partial set of instance labels are also adopted in FCN based reiterative learning framework (Liang et al., 2018), to further optimize final instance-level predictions.

Arguably, the most popular and clinically relevant MIL approach in histopathology is the global and local detection paradigm. In this approach, rather than just diagnosing cancer at whole-slide level, we can simultaneously localize the discriminative areas (instances) containing cancerous tissues or cells. In this context, the methods utilize either the bag-level label (Shujun Wang et al., 2019) or both bag-level and some coarse level instance annotations (Wang et al., 2019) to infer a global level decision. Note that the instance-level predictions are not usually validated due to lack of costly annotations, and are generally visualized as either a heatmap (Shujun Wang et al., 2019; Wang et al., 2019) or a saliency map (Huang and Chung, 2019) to highlight the diagnostically significant locations in WSIs. The main essence of this approach is to capture the instance-wise dependencies and their impact on the final image-level decision score.

There is a some disagreement among MIL methods regarding the accuracy of instance-level predictions, when trained with only bag-level labels (Cheplygina et al., 2019; Kandemir and Hamprecht, 2015). The critical and often overlooked issue among MIL methods is that even the best bag-level classifier may not be an optimal instance-level classifier for instance predictions and vice versa (Cheplygina et al., 2019). Such problems have naturally led to new solutions that integrate the visual attention models with MIL techniques to enhance the interpretability of final model predictions (Ilse et al., 2018; Huang and Chung, 2019). For instance, Huang and Chung (2019) proposed a CNN model combining multi-branch attention modules and a deep supervision mechanism (Xie and Tu, 2015), which aims to localize the discriminative evidence for the class-of-interest from a set of weakly labeled training data. Such attention-based models can precisely pinpoint the location of cancer evidence in WSI, as well as achieving a competitive slide-level accuracy, thereby enhancing the interpretability of current DL models in histopathology applications.

Not all methods identified as weakly supervised in the literature necessarily fall under the MIL category. For instance, the methods in Qu et al. (2019); Bokhorst et al. (2019); Li et al. (2019a) use the term "*weakly supervised*" to indicate that the model training has been performed on sparse set of annotations such as points inside the region of interest (Li et al., 2019a; Qu et al., 2019), bounding box (Yang et al., 2018) and also some partial pixel-level annotations of cancerous region (Bokhorst et al., 2019). These approaches alleviate the need for expensive annotations by proposing newer variants of loss functions (Li et al., 2019a), feature encoding strategies (Tellez et al., 2019b; Akbar and Martel, 2018), loss balancing mechanisms (Bokhorst et al., 2019), and methods to derive coarse labels from weak annotations (Qu et al., 2019) in order to eventually train fully-supervised models in a weakly supervised way.

3.3. Unsupervised learning

The goal of unsupervised learning is to learn something useful about the underlying data structure without the use of labels. The term "unsupervised" is sometimes used loosely among the digital pathology community for approaches that are not fully unsupervised. For instance, stain transfer without pairing, or domain adaptation via feature distribution matching are considered as unsupervised, even though the domains can be considered as labels for two separate datasets (Gadermayr et al., 2019a; de Bel et al., 2019; Ganin et al., 2016). In this survey, we examine fully unsupervised methods, where the raw data comes in the form of images without any identifiers (e.g., domain, cancerous vs. non-cancerous, tissue etc.). These approaches are rare, since the field of unsupervised learning among the machine learning community is also still in its infancy. However, it is clear why one should be interested in such approaches as the scarcity of labeled data due to regulatory concerns and labor costs (i.e., expert annotations) is a major bottleneck in achieving clinically satisfactory performance in medical imaging (Lee and Yoon, 2017).

In unsupervised learning, the learning task is ambiguous, since it is possible to map the inputs into infinitely many subsets, provided there are no restrictions. Most unsupervised approaches aim to maximize the probability distribution of the data, subject to some constraints, in order to limit the solution space and to achieve a desired grouping/clustering for the target task. A common technique is to transform the data into a lower-dimensional subspace, followed by aggregation of feature representations into mutually exclusive or hierarchical clusters, which is illustrated in Fig. 5. Autoencoders are typically utilized for the dimensionality reduction step. Recent advances in modeling the stochasticity (Kingma and Welling, 2013), and more robustly disentangling visual features (Higgins et al., 2017; Chen et al., 2018) have made autoencoders more attractive for feature modeling and dimensionality reduction. In early work, sparse autoencoders were utilized for unsupervised nuclei detection (Xu et al., 2015a). Later, detection performance was improved by modifying the receptive field of the convolutional filters to accommodate small nuclei (Hou et al., 2019b). For more complex tasks, such as tissue and cell classification, Generative Adversarial Networks (GANs) have also been employed. Specifically, InfoGANs (Chen et al., 2016b) have been used for extracting features, which maximize the mutual information between the generated images and a predefined subset of latent (noise) codes, which are then used for tasks such as cell-level classification, nuclei segmentation, and cell counting (Hu et al., 2018a).

Finally, we examine unsupervised transfer learning approaches, where instead of directly applying learned features on a target task, learned mapping functions are used as an initialization for target tasks, possibly with very few labeled training images. Using a loss term that is similar to the reconstruction objective of autoencoders, (Chang et al., 2017) trains a convolutional network using unlabeled images pertaining to a specific modality (e.g., brain MRI or kidney histology images), to learn filter banks at different scales. The resulting filters are shift-invariant, scale-specific, and can uncover intricate patterns in various tasks, such as tumour classification of glioblastoma multiforme or kidney renal clear cell carcinoma. In machine learning, this form of unsupervised learning is called *"self-supervised"* learning. Since self-supervised techniques can deal with larger images in general, they offer a promising alternative to clustering approaches in histopathology, which usually require context and a larger

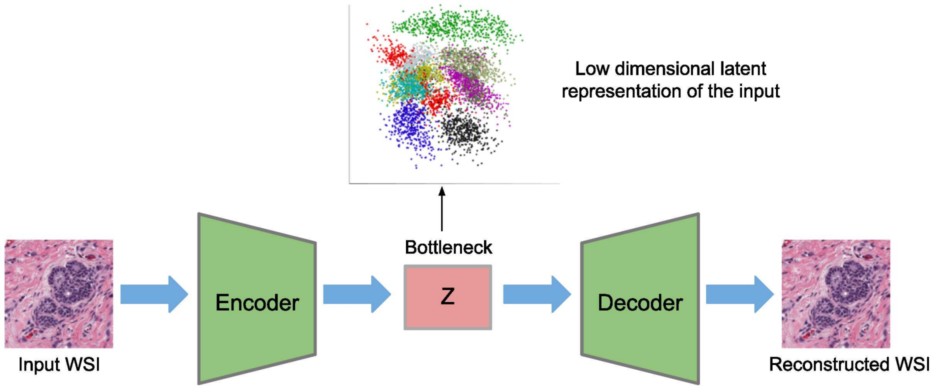


Figure 5: An overview of unsupervised learning models.

field of view. Context-based self-supervised methods which predict spatial ordering (Noroozi and Favaro, 2016) or image rotations (Gidaris et al., 2018), and generative methods such as mapping grayscale images into their RGB counterparts have been successfully used for initializing networks for faster convergence and learning target tasks with fewer labels. However, in histopathology, the rules governing the spatial location of cell structures, or the color or staining of a histology image are different to those for natural scene images. While, this makes the task of unsupervised learning more difficult for histopathology images, it also presents an opportunity for researchers to develop novel techniques that may be applicable to medical images.

Unsupervised learning methods are desirable as they allow models to be trained with little or no labeled data. Furthermore, as these methods are constructed to disentangle relationships between samples in the dataset for grouping (or clustering), a successful unsupervised learning method can also improve the interpretability of a model, by examining how the model groups items into separate categories. While fully unsupervised methods for arbitrary tasks are still uncommon, techniques used for auxiliary tasks (e.g., pre-training) such as self-supervision (Tellez et al., 2019b) can reduce the annotation burden on the expert, thereby significantly expediting the research.

3.4. Transfer learning

The most popular and widely adopted technique in digital pathology is the use of *transfer learning* approach. In transfer learning, the goal is to extract knowledge from one domain (i.e., source) and apply it to another domain (i.e., target) by relaxing the assumption that the train and test set must be independent and identically distributed. In histopathology, transfer learning is typically done using ImageNet pretrained models such as *VGGNet* (Simonyan and Zisserman, 2014), *InceptionNet* (Szegedy et al., 2015, 2016), *ResNet* (He et al., 2016), *MobileNet* (Howard et al., 2017), *DenseNet* (Huang et al., 2017), and various other variants of these models. These pre-trained models have been widely applied to various cancer grading and prognosis tasks (Refer, Table 4 and Section 4 for more details). A critical analysis of best-performing methods on various Grand Challenges is discussed thoroughly in Section 5.1.

In digital pathology, different types of staining are used depending on the application. Immunohistochemistry (IHC) allows specific molecular targets to be visualized (e.g., Ki-67 to estimate tumour cell proliferation rate (Valkonen et al., 2019), and cytokeratin to detect micrometastases (Ehteshami Bejnordi et al., 2017), whilst H&E is a widely-used general-purpose stain. The appearance of images varies widely depending on the stain used and also on the degree of staining, and this poses a unique challenge, as CNN's are highly sensitive to the data they were trained on (Ciompi et al., 2017). In the following sub-sections, we review two approaches used to overcome this problem.

Table 3: Overview of unsupervised learning models. Note: (✓) indicates the code is publicly available and the link is provided in their respective paper.

Reference	Cancer types	Staining	Application	Method	Dataset
Xu et al. (2015a)	Breast	H&E	Nuclei segmentation	Stacked sparse autoencoders	537 H&E images from Case Western Reserve University
Hu et al. (2018a) (✓)	Bone marrow	H&E	Tissue and cell classification	InfoGAN	3 separate datasets: public data with 11 patches of size 1200 × 1200, private datasets with WSIs of 24 patients + 84 images
Bulten and Litjens (2018)	Prostate	H&E, IHC	Classification of prostate into tumour vs. non-tumour	Convolutional adversarial autoencoders	94 registered WSIs from Radboud University Medical Center
Sari and Gunduz-Demir (2019)	Colon	H&E, IHC	Subtyping of intrahepatic cholangiocarcinoma (ICC)	Restricted Boltzmann Machines + Clustering	3236 images, private dataset
Quiros et al. (2019) (✓)	Breast	H&E	High resolution image generation + feature extraction	BigGAN + Relativistic GAN	248 + 328 patients from private dataset
Hou et al. (2019b)	Breast	H&E	Nuclei detection, segmentation and representation learning	Sparse autoencoder	0.5 million images of nuclei from TCGA
Gadermayr et al. (2019a)	Kidney	Stain agnostic	Segmentation of object-of-interest in WSIs	CycleGAN + UNet segmentation	23 PAS, 6 AFOG, 6 Col3 and 6 CD31 WSIs
de Bel et al. (2019)	Kidney	Stain agnostic	Tissue segmentation	CycleGAN + UNet segmentation	Private set containing 40 + 24 biopsy images
Gadermayr et al. (2019b)	Kidney	PAS, H&E	Segmentation of the glomeruli	CycleGAN	23 WSIs, private dataset

3.4.1. Domain adaptation

Domain adaptation is a sub-field of *transfer learning*, where a task is learned from one or more source domains with labeled data, and the aim is to achieve similar performance on the same task on a target domain with little or no labeled data (Wang and Deng, 2018). Domain-adversarial networks are designed to learn features that are discriminative for the main prediction task whilst being insensitive to domain shift (Ganin et al., 2016; Lafarge et al., 2017; Ren et al., 2018) and this approach has been applied to digital pathology. Ciga et al. (2019) achieved state-of-the-art performance on the BACH (BreAst Cancer Histopathology) challenge task using a multi-level domain-adversarial network. Ren et al. (2018) performed unsupervised training based on siamese networks on prostate WSIs, positing that given a WSI, different patches should be given the same Gleason score, thereby extracting common features present in different parts of the WSI. This auxiliary task also helped increase the adversarial domain adaptation performance on another target dataset.

Fake (artificially generated) images are also used in domain adaptation. Brieu et al. (2019) utilized semi-automatic labeling of the nuclei with one type of staining (IHC) to alleviate the costlier annotation of another staining method (H&E), where fake H&E images are generated from IHC images to increase the dataset size. Similarly, Gadermayr et al. (2019a) used artificial data generation with GANs for semantic segmentation in kidney histology images with multiple stains. Each work uses adversarial models (i.e., generators and discriminators) for image-to-image translation utilizing cycle consistency loss for unpaired training. The translation is performed to obtain an intermediate, stain-agnostic representation (Lahiani et al., 2019), which is then fed to a network trained on this representation to perform segmentation.

3.4.2. Stain normalization

Stain normalization, augmentation and stain transfer are popular image preprocessing techniques to improve generalization of a task by modifying the staining properties of a given image to match another image visually. In contrast to the methods described in Section 3.4.1, which modify the *features* extracted from different image distri-

Table 4: Overview of transfer learning models. Note: (✓) indicates the code is publicly available and the link is provided in their respective paper.

Reference	Cancer types	Staining	Application	Method	Dataset
Wang et al. (2016a)	Breast	H&E	Detection of breast cancer metastasis	Pre-trained GoogleNet model	Camelyon16 (400 WSIs)
Liu et al. (2017)	Breast	H&E	Detection of breast cancer metastasis	Pre-trained Inception-V3 model	Camelyon16 (400 WSIs)
Han et al. (2017)	Breast	H&E	Breast cancer multi-classification	CNN integrated with feature space distance constraints for identifying feature space similarities	BreaKHis (7,909 images)
Lee and Paeng (2018)	Breast	H&E	Detection and pN-stage classification of breast cancer metastasis	Patch based CNN for metastasis detection + Random forest classifier for lymph node classification	Camelyon17 (1,000 WSIs)
Chennamsetty et al. (2018)	Breast	H&E	Breast cancer classification	Ensemble of three pre-trained CNNs + aggregation using majority voting	BACH 2018 challenge (400 WSIs)
Kwok (2018)	Breast	H&E	Breast cancer classification	Inception-Resnet-V2 based patch classifier	BACH 2018 challenge (400 WSIs)
Bychkov et al. (2018)	Colon	H&E	Outcome prediction of colorectal cancer	A 3-layer LSTM + VGG-16 pre-trained features to predict colorectal cancer outcome	Private set (420 cases)
Arvaniti et al. (2018) (✓)	Prostate	H&E	Predicting Gleason score	Pre-trained MobileNet architecture	Private set (886 cases)
Coudray et al. (2018) (✓)	Lung	H&E	Genomics prediction from pathology images	Patch based Inception-V3 model	TCGA (1,634 WSIs) and validated on independent private set containing frozen sections (98 slides), FFPE sections (140 slides) and lung biopsies (102 slides)
Kather et al. (2019) (✓)	Colon	H&E	Survival prediction of colorectal cancer	Pre-trained VGG-19 based patch classifier	TCGA (862 WSIs) and two other public datasets (25 + 86 WSIs)
Noorbakhsh et al. (2019) (✓)	Multi-Cancers	H&E	Pan-cancer classification	Pre-trained Inception-V3 model	TCGA (27,815 WSIs)
Tabibu et al. (2019) (✓)	Kidney	H&E	Classification of Renal Cell Carcinoma subtypes and survival prediction	Pre-trained ResNet based patch classifier	TCGA (2,093 WSIs)
Akbar et al. (2019)	Breast	H&E	tumour cellularity (TC) scoring	Two separate Inception-Nets: one for classification (healthy vs. cancerous tissue) and the other outputs regression scores for TC	BreastPathQ (96 WSIs)
Valkonen et al. (2019) (✓)	Breast	ER, PR, Ki-67	Cell detection	Fine-tuning partially pre-trained CNN network	DigitalPanCK (152 - invasive breast cancer images)
Ström et al. (2020)	Prostate	H&E	Grading of prostate cancer	Ensembles of two pre-trained Inception-V3 models	Private set (8730 WSI's)

Table 5: Overview of domain adaptation and stain normalization models. Note: (✓) indicates the code is publicly available and the link is provided in their respective paper.

Reference	Cancer types	Staining	Application	Method	Dataset
Domain adaptation					
Lafarge et al. (2017)	Breast	H&E	Mitosis detection	Gradient reversal with CNNs	TUPAC16 (73 WSIs)
Ren et al. (2018)	Prostate	H&E	Feature matching of image patches	Siamese networks	TCGA + private dataset
Brieu et al. (2019)	Multi-Cancers	Multi-stain	Semi-automatic nuclei labeling using stain transfer	CycleGAN	TCGA (75 bladder cancer + 29 lung cancer + 142 tissue samples of FOVs images) + 30 FOVs of breast cancer (private set)
Gadermayr et al. (2019a)	Kidney	Stain agnostic	Segmentation of object-of-interest in WSIs	CycleGAN + UNet segmentation	23 PAS, 6 AFOG, 6 Col3 and 6 CD31 WSIs
Kapil et al. (2019)	Lung	PD-L1 + Cytokeratin	Segmentation	CycleGAN + SegNet segmentation model	56 Cytokeratin + 69 PD-L1 WSIs (private set)
Ciga et al. (2019)	Breast	H&E	Classification	Multi-layer gradient reversal	BACH
Stain variability					
Janowczyk et al. (2017) (✓)	Multi-Cancers	H&E	Stain transfer for H&E staining	Sparse autoencoders	5 breast biopsy slides + 7 gastrointestinal biopsies
Cho et al. (2017)	Breast	H&E	Stain transfer	DCGAN conditioned on a target image	CAMELYON16
BenTaieb and Hamarneh (2017)	Multi-Cancers	H&E	Stain transfer	GAN + regularization based on auxiliary task performance	ICPR2014 + GLAS challenge + 135 WSIs (private set)
Zanjani et al. (2018) (✓)	Lymph nodes	H&E	Stain transfer for H&E staining	Multiple studies with Gaussian mixture models, variational autoencoders, and InfoGAN	625 images from 125 WSIs of lymph nodes from 3 patients
de Bel et al. (2019)	Kidney	Stain agnostic	Segmentation	CycleGAN + UNet segmentation	40 + 24 biopsy images (private)
Shaban et al. (2019b) (✓)	Breast	H&E	Stain transfer	CycleGAN	ICPR2014
Rivenson et al. (2019)	Multi-Cancers	H&E, Jones, Masson's trichrome	Digital staining of multiple tissues	Custom GAN	N/A
Lahiani et al. (2019)	Liver	FAP-CK from Ki67-CD8	Virtual stain transformation between different types of staining	CycleGAN + Instance normalization	10 Ki67-CD8 + 10 FAP-CK stained colorectal carcinoma WSIs

butions so that they are indistinguishable from each other; stain normalization directly modifies the input images to obtain features that are invariant to staining variability.

One may combat staining variation by augmenting the training data by varying each pixel value per channel within a predefined range on transformed color spaces, such as HSV (hue, saturation and value) or HED (Hematoxylin, Eosin, and Diaminobenzidine) (Liu et al., 2017; Li and Ping, 2018; Tellez et al., 2018). Earlier machine learning (ML) methods (Macenko et al., 2009; Vahadane et al., 2016) assume that staining attenuates light (optical density) uniformly and decompose each optical density image into concentration (appearance) and color (stain) matrices. The uniformity assumption is relaxed in more recent ML methods, where the type of chemical staining and morphological properties of an image are considered in generating stain matrices (Khan et al., 2014; Bejnordi et al., 2015). Neural networks, such as sparse autoencoders for template matching (Janowczyk et al., 2017), and GANs are also used for stain transfer and normalization (Zanjani et al., 2018; de Bel et al., 2019; BenTaieb and Hamarneh, 2017; Cho et al., 2017). Cycle consistency loss objective (Zhu et al., 2017a) has been utilized for improved stain transfer with structure preservation, as well as for training systems without annotating the pairing between the source (to be stained) and the target (used as a reference for staining new images) (Shaban et al., 2019b; de Bel et al., 2019; Cho et al., 2017). Auxiliary tasks, such as maintaining high prediction accuracy on classification or segmentation, have led to consistent stain transfer accounting for the type or shape of the tissue present (Odena et al., 2017; BenTaieb and Hamarneh, 2017). Recently, the same techniques have also been used for virtually staining quantitative phase images of label-free tissue sections (Rivenson et al., 2019).

Although stain transfer methods produce aesthetically pleasing results, their use cases are still not entirely clear. For instance, Shaban et al. (2019b) report considerable gains compared to a baseline (no augmentations or traditional methods such as Macenko et al. (2009); Reinhard et al. (2001); Vahadane et al. (2016) in the CAMELYON16 challenge; however, the winning entry (by a margin of 21% with respect to Shaban et al. (2019b) in AUC for a binary classification task on WSIs) utilizes a traditional machine learning based normalization technique that aligns chromatic and density distributions of source and the target (Bejnordi et al., 2015). A thorough study comparing numerous approaches found that it is always advisable to apply various forms of color augmentation in HSV or HED space, and additional slight performance gains are still achievable with a network-based augmentation strategy (Tellez et al., 2019a). Similarly, Stacke et al. (2019) examined the effect of domain shift on histopathological data and automated medical imaging systems, and found that augmentation and normalization strategies drastically alter the performance of these systems. Differentiating factors such as scale-spaces, resolution, image quality, scanner imperfections are also likely to affect the performance of a model, in addition to staining, which is less explored in the community.

4. Survival models for disease prognosis

This section concentrates on methods of training survival models that can either generate a probability of an event in a certain predefined period of time, or can predict time to an event using regression from a WSI. In the context of cancer, the term prognosis refers to the likely outcome for a patient on standard treatment, and the term prediction refers to how the patient responds to a particular treatment. Since the difference between these two terms is not relevant when carrying out survival analysis, we will use the term prediction to cover both prediction and prognosis. The outcome metrics used to train a prediction model will depend on the disease. For example, in patients with very aggressive disease such as glioblastoma, the survival time in months may be used as an endpoint, whereas for breast cancer, with an average survival rate at 10 years of around 80%, the time to recurrence of the disease after surgery is a more relevant metric and at least 5 years follow-up is required. Following up patients prospectively is a time consuming and expensive process and for this reason several studies use existing clinically validated risk models, or genomics assays as a proxy for long term outcomes; for example, the use of PAM50 scores (Veta et al., 2019; Couture et al., 2018) in breast cancer and Gleason grades in prostate cancer (Nagpal et al., 2019). The survival data or risk scores may be dichotomized e.g., survival at specific time points or risk score above or below a set cutoff; this allows the survival model to be treated as a classification problem, but information is lost and new models have to be trained if the cutoff value is changed, e.g., two different models are needed to predict 5-year and 10-year survival times. Time to event models are more complicated since nothing is known about what happens to a patient after they are lost to followup; this is known as right censoring. A proportional hazards model is commonly used to model an individual's survival and can be implemented using a neural network (Katzman et al., 2018) and several groups have used this approach in digital pathology (Zhu et al., 2017b; Mobadersany et al., 2018; Tang et al., 2019).

The data used to train survival models is weakly labeled, with only one outcome label per patient. This poses a computational challenge as a WSI is so large that it has to be broken down into 100's or even 1000's of smaller patches for processing; since tumours may be very heterogeneous, only a subset of these patches may be salient for the prediction task. Although Campanella et al. (2019) recently demonstrated that a relatively simple MIL approach could produce accurate diagnostic results when more than 10,000 slides were available for training, datasets for survival analysis usually have fewer than 1000 slides available which makes the task much more difficult. Three main approaches are used to overcome the shortage of labeled data. The first is to use the image features that expert pathologists have already identified as being associated with survival. Examples include assessing tumour proliferation (Veta et al., 2019) in breast cancer, quantifying the stroma/tumour ratio in colorectal cancer (Geessink et al., 2019) and predicting the Gleason grade in prostate cancer (Nagpal et al., 2019). The role of deep learning in these cases is to provide an automatic and reproducible method for extracting these features; a vital advantage of this approach over end-to-end models is that results are more interpretable since each component can be assessed individually. In the second approach, image features are extracted from image patches using a pre-trained CNN, then feature selection or dimensionality reduction is carried out and finally a survival model is trained on the resulting feature vector. Examples include survival prediction in mesothelioma (Courtiol et al., 2019) and colorectal cancer (Kather et al., 2019; Bychkov et al., 2018), and risk of recurrence in breast cancer (Couture et al., 2018). In the third approach, unsupervised methods are used to learn a latent representation of the data which is then used to train the survival model. For example, Zhu et al. (2017b) apply K-means to small patches to identify 50 clusters or phenotypes for glioma and non-small-cell lung cancer, and Muhammad et al. (2019) use an autoencoder with an additional clustering constraint to predict survival in intrahepatic cholangiocarcinoma.

There are many possible ways of aggregating the predictions for individual patches to give a single prediction for a patient. The simplest approach is to take the mean prediction across all patches (Tang et al., 2019), but this will not work if the salient patches only represent a small fraction of the WSI; for this reason other schemes, such as taking the average of the two highest ranking patches (Mobadersany et al., 2018), may be more appropriate. Some methods generate a low dimensional feature vector that captures the distribution of scores across the patches. For example, Nagpal et al. (2019) use the distribution of Gleason scores across all patches as an input feature vector to a KNN to generate a patient score, and Couture et al. (2018) aggregate patch probabilities into a quantile function which is then used by an SVM to generate a patient level class. Methods that assign patches to discrete classes or clusters can simply use the majority class to label the WSI (Muhammad et al., 2019) or adopt a RNN to generate a single prediction from a sequence of patches (Bychkov et al., 2018).

End-to-end methods that learn features directly from the image data and allow probabilities to be associated with individual patches can be used to uncover new information how morphology is related to outcome. For example, Courtiol et al. (2019) were able to show that regions associated with stroma, inflammation, cellular diversity, and vacuolization were important in predicting survival in mesothelioma patients, and Mobadersany et al. (2018) showed that microvascular proliferation and increased cellularity is associated with poorer outcome in glioma patients. Prediction heatmaps may also allow researchers to uncover patterns of tumour heterogeneity and could be used to guide tissue extraction for genomics and proteomics assays. Deep learning survival models are, therefore, of great interest to cancer researchers as well as to pathologists and oncologists.

Table 6: Overview of survival models for disease prognosis. Note: (✓) indicates the code is publicly available and the link is provided in their respective paper.

Reference	Cancer types	Application	Method	Dataset
Zhu et al. (2017b)	Multi-Cancers	Loss function based on survival time	Raw pixel values of downsampled patches used as feature vectors; 10 clusters identified using K-means clustering. Deep survival models are trained for each cluster separately. Significant clusters are identified and corresponding scores are fed into final WSI classifier	TCIA-NLST, TCGA-LUSC, TCGA-GBM
Bychkov et al. (2018)	Colorectal	5 year disease specific survival	Extracted features using pre-trained VGG-16. Used RNN to generate WSI prediction from tiles	Private set - TMAs from 420 patients
Couture et al. (2018) (✓)	Breast	Prediction of tumour grade, ER status, PAM50 intrinsic subtype, histologic subtype and risk of recurrence score	Pre-trained VGG-16 model. Aggregate features over 800×800 regions to predict class for each patch, then frequency distribution of classes input to SVM to combine regions to predict TMA class	TMA cores (Private-1203 cases)
Mobadersany et al. (2018) (✓)	Brain	Time to event modelling	CNN integrated with a Cox proportional hazards model to predict patient outcomes using histology and genomic biomarkers. Calculate median risk for each ROI, then average 2 highest risk regions	TCGA-LGG, TCGA-GBM (1,061 WSIs)
Courtiol et al. (2019)	Mesothelioma	Loss function based on survival time	Pre-trained ResNet50 extracts features from 10000 tiles. 1-D convolutional layer generates score for each tile. 10 highest and lowest scores fed into MLP classifier for WSI prediction	MESOPATH/MESOBANK (private set-2,981 WSIs), TCGA validation set (56 WSIs)
Geessink et al. (2019)	Colorectal	Dichotomized tumour/stromal ratios	CNN based patch classifier trained to identify tissue components. Calculate tumour-stroma ratio for manually defined hot-spots	Private set-129 WSIs
Kather et al. (2019) (✓)	Colorectal	Dichotomized stromal score	VGG-19 based patch classifier trained to identify tissue component. Calculate HR for each tissue component using mean activation. Combine components with $HR > 1$ to give a “deep stromal score”	NCT-CRC-HE-100k; TCGA-READ, TCGA-COAD
Muhammad et al. (2019)	Liver ICC	HRs of clusters compared	Unsupervised method to cluster tiles using autoencoder. WSI assigned to cluster corresponding to majority of tiles	Private set - 246 ICC H&E WSIs
Nagpal et al. (2019)	Prostate	Gleason scoring	Trained Inception-V3 network to predict Gleason score on labeled patches. Then calculate % patches with each grade on the WSI and use result as a low dimensional feature vector input to k-NN classifier	TCGA-PRAD and private dataset
Qaiser et al. (2019a)	Lymphoma	Generate 4 DPC categories	Multi-task CNN model for simultaneous cell detection and classification, followed by digital proximity signature (DPS) estimation	Private set-32 IHC WSIs
Tang et al. (2019)	Multi-Cancers	Dichototomized survival time (<= 1 year and > 1 year)	A capsule network is trained using a loss function that combines a reconstruction loss, margin loss ans Cox loss. The mean of all patch-level survival predictions is calculated to achieve a final patient-level survival prediction.	TCGA-GBM and TCGA-LUSC
Veta et al. (2019)	Breast	Predict mitotic score & PAM50 proliferation score	Multiple methods from challenge teams	TUPAC 2016
Yamamoto et al. (2019)	Prostate	Predict accuracy of prostate cancer recurrence	Deep autoencoders trained at different magnifications and weighted non-hierarchical clustering, followed by SVM classifier to predict the short-term biochemical recurrence of prostate cancer	Private set - 15,464 WSI’s

Table 7: Summary of publicly available databases in computational histopathology.

Dataset / Year	Cancer types	Goal	Images / Cases (train+test)	Annotation	Link
ICPR 2012 (Cireşan et al., 2013)	Breast	Mitosis detection	50 (35+15)	Pixel-level annotation of mitotic cells	`http://ludo17.free.fr/mitos_2012/`
AMIDA 2013 (Veta et al., 2015)	Breast	Mitosis detection	23 (12+11)	Centroid pixel of mitotic cells	`http://amida13.isi.uu.nl/`
ICPR 2014 (Cireşan et al., 2013)	Breast	Mitosis detection	2112 (2016+96)	Centroid pixel of mitotic cells	`https://mitos-atypia-14.grand-challenge.org/`
GLAS 2015 (Sirinukunwattana et al., 2017)	Colon	Gland segmentation	165 (85+80)	Glandular boundaries	`https://warwick.ac.uk/fac/sci/dcs/research/tia/glascontest/`
TUPAC 2016 (Veta et al., 2019)	Breast	tumour proliferation based on mitosis counting & molecular data + two auxiliary tasks	821 (500+321) + (73/34)	Proliferation scores & ROI of mitotic cells	`http://tupac.tue-image.nl/`
HER2 Scoring 2016 (Qaiser et al., 2018)	Breast	HER2 scoring in breast cancer WSIs	86 (52+28)	HER2 score on whole-slide level	`https://warwick.ac.uk/fac/sci/dcs/research/tia/her2contest/`
BreakHis 2016 (Spanhol et al., 2015)	Breast	Breast cancer detection	82 (7909 patches)	WSL benign vs. malignant annotation	`https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/`
CRCHisto 2016 (Sirinukunwattana et al., 2016)	Colon	Nuclei detection & classification	100	29,756 nuclei centres + out of which 22,444 with associated class labels	`https://warwick.ac.uk/fac/sci/dcs/research/tia/data/crchistolabelednucleihe`
CAMELYON16 (Ehteshami Bejnordi et al., 2017)	Breast	Breast cancer metastasis detection	400 (270+130)	Contour of cancer locations	`https://camelyon16.grand-challenge.org/`
CAMELYON17 (Bandi et al., 2018)	Breast	Breast cancer metastasis detection & pN-stage prediction	1000 (500+500)	Contour of cancer locations + patient level score	`https://camelyon17.grand-challenge.org/`
MoNuSeg 2018 (Kumar et al., 2019)	Multi-Cancers	Nuclei segmentation	44(30+14)	22,000+7000 nuclear boundary annotations	`https://monuseg.grand-challenge.org/Home/`
PCam 2018 (Veeling et al., 2018)	Breast	Metastasis detection	3,27,680 patches	Patch-level binary label	`https://github.com/basveeling/pcam`
TNBC 2018 (Naylor et al., 2018)	Breast	Nuclei segmentation	50	4022 pixel-level annotated nuclei	`https://github.com/PeterJackNaylor-/DRFNS`
BACH 2018 (Aresta et al., 2019)	Breast	Breast cancer classification	500 (400+100)	Image-wise & pixel-level annotations	`https://iciar2018-challenge.grand-challenge.org/Home/`
BreastPathQ 2018 (Akbar et al., 2019)	Breast	tumour cellularity	96 (69+25) WSIs	3,700 patch-level tumour cellularity score	`https://breastpathq.grand-challenge.org/`
Post-NAT-BRCA (Martel et al., 2019)	Breast	tumour cellularity	96 WSIs	Nuclei, patch and patient level annotations	`https://doi.org/10.7937-/TCIA.2019.4YIBTJNO`
CoNSeP 2019 (Graham et al., 2019b)	Colon	Nuclei segmentation and classification	41	24,319 pixel-level annotated nuclei	`https://warwick.ac.uk/fac/sci/dcs-/research/tia/data/hovernet/`
CRAG 2019 (Graham et al., 2019a)	Colon	Gland segmentation	213 (173+40)	Gland instance-level ground truth	`https://warwick.ac.uk/fac/sci/dcs-/research/tia/data/mildnet/`
LYON 2019 (Swiderska-Chadaj et al., 2019)	Multi-Cancers	Lymphocyte detection	83 WSIs	171,166 lymphocytes in 932 ROIs were annotated	`https://lyon19.grand-challenge.org/Home/`
NCT-CRC-HE-100k (Kather et al., 2019)	Colon	Tissue classification	1,00,000 patches(86 WSIs)+7,180 patches(25 WSIs)	Patch-label for nine class tissue classification	`https://zenodo.org/record/1214456#.XffRa3VKhhE`
ACDC-LungHP 2019 (Li et al., 2018b)	Lung	Detection and classification of lung cancer subtypes	(150 + 50) WSI's	Contour of cancer locations + image-level cancer subtype scores	`https://acdc-lunghp.grand-challenge.org/`
Dataset of segmented nuclei 2020 (Hou et al., 2020)	Multi-Cancers	Nuclei segmentation	5,060 (WSI's) + 1,356 (image patches)	Patch level labels (1,356 patches) + 5 billion pixel-level nuclei labels	`https://wiki.cancerimagingarchive.net/display/DOI/Dataset+of+Segmented+Nuclei+in+Hematoxylin+and+Eosin+Stained+Histopathology+Images`
TCGA (TCGA)	Multi-Cancers	Multiple	—-	—-	`https://portal.gdc.cancer.gov/`
TCIA (TCIA)	Multi-Cancers	Multiple	—-	—-	`https://www.cancerimagingarchive.net/`

5. Discussion and future trends

5.1. Effect of deep learning architectures on task performance

In most applications, standard architectures (e.g., *VGGNet* (Simonyan and Zisserman, 2014), *InceptionNet* (Szegedy et al., 2015, 2016), *ResNet* (He et al., 2016), *MobileNet* (Howard et al., 2017), *DenseNet* (Huang et al., 2017)) can be directly employed, and custom networks should only be used if it is impossible to transform the inputs into a suitable format for the given architecture, or the transformation may cause significant information loss that may affect the task performance. For instance, if the scanner pixel scale does not match with the powers of two for nuclei segmentation, custom neural networks with varying image sizes (e.g., 71 × 71) can be utilized (Saha et al., 2017). The most standard architectures are exhaustively tested by many, where their pitfalls, convergence behaviour and weaknesses are well documented in the literature. Unlike the previous, the custom network design choices such as the type of pooling - performed for spatial dimensionality reduction, sizes of the convolutional filters, inclusion of residual connections and/or any other blocks (e.g., Squeeze-and-Excite modules (Hu et al., 2018b) or inception modules (Szegedy et al., 2016)) are left for the researchers to explore and can be critical to the performance of the network. In general, it is recommended to use larger convolutional filters if the input size is large, skip connections in segmentation tasks, and batch normalization for faster convergence and to obtain better performance (Van Eycke et al., 2018). Ideally, any change made to a standard architecture should be thoroughly explained and reasoned, and the performance improvements should be verified through ablation experiments or comparative studies between custom and conventional architectures.

Pre-trained networks are widely employed but although pre-training is known to improve convergence speed significantly, it might not always lead to a better performance compared to a network trained from scratch, given enough time for convergence (Liu et al., 2017; He et al., 2019). In pre-trained networks, natural scene image databases (e.g., ImageNet) are commonly used, and it is possible that the learned feature representations may not be accurate for histopathology images. Given a training dataset with few images, training only a few of the last decision layers (i.e., freezing the initial layers), using a nonlinear decision layer (i.e., composed of one or more hidden layers with nonlinear activations), using regularization techniques such as weight decay and dropout with ratio $p \in [0.2, 0.5]$ are recommended to avoid overfitting (Tabibu et al., 2019; Valkonen et al., 2019).

Until very recently, traditional image processing techniques such as density-based models or handcrafted features were competitive against CNNs in digital histopathology (Sharma et al., 2017), however, as neural networks have become more and more capable, state-of-the-art results in digital pathology have overwhelmingly come from CNN based methods. It is, however, not entirely clear how much of this increase can be attributed to the most recent advancements in neural networks, as opposed to proper validation, data mining and processing practices, or the general familiarity of researchers with DNN.

As such comparative studies have yet to exist for digital histopathology, we examined various public histopathology challenges (Refer, Table 7) to assess the impact of the architecture, and found that in many tasks, the specific architecture was not a determining factor in the task objective outcome. For instance, in BACH challenge on breast histology images (Aresta et al., 2019), the winning entry won the 1^{st} place (with a 14% margin compared to the next best entry) despite using significantly less data without any ensembling networks and a smaller contextual window (i.e., the patch size of the input image), while employing a hard example mining scheme, which made it possible for a network to learn from few examples to converge faster and to avoid overfitting. The winning entry from a MoNuSeg (multi-organ nucleus segmentation) challenge (Kumar et al., 2019) employed a UNet without any post-processing step (e.g., watershed transformation or morphological operations to separate nuclei), whereas, the participants using cascaded UNets, ResNets and feature pyramid networks (FPN) or DenseNets consistently scored lower. The winning entry for a TUPAC (tumour proliferation rate estimation challenge) used a hard negative mining technique with a modified ResNet with 6 or 9 residual blocks, and an SVM (support vector machine) for the decision (feature aggregation) layer for mitosis detection (Veta et al., 2019), beating architectures including GoogleNet, UNet and VGGNet. The winning entry for a CAMELYON16 challenge, including the detection of lymph node metastases (Ehteshami Bejnordi et al., 2017) utilized an ensemble of two GoogleNet networks that has given superior results compared to pathologists with time constraints. In contrast, the same network architecture without any stain standardization or data augmentation achieved 10% and 7% lower in free-response receiver operator characteristic curve (FROC) and area under curve (AUC) metrics, respectively. Three out of five of the top results used GoogleNet architecture with 7 million parameters (22 layers), and the second best entry employed a ResNet with 44 million parameters (101 layers). Results from a subsequent CAMELYON17 challenge involving detection of cancer metastases in lymph nodes and lymph node status classification (Bandi et al., 2018) suggest that using ensemble networks may help self-correct predictions by a suitable form of voting between ensemble networks. In this challenge, the top entry achieved around 2% better in quadratic weighted Cohen's kappa metric (Cohen, 1960) over the second best entry. The top two entries both used ResNet-101 networks, where the top entry used an ensemble of three, and the second best participant used a single network with image resolution four times smaller than the first entry, with about four times the patch size (960×960 versus 256×256 pixels).

It is noteworthy that all of the "*winning networks*" for the various challenges described above were invented on or before the year 2015, whereas challenge dates vary from 2016 to 2019. While challenge scores are not necessarily indicative of the use case performance, as the challenge participants tend to heavily fine-tune their models to achieve the highest possible score, these results indirectly indicate that simpler networks can still prevail, provided that appropriate training practices are applied for the specific problem at hand.

5.2. Challenges in histopathology image analysis

Standard DL architectures require their inputs (e.g., images) in a specific format with certain spatial dimensions. Furthermore, these architectures are generally designed for RGB images, whereas in digital histopathology, working with images in grayscale, HSV or HED color spaces may be desirable for a specific application. Converting images between color spaces, resizing images to fit into GPU memory, quantizing images from a higher bit representation into a lower one, deciding the best resolution for the application at hand and tiling, are some of the choices researchers need to make that will lead to varying degrees of information loss. A reasonable data processing strategy aims to achieve minimal information loss while utilizing architectures to their maximal capacity.

In most applications, it is inevitable that input images will need to be tiled or resized. Memory and computational constraints also make it necessary to find a balance between the required context and the magnification and, as CNNs learn more quickly from smaller images, one should not use images larger than the required context. The optimum trade-off between field of view (FOV) and resolution will depend on the application; for example classifying ductal carcinoma (DCIS) requires a large context to capture morphology, whilst for nuclei detection it is common to use the highest possible power as the required context is as small as one nucleus. In some cases, both high resolution and large FOV are required, for example in cellularity assessment a high power is needed to differentiate between malignant and benign nuclei and a larger FOV is needed to provide the context (Akbar et al., 2019). A considerable amount of work has been done to combine low and high-resolution inputs in making better decisions in various forms and problems (Li et al., 2019b; Chang et al., 2017; Shujun Wang et al., 2019; Li et al., 2018a). However, it is still unclear that these methods are more effective in segmentation tasks compared to selecting a single "best fit" resolution (Seth et al., 2019).

Image pre- and post-processing can be used to boost the task performance of a DL model. In addition to standard preprocessing practices (e.g., resizing an input image and normalization, noise removal, morphological operations to smooth the segmentation masks), preprocessing can also be used to eliminate the need for computationally costly post-processing steps such as iterative refinement of boundaries of segmented regions using conditional random fields Xu et al. (2016). Post-processing techniques can also be used to iteratively refine the model outputs to marginally improve the task objective. Methods based on CRFs are commonly employed to refine boundaries in segmentation tasks (e.g., nuclei) for better delineation of structure boundaries (Qu et al., 2019). Post-processing can also be used for bootstrapping, where the trained model is used for selecting *hard* examples from an unseen test set in which the model underperforms. Then, the model is trained with a subset of original training data and the hard examples obtained from the post-processing step.

This form of post-processing is especially useful in selecting a small subset of data from the majority class to prevent class imbalance, or balance the foreground and background samples (i.e., hard negative mining), and is applicable in many tasks including multiple instance learning or segmentation (Li et al., 2019c; Kwok, 2018). In digital histopathology, hard negative mining is generally used for sampling the normal, or the healthy tissue to avoid over- or under-sampling different types of background regions.

5.3. Quality of training and validation data

The success of DL depends on the availability of high-quality training sets to achieve the desired predictive performance (Madabhushi and Lee, 2016; Bera et al., 2019; Niazi et al., 2019).

It is evident from this survey that a vast majority of methods are based on fully-supervised learning. Obtaining a well-curated data set is, however, often expensive and requires significant manual expertise to obtain clean and accurate annotations. There will always be variability between pathologists so ideally the inter-observer agreement should be quantified (Bulten et al., 2020; Akbar et al., 2019; Seth et al., 2019) and if possible, a consensus between pathologists reached (Veta et al., 2015). Some attempts have been made to generate additional annotated data by using alternative techniques like data augmentation (Tellez et al., 2019a), image synthesis (Hou et al., 2019a) and crowdsourcing (Albarqouni et al., 2016), but it is not yet clear that they are appropriate for digital pathology. In some cases, it is possible to acquire additional information to provide definitive ground truth labels, for example, cytokeratin-stained slides were used to resolve diagnostic uncertainty in the CAMELYON16 challenge (Ehteshami Bejnordi et al., 2017). It is important for researchers to understand how labels are generated and to have some measure of label accuracy.

One way to increase model robustness and improve generalization ability is to include diversity in the training data such as images from multiple scan centres (Campanella et al., 2019), images containing heterogeneous tissue types (Hosseini et al., 2019) with variations in staining protocols (Bulten et al., 2019). For instance, Campanella et al. (2019) trained their DL model on an extensive training set containing more than 15,000 patients of various cancer types, obtained across 45 countries. The authors achieved an excellent performance of AUC greater than 0.98 for three histology task, which demonstrates the importance of a large diverse dataset on model performance. With an increase in the number of well-curated open-source datasets hosted by the Cancer Genome Atlas (TCGA), the Cancer Imaging Archive (TCIA) and various Biomedical Grand Challenges (Refer, Table 7), it is increasingly possible to test methods on a standard benchmark dataset. There is, however, a need for more clinically relevant datasets which capture the complexity of real clinical tasks. The expansion of the breast cancer metastases dataset, CAMELYON16, to CAMELYON17 provides a good illustration of how much larger datasets are needed to assess an algorithm in a more meaningful clinical context (Litjens et al., 2018); in CAMELYON16 399 WSIs from 2 centres were made available but slides containing only isolated tumour cells were excluded and only slide level labels were provided; in CAMELYON17 an additional 1000 WSIs were added from 500 patients and five centres and the total dataset grew to 2.95 terabytes. Even this large dataset does not capture the scale of the clinical task where patients may have multiple WSIs from many more lymph nodes (the CAMELYON set excludes patients with > 5 dissected nodes), and it also excludes patients who have undergone neoadjuvant therapy which is known to adversely affect classification accuracy (Campanella et al., 2019).

As the number of clinical centres adopting a fully digital workflow increases, it is likely that the expectation will be that all digital pathology models should be trained and tested on large, clinically relevant datasets. Making such large datasets of WSIs and associated clinical data available publicly poses significant challenges and one way of addressing this may be to move away from the current approach of moving data to the model, and instead, to create mechanisms for researchers (and companies) to move the training and testing of models to the data. A recent example of this was the DREAM mammography challenge (DREAM, 2016), where only a small subset of data was released to allow developers to test software, and developers then had to submit docker containers to a central server to access the primary dataset for training and testing.

5.4. Model interpretability

In recent years, DL based methods have achieved near human-level performance in many different histology applications (Campanella et al., 2019; Noorbakhsh et al., 2019; Coudray et al., 2018), however, the main issue with DL models is that they are generally regarded as a "black box". Interpretability is less important when networks are carrying out tasks such as mitosis detection or nuclear pleomorphism classification since a pathologist can readily validate performance by visual inspection. Survival models based on a small number of image features that are familiar to pathologists may, therefore, be more acceptable to clinicians than end-to-end deep survival models where it is difficult to understand how a particular prediction is made (Holzinger et al., 2017). Consequently, several *explainable* AI systems (Samek et al., 2017; Chen et al., 2019) have been developed, which attempt to gain deeper insights into the working of DL models. In histology, interpretability of DL models has been addressed by using visual attention maps (Huang and Chung, 2019; BenTaieb and Hamarneh, 2018), saliency maps (Tellez et al., 2019b), heatmaps (Paschali et al., 2019) and image captioning (Zhang et al., 2019; Weng et al., 2019) techniques. These methods aim to highlight discriminative evidence locations in WSIs by providing pathologists with more clinically interpretable results. For instance, Zhang et al. (2019) presented a biologically inspired multimodal DL model capable of visualizing learned representations to produce rich interpretable predictions using network visual attention maps. Furthermore, their model also learns to generate diagnostic reports based on natural language descriptions of histologic findings in a way understandable to a pathologist. Such multimodal models trained on metadata (such as pathology images, clinical reports and genomic sequences) have the potential to offer reliable diagnosis, strong generalizability and objective second opinions, while simultaneously encouraging consensus in routine clinical histopathology practices.

One of the most overlooked issue with current DL models is the vulnerability to adversarial attacks (Papernot et al., 2017). Several recent studies (Finlayson et al., 2019; Jungo and Reyes, 2019; Ma et al., 2019) have demonstrated that DL system can be compromised by carefully designed adversarial examples, i.e., even small imperceptible perturbations can deceive neural networks in predicting wrong outputs with high certainty. This behaviour has raised concerns in successful real-time integration of these DL systems in critical applications like face recognition (Sharif et al., 2016), autonomous driving (Eykholt et al., 2017) and medical diagnosis (Ma et al., 2019).

Uncertainty maps have been used to identify the failure points of neural networks, and to increase the model interpretability (DeVries and Taylor, 2018; Jungo and Reyes, 2019). In histopathology, uncertainty estimates can also be used to identify rare findings in a slide (e.g., locating lymphoma through high uncertainty regions given a breast cancer metastasis classifier), or as a signal for human interference in labeling the low confidence regions (e.g., active learning), which the network is uncertain (Raczkowski et al., 2019; Graham et al., 2019a).

5.5. *Clinical translation*

There has been a rapid growth in artificial intelligence (AI) research applied to medical imaging, and its potential impact has been demonstrated by applications which include detection of breast cancer metastasis in lymph nodes (Steiner et al., 2018), interpreting chest X-rays (Nam et al., 2018), detecting brain tumours in MRI (Kamnitsas et al., 2016), detecting skin cancers (Esteva et al., 2017), diagnosing diseases in retinal images (Gulshan et al., 2016), and so on. Despite this impressive array of applications, the real and impactful deployment of AI in clinical practice still has a far way to go.

The main challenges and potential implications in transforming AI technologies from research to clinical use are as follows. First, the major bottleneck is the regulatory and privacy concern in getting ownership of the patient data such as images and personal health records (Bera et al., 2019; Kelly et al., 2019). This makes it challenging to train, develop and test safe AI solutions for clinical use. Furthermore, the comparison of DL algorithms in an objective manner is challenging due to variability in design methodologies, which are specifically targeted for a small group of populations. To make fair comparisons, the AI models need to be tested on the same independent test set, which represents the same target population with similar performance metrics. Second, most AI algorithms suffer from inapplicability outside of the training domain, algorithmic bias and can be easily fooled by adversarial attacks (Kelly et al., 2019) or by the inclusion of disease subtypes not considered during training. These issues can be partly addressed by developing "interpretable" AI systems (Liu et al., 2018; Rudin, 2019) which provide a reliable measure of model confidence and also generalization to different multi-cohort datasets. Developing human-centred AI models that can meaningfully represent clinical knowledge and provide a clear explanation for model prediction to facilitate improved interactions with clinicians and machines is of paramount importance. Finally, the algorithms need to be integrated into the clinical workflow. This may be the biggest challenge as very few hospitals have made the significant investment required to implement a fully digital workflow, which means that microscope slides are not routinely scanned. Transitioning to a digital workflow does, however, result in significant improvements in turnaround time and cost savings (Hanna et al., 2019), and this is helping to drive increased adoption of digital pathology. If the above challenges are taken into consideration while designing AI solutions, then they are most likely to be transformational in routine patient health care system.

6. Conclusions

In this survey, we have presented a comprehensive overview of deep neural network models developed in the context of computational histopathology image analysis. The availability of large-scale whole-slide histology image databases and recent advancements in technology have triggered the development of complex deep learning models in computational pathology. From the survey of over 130 papers, we have identified that the automatic analysis of histopathology images has been tackled by different deep learning perspectives (e.g., supervised, weakly-supervised, unsupervised and transfer learning) for a wide variety of histology tasks (e.g., cell or nuclei segmentation, tissue classification, tumour detection, disease prediction and prognosis), and has been applied to multiple cancer types (e.g., breast, kidney, colon, lung). The categorization of methodological approaches presented in this survey acts as a reference guide to current techniques available in the literature for computational histopathology. We have also discussed the critical analysis of deep learning architectures on task performance, along with the importance of training data and model interpretability for successful clinical translation. Finally, we have outlined some open issues and future trends for the progress of this field.

Conflict of interest

ALM is co-founder and CSO of Pathcore. CS and OC have no conflicts.

Acknowledgment

This research is funded by: Canadian Cancer Society (grant number 705772); National Cancer Institute of the National Institutes of Health [grant number U24CA199374-01]; Canadian Institutes of Health Research.

References

Agarwalla, A., Shaban, M., Rajpoot, N.M., 2017. Representation-aggregation networks for segmentation of multi-gigapixel histology images. arXiv preprint arXiv:1707.08814 .

Akbar, S., Martel, A.L., 2018. Cluster-based learning from weakly labeled bags in digital pathology. Machine Learning for Health (ML4H) Workshop, NeurIPS 2018 .

Akbar, S., Peikari, M., Salama, S., Panah, A.Y., Nofech-Momes, S., Martel, A.L., 2019. Automated and manual quantification of tumour cellularity in digital slides for tumour burden assessment. Scientific Reports 9, 14099.

Albarqouni, S., Baur, C., Achilles, F., Belagiannis, V., Demirci, S., Navab, N., 2016. Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images. IEEE Transactions on Medical Imaging 35, 1313–1321.

Amgad, M., Elfandy, H., Hussein, H., Atteya, L.A., Elsebaie, M.A., Abo Elnasr, L.S., Sakr, R.A., Salem, H.S., Ismail, A.F., Saad, A.M., et al., 2019. Structured crowdsourcing enables convolutional segmentation of histology images. Bioinformatics 35, 3461–3467.

Aresta, G., Araújo, T., Kwok, S., Chennamsetty, S.S., Safwan, M., Alex, V., Marami, B., Prastawa, M., Chan, M., Donovan, M., Fernandez, G., Zeineh, J., Kohl, M., Walz, C., Ludwig, F., Braunewell, S., Baust, M., Vu, Q.D., To, M.N.N., Kim, E., Kwak, J.T., Galal, S., Sanchez-Freire, V., Brancati, N., Frucci, M., Riccio, D., Wang, Y., Sun, L., Ma, K., Fang, J., Kone, I., Boulmane, L., Campilho, A., Eloy, C., Polónia, A., Aguiar, P., 2019. BACH: Grand challenge on breast cancer histology images. Medical Image Analysis 56, 122–139.

Artieres, T., et al., 2010. Neural conditional random fields, in: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, pp. 177–184.

Arvaniti, E., Fricker, K.S., Moret, M., Rupp, N., Hermanns, T., Fankhauser, C., Wey, N., Wild, P.J., Rueschoff, J.H., Claassen, M., 2018. Automated gleason grading of prostate cancer tissue microarrays via deep learning. Scientific Reports 8.

Awan, R., Koohbanani, N.A., Shaban, M., Lisowska, A., Rajpoot, N., 2018. Context-aware learning using transferable features for classification of breast cancer histology images, in: International Conference Image Analysis and Recognition, pp. 788–795.

Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B.E., Lee, B., Paeng, K., Zhong, A., et al., 2018. From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. IEEE Transactions on Medical Imaging 38, 550–560.

Beck, A.H., Sangoi, A.R., Leung, S., Marinelli, R.J., Nielsen, T.O., van de Vijver, M.J., West, R.B., van de Rijn, M., Koller, D., 2011. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. Science Translational Medicine 3, 108ra113.

Bejnordi, B.E., Litjens, G., Timofeeva, N., Otte-Höller, I., Homeyer, A., Karssemeijer, N., van der Laak, J.A., 2015. Stain specific standardization of whole-slide histopathological images. IEEE Transactions on Medical Imaging 35, 404–415.

Bejnordi, B.E., Mullooly, M., Pfeiffer, R.M., Fan, S., Vacek, P.M., Weaver, D.L., Herschorn, S., Brinton, L.A., van Ginneken, B., Karssemeijer, N., et al., 2018. Using deep convolutional neural networks to identify and classify tumor-associated stroma in diagnostic breast biopsies. Modern Pathology 31, 1502.

Bejnordi, B.E., Zuidhof, G., Balkenhol, M., Hermsen, M., Bult, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J., 2017. Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images. Journal of Medical Imaging 4, 044504.

de Bel, T., Hermsen, M., Kers, J., van der Laak, J., Litjens, G., 2019. Stain-transforming cycle-consistent generative adversarial networks for improved segmentation of renal histopathology, in: International Conference on Medical Imaging with Deep Learning, pp. 151–163.

de Bel, T., Hermsen, M., Smeets, B., Hilbrands, L., van der Laak, J., Litjens, G., 2018. Automatic segmentation of histopathological slides of renal tissue using deep learning, in: Medical Imaging 2018: Digital Pathology, p. 1058112.

BenTaieb, A., Hamarneh, G., 2016. Topology aware fully convolutional networks for histology gland segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 460–468.

BenTaieb, A., Hamarneh, G., 2017. Adversarial stain transfer for histopathology image analysis. IEEE Transactions on Medical Imaging 37, 792–802.

BenTaieb, A., Hamarneh, G., 2018. Predicting cancer with a recurrent visual attention model for histopathology images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 129–137.

Bera, K., Schalper, K.A., Rimm, D.L., Velcheti, V., Madabhushi, A., 2019. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. Nature Reviews Clinical Oncology 16, 703–715.

Bokhorst, J., Pinckaers, H., van Zwam, P., Nagtegaal, I., van der Laak, J., Ciompi, F., 2019. Learning from sparsely annotated data for semantic segmentation in histopathology images, in: Proceedings of the 2nd International Conference on Medical Imaging with Deep Learning, pp. 84–91.

Brieu, N., Meier, A., Kapil, A., Schoenmeyer, R., Gavriel, C.G., Caie, P.D., Schmidt, G., 2019. Domain adaptation-based augmentation for weakly supervised nuclei detection. arXiv preprint arXiv:1907.04681 .

Bulten, W., Bándi, P., Hoven, J., van de Loo, R., Lotz, J., Weiss, N., van der Laak, J., van Ginneken, B., Hulsbergen-van de Kaa, C., Litjens, G., 2019. Epithelium segmentation using deep learning in H&E-stained prostate specimens with immunohistochemistry as reference standard. Scientific Reports 9, 864.

Bulten, W., Litjens, G., 2018. Unsupervised prostate cancer detection on H&E using convolutional adversarial autoencoders. arXiv preprint arXiv:1804.07098 .

Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., van der Laak, J., Hulsbergen-van de Kaa, C., Litjens, G., 2020. Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study. The Lancet Oncology .

Bychkov, D., Linder, N., Turkki, R., Nordling, S., Kovanen, P.E., Verrill, C., Walliander, M., Lundin, M., Haglund, C., Lundin, J., 2018. Deep learning based tissue analysis predicts outcome in colorectal cancer. Scientific Reports 8, 3395.

Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Silva, V.W.K., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J., 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine 25, 1301–1309.

Campanella, G., Silva, V.W.K., Fuchs, T.J., 2018. Terabyte-scale deep multiple instance learning for classification and localization in pathology. arXiv preprint arXiv:1805.06983 .

Chang, H., Han, J., Zhong, C., Snijders, A.M., Mao, J.H., 2017. Unsupervised transfer learning via multiscale convolutional sparse coding for biomedical applications. IEEE Transactions on Pattern Analysis and Machine Intelligence 40, 1182–1194.

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K., 2019. This looks like that: deep learning for interpretable image recognition, in: Advances in Neural Information Processing Systems, pp. 8928–8939.

Chen, H., Qi, X., Yu, L., Dou, Q., Qin, J., Heng, P.A., 2017a. DCAN: Deep contour-aware networks for object instance segmentation from histology images. Medical Image Analysis 36, 135–146.

Chen, H., Wang, X., Heng, P.A., 2016a. Automated mitosis detection with deep regression networks, in: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), pp. 1204–1207.

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017b. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence 40, 834–848.

Chen, T.Q., Li, X., Grosse, R.B., Duvenaud, D.K., 2018. Isolating sources of disentanglement in variational autoencoders, in: Advances in Neural Information Processing Systems, pp. 2610–2620.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P., 2016b. Infogan: Interpretable representation learning by information maximizing generative adversarial nets, in: Advances in Neural Information Processing Systems, pp. 2172–2180.

Chennamsetty, S.S., Safwan, M., Alex, V., 2018. Classification of breast cancer histology image using ensemble of pre-trained neural networks, in: International Conference Image Analysis and Recognition, pp. 804–811.

Cheplygina, V., de Bruijne, M., Pluim, J.P., 2019. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical Image Analysis 54, 280–296.

Cho, H., Lim, S., Choi, G., Min, H., 2017. Neural stain-style transfer learning using GAN for histopathological images. arXiv preprint arXiv:1710.08543 .

Ciga, O., Chen, J., Martel, A., 2019. Multi-layer domain adaptation for deep convolutional networks, in:

Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data. Springer, pp. 20–27.

Ciompi, F., Geessink, O., Bejnordi, B.E., De Souza, G.S., Baidoshvili, A., Litjens, G., Van Ginneken, B., Nagtegaal, I., Van Der Laak, J., 2017. The importance of stain normalization in colorectal tissue classification with convolutional networks, in: IEEE 14th International Symposium on Biomedical Imaging, pp. 160–163.

Cireşan, D., Giusti, A., Gambardella, L.M., Schmidhuber, J., 2012. Deep neural networks segment neuronal membranes in electron microscopy images, in: Advances in Neural Information Processing Systems, pp. 2843–2851.

Cireşan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J., 2013. Mitosis detection in breast cancer histology images with deep neural networks, in: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 411–418.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37–46.

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

Courtiol, P., Maussion, C., Moarii, M., Pronier, E., Pilcer, S., Sefta, M., Manceron, P., Toldo, S., Zaslavskiy, M., Le Stang, N., et al., 2019. Deep learning-based classification of mesothelioma improves prediction of patient outcome. Nature Medicine 25, 1519–1525.

Couture, H.D., Williams, L.A., Geradts, J., Nyante, S.J., Butler, E.N., Marron, J., Perou, C.M., Troester, M.A., Niethammer, M., 2018. Image analysis with deep learning to predict breast cancer grade, er status, histologic subtype, and intrinsic subtype. NPJ Breast Cancer 4, 30.

Cruz-Roa, A., Basavanhally, A., González, F., Gilmore, H., Feldman, M., Ganesan, S., Shih, N., Tomaszewski, J., Madabhushi, A., 2014. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks, in: Medical Imaging 2014: Digital Pathology, p. 904103.

Cruz-Roa, A., Gilmore, H., Basavanhally, A., Feldman, M., Ganesan, S., Shih, N., Tomaszewski, J., Madabhushi, A., González, F., 2018. High-throughput adaptive sampling for whole-slide histopathology image analysis (HASHI) via convolutional neural networks: Application to invasive breast cancer detection. PloS one 13, e0196828.

Cruz-Roa, A., Gilmore, H., Basavanhally, A., Feldman, M., Ganesan, S., Shih, N.N., Tomaszewski, J., González, F.A., Madabhushi, A., 2017. Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent. Scientific Reports 7, 46450.

Dai, J., He, K., Sun, J., 2016. Instance-aware semantic segmentation via multi-task network cascades, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3150–3158.

DeVries, T., Taylor, G.W., 2018. Leveraging uncertainty estimates for predicting segmentation quality. arXiv preprint arXiv:1807.00502 .

Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T., 1997. Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence 89, 31–71.

Ding, H., Pan, Z., Cen, Q., Li, Y., Chen, S., 2019. Multi-scale fully convolutional network for gland segmentation using three-class classification. Neurocomputing .

Dov, D., Kovalsky, S.Z., Cohen, J., Range, D.E., Henao, R., Carin, L., 2019. A deep-learning algorithm for thyroid malignancy prediction from whole slide cytopathology images. arXiv preprint arXiv:1904.12739 .

DREAM, 2016. The Digital Mammography DREAM Challenge. `https://www.synapse.org/#!Synapse:syn4224222/wiki/`.

Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J.A.W.M., , the CAMELYON16 Consortium, 2017. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA 318, 2199–2210.

Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J.A.W.M., Hermsen, M., Manson, Q.F., Balkenhol, M., Geessink, O., Stathonikos, N., van Dijk, M.C., Bult, P., Beca, F., Beck, A.H., Wang, D., Khosla, A., Gargeya, R., Irshad, H., Zhong, A., Dou, Q., Li, Q., Chen, H., Lin, H.J., Heng, P.A., Haß, C., Bruni, E., Wong, Q., Halici, U., Öner, M.Ü., Cetin-Atalay, R., Berseth, M., Khvatkov, V., Vylegzhanin, A., Kraus, O., Shaban, M., Rajpoot, N., Awan, R., Sirinukunwattana, K., Qaiser, T., Tsang, Y.W., Tellez, D., Annuscheit, J., Hufnagl, P., Valkonen, M., Kartasalo, K., Latonen, L., Ruusuvuori, P., Liimatainen, K., Albarqouni, S., Mungal, B., George, A., Demirci, S., Navab, N., Watanabe, S., Seno, S., Takenaka, Y., Matsuda, H., Ahmady Phoulady, H., Kovalev, V., Kalinovsky, A., Liauchuk, V., Bueno, G., Fernandez-Carrobles, M.M., Serrano, I., Deniz, O., Racoceanu, D., Venâncio, R., 2017. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA 318, 2199.

Epstein, J., Allsbrook, W.J., Amin, M., Egevad, L., 2005. The 2005 international society of urological pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma. Am J Surg Pathol 29, 1228–1242.

Ertosun, M.G., Rubin, D.L., 2015. Automated grading of gliomas using deep learning in digital pathology images: A modular approach with ensemble of convolutional neural networks, in: AMIA Annual Symposium Proceedings, p. 1899.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D., 2017. Robust physical-world attacks on deep learning models. arXiv preprint arXiv:1707.08945 .

Finlayson, S.G., Bowers, J.D., Ito, J., Zittrain, J.L., Beam, A.L., Kohane, I.S., 2019. Adversarial attacks on medical machine learning. Science 363, 1287–1289.

Gadermayr, M., Gupta, L., Appel, V., Boor, P., Klinkhammer, B.M., Merhof, D., 2019a. Generative adversarial networks for facilitating stain-independent supervised & unsupervised segmentation: A study on kidney histology. IEEE Transactions on Medical Imaging .

Gadermayr, M., Gupta, L., Klinkhammer, B.M., Boor, P., Merhof, D., 2019b. Unsupervisedly training GANs for segmenting digital pathology with automatically generated annotations, in: Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning, pp. 175–184.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2016. Domain-adversarial training of neural networks. The Journal of Machine Learning Research 17, 2096–2030.

Gao, Z., Wang, L., Zhou, L., Zhang, J., 2017. Hep-2 cell image classification with deep convolutional neural networks. IEEE Journal of Biomedical and Health Informatics 21, 416–428.

Gecer, B., Aksoy, S., Mercan, E., Shapiro, L.G., Weaver, D.L., Elmore, J.G., 2018. Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks. Pattern Recognition 84, 345–356.

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

Gidaris, S., Singh, P., Komodakis, N., 2018. Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728 .

Girshick, R., 2015. Fast R-CNN, in: The IEEE International Conference on Computer Vision (ICCV).

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep learning. MIT press.

Graham, S., Chen, H., Gamper, J., Dou, Q., Heng, P.A., Snead, D., Tsang, Y.W., Rajpoot, N., 2019a. MILD-Net: Minimal information loss dilated network for gland instance segmentation in colon histology images. Medical Image Analysis 52, 199–211.

Graham, S., Vu, Q.D., Raza, S.E.A., Azam, A., Tsang, Y.W., Kwak, J.T., Rajpoot, N., 2019b. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Medical Image Analysis 58, 101563.

Gu, F., Burlutskiy, N., Andersson, M., Wilén, L.K., 2018. Multi-resolution networks for semantic segmentation in whole slide images, in: Computational Pathology and Ophthalmic Medical Image Analysis, pp. 11–18.

Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al., 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 316, 2402–2410.

Guo, Z., Liu, H., Ni, H., Wang, X., Su, M., Guo, W., Wang, K., Jiang, T., Qian, Y., 2019. A fast and refined cancer regions segmentation framework in whole-slide breast pathological images. Scientific Reports 9, 882.

Gurcan, M.N., Boucheron, L.E., Can, A., Madabhushi, A., Rajpoot, N.M., Yener, B., 2009. Histopathological image analysis: a review.

Halicek, M., Shahedi, M., Little, J.V., Chen, A.Y., Myers, L.L., Sumer, B.D., Fei, B., 2019. Head and neck cancer detection in digitized whole-slide histology using convolutional neural networks. Scientific Reports 9, 1–11.

Han, Z., Wei, B., Zheng, Y., Yin, Y., Li, K., Li, S., 2017. Breast cancer multi-classification from histopathological images with structured deep learning model. Scientific Reports 7, 4172.

Hanna, M.G., Reuter, V.E., Samboy, J., England, C., Corsale, L., Fine, S.W., Agaram, N.P., Stamelos, E., Yagi, Y., Hameed, M., Klimstra, D.S., Sirintrapun, S.J., 2019. Implementation of digital pathology offers clinical and operational increase in efficiency and cost savings. Archives of Pathology & Laboratory Medicine 143, 1545–1555.

Hariharan, B., Arbeláez, P., Girshick, R., Malik, J., 2014. Simultaneous detection and segmentation, in: European Conference on Computer Vision, pp. 297–312.

He, K., Girshick, R., Dollár, P., 2019. Rethinking Imagenet pre-training , 4918–4927.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M.M., Mohamed, S., Lerchner, A., 2017. beta-VAE: Learning basic visual concepts with a constrained variational framework, in: ICLR.

Ho, D.J., Yarlagadda, D.V., D'Alfonso, T.M., Hanna, M.G., Grabenstetter, A., Ntiamoah, P., Brogi, E., Tan, L.K., Fuchs, T.J., 2019. Deep multi-magnification networks for multi-class breast cancer image segmentation. arXiv preprint arXiv:1910.13042 .

Holzinger, A., Biemann, C., Pattichis, C.S., Kell, D.B., 2017. What do we need to build explainable ai systems for the medical domain? arXiv preprint arXiv:1712.09923 .

Hosseini, M.S., Chan, L., Tse, G., Tang, M., Deng, J., Norouzi, S., Rowsell, C., Plataniotis, K.N., Damaskinos, S., 2019. Atlas of digital pathology: A generalized hierarchical histological tissue type-annotated database for deep learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 11747–11756.

Hou, L., Agarwal, A., Samaras, D., Kurc, T.M., Gupta, R.R., Saltz, J.H., 2019a. Robust histopathology image analysis: To label or to synthesize?, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8533–8542.

Hou, L., Gupta, R., Van Arnam, J.S., Zhang, Y., Sivalenka, K., Samaras, D., Kurc, T.M., Saltz, J.H., 2020. Dataset of segmented nuclei in hematoxylin and eosin stained histopathology images of 10 cancer types. arXiv preprint arXiv:2002.07913 .

Hou, L., Nguyen, V., Kanevsky, A.B., Samaras, D., Kurc, T.M., Zhao, T., Gupta, R.R., Gao, Y., Chen, W., Foran, D., et al., 2019b. Sparse autoencoder for unsupervised nucleus detection and representation in histopathology images. Pattern Recognition 86, 188–200.

Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H., 2015. Efficient multiple instance convolutional neural networks for gigapixel resolution image classification. arXiv preprint arXiv:1504.07947 , 7.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 .

Hu, B., Tang, Y., Eric, I., Chang, C., Fan, Y., Lai, M., Xu, Y., 2018a. Unsupervised learning for cell-level visual representation in histopathology images with generative adversarial networks. IEEE Journal of Biomedical and Health Informatics 23, 1316–1328.

Hu, J., Shen, L., Sun, G., 2018b. Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708.

Huang, Y., Chung, A., 2019. CELNet: Evidence localization for pathology images using weakly supervised learning. arXiv preprint arXiv:1909.07097 .

Ilse, M., Tomczak, J., Welling, M., 2018. Attention-based deep multiple instance learning, in: Proceedings of the 35th International Conference on Machine Learning, pp. 2127–2136.

Irshad, H., Oh, E.Y., Schmolze, D., Quintana, L., Collins, L., Tamimi, R.M., Beck, A.H., 2017. Crowdsourcing scoring of immunohistochemistry images: Evaluating performance of the crowd and an automated computational method. Scientific Reports 7, 43286.

Janowczyk, A., Basavanhally, A., Madabhushi, A., 2017. Stain normalization using sparse autoencoders (stanosa): Application to digital pathology. Computerized Medical Imaging and Graphics 57, 50–61.

Jia, Z., Huang, X., Eric, I., Chang, C., Xu, Y., 2017. Constrained deep weak supervision for histopathology image segmentation. IEEE Transactions on Medical Imaging 36, 2376–2388.

Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution, in: European Conference on Computer Vision, pp. 694–711.

Jungo, A., Reyes, M., 2019. Assessing reliability and challenges of uncertainty estimations for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 48–56.

Kamnitsas, K., Ferrante, E., Parisot, S., Ledig, C., Nori, A.V., Criminisi, A., Rueckert, D., Glocker, B., 2016. Deepmedic for brain tumor segmentation, in: International workshop on Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries, pp. 138–149.

Kandemir, M., Hamprecht, F.A., 2015. Computer-aided diagnosis from weak supervision: A benchmarking study. Computerized Medical Imaging and Graphics 42, 44–50.

Kapil, A., Wiestler, T., Lanzmich, S., Silva, A., Steele, K., Rebelatto, M., Schmidt, G., Brieu, N., 2019. DASGAN - joint domain adaptation and segmentation for the analysis of epithelial regions in histopathology PD-L1 images. CoRR abs/1906.11118.

Kashif, M.N., Raza, S.E.A., Sirinukunwattana, K., Arif, M., Rajpoot, N., 2016. Handcrafted features with convolutional neural networks for detection of tumor cells in histology images, in: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), pp. 1029–1032.

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

Katzman, J.L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y., 2018. Deepsurv: Personalized treatment recommender system using a cox proportional hazards deep neural network. BMC Medical Research Methodology 18.

Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G., King, D., 2019. Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine 17, 195.

Khan, A.M., Rajpoot, N., Treanor, D., Magee, D., 2014. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. IEEE Transactions on Biomedical Engineering 61, 1729–1738.

Kingma, D.P., Welling, M., 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 .

Kong, B., Wang, X., Li, Z., Song, Q., Zhang, S., 2017. Cancer metastasis detection via spatially structured deep network, in: International Conference on Information Processing in Medical Imaging, pp. 236–248.

Krause, J., Johnson, J., Krishna, R., Fei-Fei, L., 2017. A hierarchical approach for generating descriptive image paragraphs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 317–325.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, pp. 1097–1105.

Kumar, N., Sethi, A., and Others, 2019. A multi-organ nucleus segmentation challenge. IEEE Transactions on Medical Imaging , 1–1.

Kwok, S., 2018. Multiclass classification of breast cancer in whole-slide images, in: International Conference Image Analysis and Recognition, pp. 931–940.

Lafarge, M.W., Pluim, J.P., Eppenhof, K.A., Moeskops, P., Veta, M., 2017. Domain-adversarial neural networks to address the appearance variability of histopathology images, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 83–91.

Lahiani, A., Gildenblat, J., Klaman, I., Albarqouni, S., Navab, N., Klaiman, E., 2019. Virtualization of tissue staining in digital pathology using an unsupervised deep learning approach, in: European Congress on Digital Pathology, pp. 47–55.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.

Lee, B., Paeng, K., 2018. A robust and effective approach towards accurate metastasis detection and pn-stage classification in breast cancer, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 841–850.

Lee, C.H., Yoon, H.J., 2017. Medical big data: promise and challenges. Kidney Research and Clinical Practice 36, 3.

Li, C., Wang, X., Liu, W., Latecki, L.J., Wang, B., Huang, J., 2019a. Weakly supervised mitosis detection in breast histopathology images using concentric loss. Medical Image Analysis 53, 165–178.

Li, J., Li, W., Gertych, A., Knudsen, B.S., Speier, W., Arnold, C.W., 2019b. An attention-based multi-resolution model for prostate whole slide image classification and localization. arXiv preprint arXiv:1905.13208 .

Li, J., Speier, W., Ho, K.C., Sarma, K.V., Gertych, A., Knudsen, B.S., Arnold, C.W., 2018a. An EM-based semi-supervised deep learning approach for semantic segmentation of histopathological images from radical prostatectomies. Computerized Medical Imaging and Graphics 69, 125–133.

Li, M., Wu, L., Wiliem, A., Zhao, K., Zhang, T., Lovell, B.C., 2019c. Deep instance-level hard negative mining model for histopathology images. arXiv preprint arXiv:1906.09681 .

Li, Y., Ping, W., 2018. Cancer metastasis detection with neural conditional random field.

Li, Y., Qi, H., Dai, J., Ji, X., Wei, Y., 2017. Fully convolutional instance-aware semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2359–2367.

Li, Z., Hu, Z., Xu, J., Tan, T., Chen, H., Duan, Z., Liu, P., Tang, J., Cai, G., Ouyang, Q., et al., 2018b. Computer-aided diagnosis of lung carcinoma using deep learning-a pilot study. arXiv preprint arXiv:1803.05471 .

Liang, Q., Nan, Y., Coppola, G., Zou, K., Sun, W., Zhang, D., Wang, Y., Yu, G., 2018. Weakly supervised biomedical image segmentation by reiterative learning. IEEE Journal of Biomedical and Health Informatics 23, 1205–1214.

Lin, H., Chen, H., Dou, Q., Wang, L., Qin, J., Heng, P.A., 2018. Scannet: A fast and dense scanning framework for metastastic breast cancer detection from whole-slide image, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 539–546.

Lin, H., Chen, H., Graham, S., Dou, Q., Rajpoot, N., Heng, P.A., 2019. Fast Scannet: Fast and dense analysis of multi-gigapixel whole-slide images for cancer metastasis detection. IEEE Transactions on Medical Imaging 38, 1948–1958.

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125.

Litjens, G., Bandi, P., Bejnordi, B.E., Geessink, O., Balkenhol, M., Bult, P., Halilovic, A., Hermsen, M., van de Loo, R., Vogels, R., Manson, Q.F., Stathonikos, N., Baidoshvili, A., van Diest, P., Wauters, C., van Dijk, M., van der Laak, J., 2018. 1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset. GigaScience 7, 1–8.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical Image Analysis 42, 60–88.

Litjens, G., Sánchez, C.I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., Hulsbergen-Van De Kaa, C., Bult, P., Van Ginneken, B., Van Der Laak, J., 2016. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Scientific Reports 6, 26286.

Liu, J., Xu, B., Zheng, C., Gong, Y., Garibaldi, J., Soria, D., Green, A., Ellis, I.O., Zou, W., Qiu, G., 2019. An end-to-end deep learning histochemical scoring system for breast cancer TMA. IEEE Transactions on Medical Imaging 38, 617–628.

Liu, X., Xia, T., Wang, J., Yang, Y., Zhou, F., Lin, Y., 2016. Fully convolutional attention networks for fine-grained recognition. arXiv preprint arXiv:1603.06765 .

Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G.E., Kohlberger, T., Boyko, A., Venugopalan, S., Timofeev, A., Nelson, P.Q., Corrado, G.S., et al., 2017. Detecting cancer metastases on gigapixel pathology images. arXiv preprint arXiv:1703.02442 .

Liu, Y., Kohlberger, T., Norouzi, M., Dahl, G.E., Smith, J.L., Mohtashamian, A., Olson, N., Peng, L.H., Hipp, J.D., Stumpe, M.C., 2018. Artificial intelligence–based breast cancer nodal metastasis detection: Insights into the black box for pathologists. Archives of Pathology & Laboratory Medicine .

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., Lu, F., 2019. Understanding adversarial attacks on deep learning based medical image analysis systems. arXiv preprint arXiv:1907.10456 .

Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E., 2009. A method for normalizing histology slides for quantitative analysis, in: 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 1107–1110.

Madabhushi, A., Lee, G., 2016. Image analysis and machine learning in digital pathology: Challenges and opportunities. Medical Image Analysis 33, 170–175.

Martel, A.L., Nofech-Mozes, S., Salama, S., Akbar, S., Peikari, M., 2019. Assessment of residual breast cancer cellularity after neoadjuvant chemotherapy using digital pathology [data set]. https://doi.org/10.7937/TCIA.2019.4YIBTJNO.

Marzahl, C., Aubreville, M., Bertram, C.A., Gerlach, S., Maier, J., Voigt, J., Hill, J., Klopfleisch, R., Maier, A., 2019. Fooling the crowd with deep learning-based methods. arXiv preprint arXiv:1912.00142 .

Mnih, V., Heess, N., Graves, A., et al., 2014. Recurrent models of visual attention, in: Advances in Neural Information Processing Systems, pp. 2204–2212.

Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D.A., Barnholtz-Sloan, J.S., Vega, J.E.V., Brat, D.J., Cooper, L.A., 2018. Predicting cancer outcomes from histology and genomics using convolutional networks. Proceedings of the National Academy of Sciences 115, E2970–E2979.

Muhammad, H., Sigel, C.S., Campanella, G., Boerner, T., Pak, L.M., Büttner, S., IJzermans, J.N., Koerkamp, B.G., Doukas, M., Jarnagin, W.R., et al., 2019. Unsupervised subtyping of cholangiocarcinoma using a deep clustering convolutional autoencoder, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 604–612.

Mukhopadhyay, S., Feldman, M.D., Abels, E., Ashfaq, R., Beltaifa, S., Cacciabeve, N.G., Cathro, H.P., Cheng, L., Cooper, K., Dickey, G.E., et al., 2018. Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: A multicenter blinded randomized noninferiority study of 1992 cases (pivotal study). The American Journal of Surgical Pathology 42, 39.

Nagpal, K., Foote, D., Liu, Y., Chen, P.H.C., Wulczyn, E., Tan, F., Olson, N., Smith, J.L., Mohtashamian, A., Wren, J.H., et al., 2019. Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. npj Digital Medicine 2, 48.

Nam, J.G., Park, S., Hwang, E.J., Lee, J.H., Jin, K.N., Lim, K.Y., Vu, T.H., Sohn, J.H., Hwang, S., Goo, J.M., et al., 2018. Development and validation of deep learning–based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. Radiology 290, 218–228.

Naylor, P., Laé, M., Reyal, F., Walter, T., 2018. Segmentation of nuclei in histopathology images by deep regression of the distance map. IEEE Transactions on Medical Imaging 38, 448–459.

Niazi, M.K.K., Parwani, A.V., Gurcan, M.N., 2019. Digital pathology and artificial intelligence. The Lancet Oncology 20, e253–e261.

Noorbakhsh, J., Farahmand, S., Soltanieh-ha, M., Namburi, S., Zarringhalam, K., Chuang, J., 2019. Pan-cancer classifications of tumor histological images using deep learning. BioRxiv , 715656.

Noroozi, M., Favaro, P., 2016. Unsupervised learning of visual representations by solving jigsaw puzzles, in: European Conference on Computer Vision, pp. 69–84.

Odena, A., Olah, C., Shlens, J., 2017. Conditional image synthesis with auxiliary classifier gans, in: Proceedings of the 34th International Conference on Machine Learning, pp. 2642–2651.

Ørting, S., Doyle, A., van Hilten, M.H.A., Inel, O., Madan, C.R., Mavridis, P., Spiers, H., Cheplygina, V., 2019. A survey of crowdsourcing in medical image analysis. arXiv preprint arXiv:1902.09159 .

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A., 2017. Practical black-box attacks against machine learning, in: Proceedings of the 2017 ACM on Asia conference on computer and communications security, pp. 506–519.

Paschali, M., Naeem, M.F., Simson, W., Steiger, K., Mollenhauer, M., Navab, N., 2019. Deep learning under the microscope: Improving the interpretability of medical imaging neural networks. arXiv preprint arXiv:1904.03127 .

Peng, J., Bo, L., Xu, J., 2009. Conditional neural fields, in: Advances in Neural Information Processing Systems, pp. 1419–1427.

Pinckaers, H., Litjens, G., 2019. Neural ordinary differential equations for semantic segmentation of individual colon glands. arXiv preprint arXiv:1910.10470 .

Prewitt, J.M., Mendelsohn, M.L., 1966. The analysis of cell images. Annals of the New York Academy of Sciences 128, 1035–1053.

Qaiser, T., Mukherjee, A., Reddy Pb, C., Munugoti, S.D., Tallam, V., Pitkäaho, T., Lehtimäki, T., Naughton, T., Berseth, M., Pedraza, A., et al., 2018. Her-2 challenge contest: a detailed assessment of automated her 2 scoring algorithms in whole slide images of breast cancer tissues. Histopathology 72, 227–238.

Qaiser, T., Pugh, M., Margielewska, S., Hollows, R., Murray, P., Rajpoot, N., 2019a. Digital tumor-collagen proximity signature predicts survival in diffuse large B-cell lymphoma, in: European Congress on Digital Pathology, pp. 163–171.

Qaiser, T., Rajpoot, N.M., 2019. Learning where to see: A novel attention model for automated immunohistochemical scoring. IEEE Transactions on Medical Imaging , 1–1.

Qaiser, T., Tsang, Y.W., Taniyama, D., Sakamoto, N., Nakane, K., Epstein, D., Rajpoot, N., 2019b. Fast and accurate tumor segmentation of histology images using persistent homology and deep convolutional features. Medical Image Analysis 55, 1–14.

Qu, H., Riedlinger, G., Wu, P., Huang, Q., Yi, J., De, S., Metaxas, D., 2019. Joint segmentation and fine-grained classification of nuclei in histopathology images, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI), pp. 900–904.

Qu, H., Wu, P., Huang, Q., Yi, J., Riedlinger, G.M., De, S., Metaxas, D.N., 2019. Weakly supervised deep nuclei segmentation using points annotation in histopathology images, in: International Conference on Medical Imaging with Deep Learning, pp. 390–400.

Quellec, G., Cazuguel, G., Cochener, B., Lamard, M., 2017. Multiple-instance learning for medical image and video analysis. IEEE Reviews in Biomedical Engineering 10, 213–234.

Quiros, A.C., Murray-Smith, R., Yuan, K., 2019. Pathology gan: Learning deep representations of cancer tissue. ArXiv abs/1907.02644.

Raczkowski, Ł., Możejko, M., Zambonelli, J., Szczurek, E., 2019. Ara: accurate, reliable and active histopathological image classification framework with bayesian deep learning. Scientific Reports 9, 1–12.

Rakha, E., El-Sayed, M., Lee, A., Elston, C., Grainge, M., Hodi, Z., Blamey, R., Ellis, I., 2008. Prognostic significance of nottingham histologic grade in invasive breast carcinoma. J Clin Oncol 26, 3153–3158.

Ranzato, M., 2014. On learning where to look. arXiv preprint arXiv:1405.5488 .

Rao, S., 2018. Mitos-rcnn: A novel approach to mitotic figure detection in breast cancer histopathology images using region based convolutional neural networks. arXiv preprint arXiv:1807.01788 .

Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P., 2001. Color transfer between images. IEEE Computer Graphics and Applications 21, 34–41.

Ren, J., Hacihaliloglu, I., Singer, E.A., Foran, D.J., Qi, X., 2018. Adversarial domain adaptation for classification of prostate histopathology whole-slide images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 201–209.

Rivenson, Y., Liu, T., Wei, Z., Zhang, Y., de Haan, K., Ozcan, A., 2019. Phasestain: the digital staining of label-free quantitative phase microscopy images using deep learning. Light: Science & Applications 8, 23.

Romo-Bucheli, D., Janowczyk, A., Gilmore, H., Romero, E., Madabhushi, A., 2016. Automated tubule nuclei quantification and correlation with oncotype DX risk categories in ER+ breast cancer whole slide images. Scientific Reports 6, 32706.

Ronneberger, O., Fischer, P., Brox, T., 2015. UNet: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241.

Rony, J., Belharbi, S., Dolz, J., Ben Ayed, I., McCaffrey, L., Granger, E., 2019. Deep weakly-supervised learning methods for classification and localization in histology images: a survey. arXiv preprint arXiv:1909.03354 .

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1, 206–215.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. International Journal of Computer Vision 115, 211–252.

Sabour, S., Frosst, N., Hinton, G.E., 2017. Dynamic routing between capsules, in: Advances in Neural Information Processing Systems, pp. 3856–3866.

Saha, M., Chakraborty, C., Arun, I., Ahmed, R., Chatterjee, S., 2017. An advanced deep learning approach for Ki-67 stained hotspot detection and proliferation rate scoring for prognostic evaluation of breast cancer. Scientific Reports 7, 3213.

Samek, W., Wiegand, T., Müller, K.R., 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296 .

Sari, C.T., Gunduz-Demir, C., 2019. Unsupervised feature extraction via deep learning for histopathological classification of colon tissue images. IEEE Transactions on Medical Imaging 38, 1139–1149.

Seth, N., Akbar, S., Nofech-Mozes, S., Salama, S., Martel, A.L., 2019. Automated segmentation of DCIS in whole slide images, in: European Congress on Digital Pathology ECDP 2019, pp. 67–74.

Shaban, M., Khurram, S.A., Fraz, M.M., Alsubaie, N., Masood, I., Mushtaq, S., Hassan, M., Loya, A., Rajpoot, N.M., 2019a. A novel digital score for abundance of tumour infiltrating lymphocytes predicts disease free survival in oral squamous cell carcinoma. Scientific Reports 9, 1–13.

Shaban, M.T., Baur, C., Navab, N., Albarqouni, S., 2019b. Staingan: Stain style transfer for digital histological images, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 953–956.

Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M.K., 2016. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition, in: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 1528–1540.

Sharma, H., Zerbe, N., Klempert, I., Hellwich, O., Hufnagl, P., 2017. Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology. Computerized Medical Imaging and Graphics 61, 2–13.

Sharma, S., Kiros, R., Salakhutdinov, R., 2015. Action recognition using visual attention. arXiv preprint arXiv:1511.04119 .

Shen, D., Wu, G., Suk, H.I., 2017. Deep learning in medical image analysis. Annual Review of Biomedical Engineering 19, 221–248.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

Sirinukunwattana, K., e Ahmed Raza, S., Tsang, Y.W., Snead, D.R., Cree, I.A., Rajpoot, N.M., 2016. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE Transactions on Medical Imaging 35, 1196–1206.

Sirinukunwattana, K., Pluim, J.P., Chen, H., Qi, X., Heng, P.A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., Böhm, A., Ronneberger, O., Cheikh, B.B., Racoceanu, D., Kainz, P., Pfeiffer, M., Urschler, M., Snead, D.R., Rajpoot, N.M., 2017. Gland segmentation in colon histology images: The glas challenge contest. Medical Image Analysis 35, 489–502.

Song, Y., Tan, E., Jiang, X., Cheng, J., Ni, D., Chen, S., Lei, B., Wang, T., 2017. Accurate cervical cell segmentation from overlapping clumps in pap smear images. IEEE Transactions on Medical Imaging 36, 288–300.

Song, Y., Zhang, L., Chen, S., Ni, D., Lei, B., Wang, T., 2015. Accurate segmentation of cervical cytoplasm and nuclei based on multiscale convolutional network and graph partitioning. IEEE Transactions on Biomedical Engineering 62, 2421–2433.

Spanhol, F.A., Oliveira, L.S., Petitjean, C., Heutte, L., 2015. A dataset for breast cancer histopathological image classification. IEEE Transactions on Biomedical Engineering 63, 1455–1462.

Stacke, K., Eilertsen, G., Unger, J., Lundström, C., 2019. A closer look at domain shift for deep learning in histopathology. CoRR abs/1909.11575.

Steiner, D.F., MacDonald, R., Liu, Y., Truszkowski, P., Hipp, J.D., Gammage, C., Thng, F., Peng, L., Stumpe, M.C., 2018. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. The American Journal of Surgical Pathology 42, 1636.

Ström, P., Kartasalo, K., Olsson, H., Solorzano, L., Delahunt, B., Berney, D.M., Bostwick, D.G., Evans, A.J., Grignon, D.J., Humphrey, P.A., et al., 2020. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. The Lancet Oncology .

Swiderska-Chadaj, Z., Pinckaers, H., van Rijthoven, M., Balkenhol, M., Melnikova, M., Geessink, O., Manson, Q., Sherman, M., Polonia, A., Parry, J., et al., 2019. Learning to detect lymphocytes in immunohistochemistry with deep learning. Medical Image Analysis 58, 101547.

Symmans, W.F., Peintinger, F., Hatzis, C., Rajan, R., Kuerer, H., Valero, V., Assad, L., Poniecka, A., Hennessy, B., Green, M., et al., 2007. Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy. Journal of Clinical Oncology 25, 4414–4422.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9.

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

Tabibu, S., Vinod, P., Jawahar, C., 2019. Pan-renal cell carcinoma classification and survival prediction from histopathology images using deep learning. Scientific Reports 9, 1–9.

Tang, B., Li, A., Li, B., Wang, M., 2019. Capsurv: Capsule network for survival analysis with whole slide pathological images. IEEE Access 7, 26022–26030.

TCGA, . The cancer genome atlas. `https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga`.

TCIA, . The cancer imaging archive. `https://www.cancerimagingarchive.net/`.

Tellez, D., Balkenhol, M., Otte-Höller, I., van de Loo, R., Vogels, R., Bult, P., Wauters, C., Vreuls, W., Mol, S., Karssemeijer, N., Litjens, G., van der Laak, J., Ciompi, F., 2018. Whole-slide mitosis detection in H&E breast histology using PHH3 as a reference to train distilled stain-invariant convolutional networks. IEEE Transactions on Medical Imaging 37, 2126–2136.

Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.M., Ciompi, F., van der Laak, J., 2019a. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Medical Image Analysis 58, 101544.

Tellez, D., Litjens, G., van der Laak, J., Ciompi, F., 2019b. Neural image compression for gigapixel histopathology image analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence , 1–1.

Tokunaga, H., Teramoto, Y., Yoshizawa, A., Bise, R., 2019. Adaptive weighting multi-field-of-view cnn for semantic segmentation in pathology, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 12597–12606.

Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., Baust, M., Steiger, K., Schlitter, A.M., Esposito, I., Navab, N., 2016. Structure-preserving color normalization and sparse stain separation for histological images. IEEE Transactions on Medical Imaging 35, 1962–1971.

Valkonen, M., Isola, J., Isola, J., Ylinen, O., Muhonen, V., Saxlin, A., Tolonen, T., Nykter, M., Ruusuvuori, P., 2019. Cytokeratin-supervised deep learning for automatic recognition of epithelial cells in breast cancers stained for ER, PR, and Ki-67. IEEE Transactions on Medical Imaging , 1–1.

Van Eycke, Y.R., Balsat, C., Verset, L., Debeir, O., Salmon, I., Decaestecker, C., 2018. Segmentation of glandular epithelium in colorectal tumours to automatically compartmentalise ihc biomarker quantification: A deep learning approach. Medical Image Analysis 49, 35–45.

Vandenberghe, M.E., Scott, M.L., Scorer, P.W., Söderberg, M., Balcerzak, D., Barker, C., 2017. Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer. Scientific Reports 7, 45938.

Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M., 2018. Rotation equivariant cnns for digital pathology, in: International Conference on Medical image computing and computer-assisted intervention, pp. 210–218.

Veta, M., Heng, Y.J., Stathonikos, N., Bejnordi, B.E., Beca, F., Wollmann, T., Rohr, K., Shah, M.A., Wang, D., Rousson, M., Hedlund, M., Tellez, D., Ciompi, F., Zerhouni, E., Lanyi, D., Viana, M., Kovalev, V., Liauchuk, V., Phoulady, H.A., Qaiser, T., Graham, S., Rajpoot, N., Sjöblom, E., Molin, J., Paeng, K., Hwang, S., Park, S., Jia, Z., Chang, E.I.C., Xu, Y., Beck, A.H., van Diest, P.J., Pluim, J.P., 2019. Predicting breast tumor proliferation from whole-slide images: The TUPAC16 challenge. Medical Image Analysis 54, 111–121.

Veta, M., Van Diest, P.J., Willems, S.M., Wang, H., Madabhushi, A., Cruz-Roa, A., Gonzalez, F., Larsen, A.B., Vestergaard, J.S., Dahl, A.B., et al., 2015. Assessment of algorithms for mitosis detection in breast cancer histopathology images. Medical Image Analysis 20, 237–248.

Shujun Wang, Zhu, Y., Yu, L., Chen, H., Lin, H., Wan, X., Fan, X., Heng, P.A., 2019. RMDL: Recalibrated multi-instance deep learning for whole slide gastric image classification. Medical Image Analysis 58, 101549.

Wang, D., Khosla, A., Gargeya, R., Irshad, H., Beck, A.H., 2016a. Deep learning for identifying metastatic breast cancer. arXiv preprint arXiv:1606.05718 .

Wang, H., Roa, A.C., Basavanhally, A.N., Gilmore, H.L., Shih, N., Feldman, M., Tomaszewski, J., Gonzalez, F., Madabhushi, A., 2014. Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. Journal of Medical Imaging 1.

Wang, M., Deng, W., 2018. Deep visual domain adaptation: A survey. Neurocomputing 312, 135–153.

Wang, S., Yao, J., Xu, Z., Huang, J., 2016b. Subtype cell detection with an accelerated deep convolution neural network, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 640–648.

Wang, X., Chen, H., Gan, C., Lin, H., Dou, Q., Tsougenis, E., Huang, Q., Cai, M., Heng, P., 2019. Weakly supervised deep learning for whole slide lung cancer image analysis. IEEE Transactions on Cybernetics , 1–13.

Wei, J.W., Tafe, L.J., Linnik, Y.A., Vaickus, L.J., Tomita, N., Hassanpour, S., 2019. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. Scientific Reports 9, 3358.

Weng, W.H., Cai, Y., Lin, A., Tan, F., Chen, P.H.C., 2019. Multimodal multitask representation learning for pathology biobank metadata prediction. arXiv preprint arXiv:1909.07846 .

Xie, S., Tu, Z., 2015. Holistically-nested edge detection, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1395–1403.

Xie, W., Noble, J.A., Zisserman, A., 2018a. Microscopy cell counting and detection with fully convolutional regression networks. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 6, 283–292.

Xie, Y., Kong, X., Xing, F., Liu, F., Su, H., Yang, L., 2015a. Deep voting: A robust approach toward nucleus localization in microscopy images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 374–382.

Xie, Y., Xing, F., Kong, X., Su, H., Yang, L., 2015b. Beyond classification: structured regression for robust cell detection using convolutional neural network, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 358–365.

Xie, Y., Xing, F., Shi, X., Kong, X., Su, H., Yang, L., 2018b. Efficient and robust cell detection: A structured regression approach. Medical Image Analysis 44, 245–254.

Xing, F., Cornish, T.C., Bennett, T., Ghosh, D., Yang, L., 2019. Pixel-to-pixel learning with weak supervision for single-stage nucleus recognition in Ki-67 images. IEEE Transactions on Biomedical Engineering 66, 3088–3097.

Xing, F., Xie, Y., Yang, L., 2016. An automatic learning-based framework for robust nucleus segmentation. IEEE Transactions on Medical Imaging 35, 550–566.

Xu, B., Liu, J., Hou, X., Liu, B., Garibaldi, J., Ellis, I.O., Green, A., Shen, L., Qiu, G., 2019. Look, investigate, and classify: A deep hybrid attention method for breast cancer classification, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 914–918.

Xu, G., Song, Z., Sun, Z., Ku, C., Yang, Z., Liu, C., Wang, S., Ma, J., Xu, W., 2019. CAMEL: A weakly supervised learning framework for histopathology image segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 10682–10691.

Xu, J., Xiang, L., Liu, Q., Gilmore, H., Wu, J., Tang, J., Madabhushi, A., 2015a. Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images. IEEE Transactions on Medical Imaging 35, 119–130.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y., 2015b. Show, attend and tell: Neural image caption generation with visual attention, in: International Conference on Machine Learning, pp. 2048–2057.

Xu, Y., Li, Y., Liu, M., Wang, Y., Lai, M., Eric, I., Chang, C., 2016. Gland instance segmentation by deep multichannel side supervision, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 496–504.

Xu, Y., Li, Y., Wang, Y., Liu, M., Fan, Y., Lai, M., Chang, E.I., 2017. Gland instance segmentation using deep multichannel neural networks. IEEE Transactions on Biomedical Engineering 64, 2901–2912.

Xu, Y., Zhu, J.Y., Chang, E.I.C., Lai, M., Tu, Z., 2014. Weakly supervised histopathology cancer image segmentation and classification. Medical Image Analysis 18, 591–604.

Yamamoto, Y., Tsuzuki, T., Akatsuka, J., Ueki, M., Morikawa, H., Numata, Y., Takahara, T., Tsuyuki, T., Tsutsumi, K., Nakazawa, R., et al., 2019. Automated acquisition of explainable knowledge from unannotated histopathology images. Nature Communications 10, 1–9.

Yang, L., Zhang, Y., Zhao, Z., Zheng, H., Liang, P., Ying, M.T., Ahuja, A.T., Chen, D.Z., 2018. Boxnet: Deep learning based biomedical image segmentation using boxes only annotation. arXiv preprint arXiv:1806.00593 .

Yi, X., Walia, E., Babyn, P., 2019. Generative adversarial network in medical imaging: A review. Medical Image Analysis , 101552.

Yuan, Y., Failmezger, H., Rueda, O.M., Ali, H.R., Gräf, S., Chin, S.F., Schwarz, R.F., Curtis, C., Dunning, M.J., Bardwell, H., Johnson, N., Doyle, S., Turashvili, G., Provenzano, E., Aparicio, S., Caldas, C., Markowetz, F., 2012. Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. Science Translational Medicine 4, 157ra143.

Zanjani, F.G., Zinger, S., Bejnordi, B.E., van der Laak, J.A., et al., 2018. Histopathology stain-color normalization using deep generative models, in: 1st Conference on Medical Imaging with Deep Learning (MIDL), Amsterdam, The Netherlands.

Zhang, Z., Chen, P., McGough, M., Xing, F., Wang, C., Bui, M., Xie, Y., Sapkota, M., Cui, L., Dhillon, J., et al., 2019. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. Nature Machine Intelligence 1, 236.

Zhao, Z., Lin, H., Chen, H., Heng, P.A., 2019. PFA-ScanNet: Pyramidal feature aggregation with synergistic learning for breast cancer metastasis analysis, in: Medical Image Computing and Computer Assisted Intervention, pp. 586–594.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H., 2015. Conditional random fields as recurrent neural networks, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1529–1537.

Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2017a. Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232.

Zhu, X., Yao, J., Zhu, F., Huang, J., 2017b. Wsisa: Making survival prediction from whole slide histopathological images, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6855–6863.