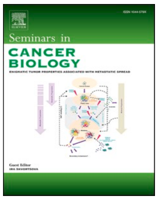
Seminars in Cancer Biology 72 (2021) 226–237

Contents lists available at ScienceDirect

Seminars in Cancer Biology

journal homepage: www.elsevier.com/locate/semcancer



Review

Deep computational pathology in breast cancer

Andrea Duggento [a,*], Allegra Conti [a,*], Alessandro Mauriello [b], Maria Guerrisi [a], Nicola Toschi [a,c]

[a] *Department of Biomedicine and prevention, University of Rome Tor Vergata, Rome, Italy*
[b] *Department of Experimental Medicine, University of Rome Tor Vergata, Rome, Italy*
[c] *A.A. Martinos Center for Biomedical Imaging – Harvard Medical School/MGH, Boston, MA, USA*

ARTICLE INFO

Keywords:
Digital pathology
Deep Learning
Breast cancer
Deep histology
Deep Neural Networks

ABSTRACT

Deep Learning (DL) algorithms are a set of techniques that exploit large and/or complex real-world datasets for cross-domain and cross-discipline prediction and classification tasks. DL architectures excel in computer vision tasks, and in particular image processing and interpretation. This has prompted a wave of disruptingly innovative applications in medical imaging, where DL strategies have the potential to vastly outperform human experts. This is particularly relevant in the context of histopathology, where whole slide imaging (WSI) of stained tissue in conjuction with DL algorithms for their interpretation, selection and cancer staging are beginning to play an ever increasing role in supporting human operators in visual assessments. This has the potential to reduce everyday workload as well as to increase precision and reproducibility across observers, centers, staining techniques and even pathologies. In this paper we introduce the most common DL architectures used in image analysis, with a focus on histopathological image analysis in general and in breast histology in particular. We briefly review how, state-of-art DL architectures compare to human performance on across a number of critical tasks such as mitotic count, tubules analysis and nuclear pleomorphism analysis. Also, the development of DL algorithms specialized to pathology images have been enormously fueled by a number of world-wide challenges based on large, multicentric image databases which are now publicly available. In turn, this has allowed most recent efforts to shift more and more towards semi-supervised learning methods, which provide greater flexibility and applicability. We also review all major repositories of manually labelled pathology images in breast cancer and provide an in-depth discussion of the challenges specific to training DL architectures to interpret WSI data, as well as a review of the state-of-the-art methods for interpretation of images generated from immunohistochemical analysis of breast lesions. We finally discuss the future challenges and opportunities which the adoption of DL paradigms is most likely to pose in the field of pathology for breast cancer detection, diagnosis, staging and prognosis. This review is intended as a comprehensive stepping stone into the field of modern computational pathology for a transdisciplinary readership across technical and medical disciplines.

1. Introduction

The terms Deep Learning (DL) and Neural Network (NN) have become ubiquitous across science and society. NN-based Artificial Intelligence (AI) algorithms have demonstrated superior capabilities with respect to classical AI, especially in those tasks which require complex data integration and decision-making. DL algorithms are a specific subset of Machine Learning (ML) algorithms (in turn a member of the AI family), which are designed to produce a desired (typically predictive) output given a certain input by learning through examples, without explicit human intervention when determining the input features. DL provides the ability to analyze large and heterogeneous data and, if provided with enough training and information, can outperform human experts in many cognitive task. Accordingly, DL has generated a revolution ranging from autonomous car-driving [1] to poker playing [2], through image recognition [3] to automated speech recognition [4], translation and synthesis [5,6]. Overall, DL techniques have proven to be particularly advantageous in image-based tasks (e.g. image recognition, segmentation and classification), and since early 2000s such architectures have been extensively trained on data crowdsourced from the web. The generality and portability embedded in those tools makes the transition to image data from other disciplines virtually effortless, provided a sufficiently large database of labeled data is available for training. In view of the above, it has been recognized that NN-based algorithm are, or soon will become, the standard tool for drug development [7], genome research [8], and medical imaging diagnosis [9,

[*] Corresponding authors.

https://doi.org/10.1016/j.semcancer.2020.08.006
Received 8 April 2020; Received in revised form 13 August 2020; Accepted 13 August 2020
Available online 17 August 2020

10]. Indeed, major medical disciplines such as oncology [11], radiology [12], neurology [13] and cardiology [14] have benefited from DL techniques in terms of detecting aberrations, supporting diagnosis, guiding treatment, predicting outcome and evaluating prognosis [15]. Interestingly, this disruptive, transdisciplinary innovation has not come unexpected, and scientists as well as physicians have been speculating about the technological transfer of AI to biomedical fields [14] since the late ‘80s. However, only recently media coverage and the use of evocative terms such as “the raise of the machine” has amplified [16] to describe the AI-fueled revolution currently underway in biomedicine. The diagnostic capabilities of DL have already outperformed pools of board-certified human experts in terms of detection accuracy [17–19], with additional benefits such as reproducibility and time-efficiency. In addition, it is likely that DL architectures will soon be able to replicate the types of processes, which commonly takes place in the mind of medical practitioners (diagnosis formulation, therapeutic path evaluation, and prognosis prediction) in a human-interpretable way. In view of the above, diagnostic imaging is a natural candidate for the deployment of DL strategies because of (i) the widespread availability of picture archiving and communication systems (PACS) and (ii) the fact that most of image data stored in PACS system is inherently labeled through the existence of a diagnosis. Historically, this has led to a first push of AI-based algorithms implemented as a computer-aided detection (CAD) systems for radiological image analysis employed for detecting e.g. pulmonary nodules [20], intracranial bleeds [21], or breast lesions [22]. Nowadays, however, any type of medical image can access the benefits of NN-based AI tools with little additional overhead. Interestingly, a number of tasks performed by pathologists have several commonalities with diagnostic radiology. The pathologist is often required to undertake extensive searches across a vast number of images – typically within the space of Whole Slide Imaging (WSI) – to extract clinically relevant information and formulate or confirm a diagnosis. In this context, the progressive adoption of certified whole-slide scanners and digital WSI infrastructure (to be contrasted with traditional microscopy [23]) has laid the groundwork for a fruitful adoption of automated AI-based systems [24] in the field of (digital) pathology. It is also worth mentioning that the steep rise of AI techniques in processing and classification of medical images has sparked suspicion and professional concern amongst medical specialists. The main concerns so far are both in the role AI could/should assume in shaping the future role of the professional practice [25], and in the degree of trust which should be placed in a diagnostic framework whose inner workings are (with currently available architectures), inscrutable. However, there is general agreement that AI tools represents more of an opportunity rather than a threat [26,12,27–30]. AI is poised to provide added value to practitioners in performing professional tasks, by e.g. lightening the heavy-lifting workload and hence freeing up resources which could be dedicated to important aspects such as inter-professional interactions, patient-physician relationship, and in general playing a more rewarding role in the improvement of patient care.

Recently, a vast number of DL approaches have been developed to improve decision making based on high-volume, complex healthcare data [31]. In the following section, we discuss the overall architecture of DL methods and how they differ from general ML methods. We will outline several DL approaches suitable for computer-aided detection in breast cancer, as well as several major publicly available data repositories of manually labelled images in breast cancer that are commonly employed for training DL models. As a result, we discuss the insight, difficulties and challenges which have resulted from recently intensified research efforts in these domains in general and in the filed of interpreting WSI in particular.

2. DL architectures

With respect to ML algorithms, in DL algorithms the computational architecture is structured in a multi-layer fashion, and each layer is capable of distilling, extracting and reorganizing information which is passed on from previous layers.

Layers are often composed of many parallel units which perform a single, simple mathematical operation. As it progresses through the layers, the data undergoes successive abstraction processes which extract information. Due to the striking analogy of DL architectures to way mammalian perception or cognitive processes operate, the units are often called ‘artificial neurons’. Of note, the distinction between “artificial NN” and “DL” is often unclear: a NN-based architecture indicates a (not necessary ‘deep’) ML architecture composed of many simple, unioperational units; if the artificial NN is organized in multiple input-output layers, it is termed ‘deep’. With respect to traditional ML systems, NNs and deep NNs can accomplish more sophisticated tasks, and with higher accuracy. On the flipside, they typically require tuning of many more parameters and therefore much more data to learn from. A complete and in-depth introduction to the principles of AI can be found in [32]. The procedure of tuning the parameters by analyzing example data is termed ‘training’. The training data is composed by the input and output data (the latter will be produced by the algorithm once trained). The input–output nature of training data is essential for the machine to learn the generality needed to produce the output given the ‘unseen’ input (input data which was not employed during training and hence to which the machine has never been exposed). In ML applications it is useful to keep part of the original dataset as a ‘test’-set, i.e. a portion of the ground true dataset which is set aside and ‘unexposed’ to the learning process. The test set can be used to obtain an unbiased evaluation of the final performance of the trained algorithm. It should be noted that the amount and quality of training data necessary to achieve optimal performance is also dependent on network architecture as well as so called network ‘capacity’. The latter can be formally defined [33] and generally refers to the amount of complexity in the patterns that the network is able to learn. If the training data is too few with respect to the capacity of the network, or if it contains too many imperfections (e.g. artifacts), the learning process will incur in the so called “over-fitting” situation. In this case, the machine will not “generalize” well, i.e. it will perform poorly on unseen data. Importantly, the quality of training data may be degraded not only by artifacts, but also by ‘contamination’ of information from outside the training dataset. This is commonly referred to as ‘data leakage’. Data leakage might arise in different, more or less subtle forms, including e.g. from biased preprocessing of training data (e.g. samples from two different classes have different probabilities of undergoing distinct preprocessing pipelines), to improper handling/separations of training and test sets (e.g. images of the two sets share some degree of information, such as when biomedical images come from the same subject). For a non-technical discussion on data leakage and other problems that might render biomedical data sub-optimal for machine learning, see [34]. Fig. 1 depicts classical “learning curves” i.e. performance as a function of training data availability, for simplest ML methods, to NN-based methods and deep NN-based methods. The more complex the architecture, the highest (typically) the performance, provided enough training data is employed.

Like any other ML algorithm, deep NNs can learn in a supervised or in an unsupervised way. Supervised learning uses labels (i.e. output data categories or attributes) which are provided available in advance. In unsupervised learning, also called self-organized learning, labels are defined without *a priori* knowledge about the output. As an example, clustering strategies – i.e. grouping observables into groups (called clusters) to minimize intra-group differences and maximize inter-groups differences – are simple forms of unsupervised learning. In [35] a non-technical introduction to supervised and unsupervised learning strategies for biomedical applications as well as other mixed strategies such as semi-supervised, multi-modal and multi-tasking learning, can be found.

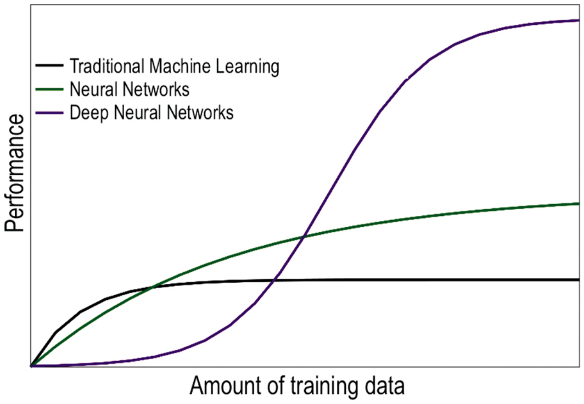


Fig. 1. Example performances in terms of accuracy on the test set as a function of available training data. Comparison between deep NN and traditional ML algorithms.

2.1. *A simple example of image analysis architecture*

In order to introduce some technical nomenclature, and for the benefit of the non-technical reader, before discussing more general classes of architectures, we now illustrate in some detail the inner workings of a simple and popular, network architecture termed Convolutional NN (CNN).

Fig. 2 describes a very simple CNN. Neurons composing the input layer receive the data (see Fig. 2). Further processing then occurs in the 'hidden' layers of the network, which in turn are arranged and connected one to each other. The strength of a connection is called weight and represents the influence that one neuron, or node, has to another [36]. In this type of architecture, the hidden layers are commonly named convolutional, pooling and fully connected layers. A convolutional layer employs convolutional operations, e.g. a multiplication of the input by an array of spatially arranged weight (i.e. a 'kernel'). In particular, each neuron of this layer performs this operation on a spatially clustered, often strictly contiguous, group of the data it receives [37–39]. A pooling (downsampling) layer reduces the spatial dimensions of its input, in order to increase computational performance and have less chance to over-fit the data. Pooling layers may perform different operations, such averaging or extracting a maximum [40]. In so called 'fully connected' layers, all artificial neurons are connected to each other. Often, the first fully connected layer within a network processes the features found by the previous steps of the system and applies weights to produce an output, i.e. in a binary classification task the fully connected layer will estimate the final probabilities for the input to belong to each of the classes. Typically, DL architectures concatenate several hidden layers to extract and transform the information contained in raw data into an output (named 'features') that is not initially visible in the raw data. In this context, the higher performance of DL systems with respect to typical ML algorithms are often due to the large number of hidden layers. While typical ML algorithms rely on one input and one output layer (with no more than one hidden layer between the two) DL systems are often characterized by a larger number of hidden layers: the larger the hidden part of the network, the deeper the learning [41].

2.2. *Most common DL architectures*

Among all the existing strategies, typical/most commonly used DL algorithms are CNN, Recurrent NN (RNN), Restricted Boltzmann Machines (RBM), autoencoders, Adversarial Networks (AN) and Deep Belief Networks (DBN) [42], all characterized by different architectures [43]. Convolutional NNs are DL algorithms with an architecture inspired by the connectivity between neurons in the human Visual Cortex. These networks are composed by at least one convolutional layer as well as pooling and fully connected layers [39]. Convolutional and pooling layers are used to extract features, while the fully connected layers convert these features into a final output. CNN are often used in biomedical applications to recognize features in radiological images [39] through both supervised and unsupervised approaches. Recurrent NN are based on directed connections between layers and nodes which get updates in a discretized fashion at every time increment, or time-step. The ordered and directed architecture of RNNs employs the output from the previous step as input to the current step [44], hence mimicking RNNs mimic temporal dynamic behaviors. This makes RNN suitable to study sequences of inputs, such as timeseries (e.g. biomedical signals). RNN can be trained both in supervised and unsupervised manners. RBM are single-layer, undirected models formed by a visible and an hidden layer with no intra-layer connections between nodes [45]. The algorithm learns thanks to the probability distributions which are associated to the inputs through the interaction of the hidden layer with other units of the network. RBM can be used to implement both supervised and unsupervised learning methods, and they have been employed to e.g. discriminate between healthy subjects from patients affected by a number of pathologies such as e.g. cardiovascular diseases, diabetes, liver diseases [46]. An autoencoder is a type of unsupervised NN composed by visible input and output layers connected through a hidden layer [47]. This latter layer represents the core of the algorithm. An encoder maps the input onto the hidden layer, while a decoder works in the opposite direction, hence reconstructing the original inputs. Autoencoders are commonly used in medicine to label features in radiological images [48]. By AN we usually refer to a combination of a classifying network and a generative network that compete against each other: the generative model is trained to generate synthetic samples similar to some real examples, aiming to render the classifying network (aptly called discriminator) unable to discriminate the synthetic sample from the real examples. AN are most often employed in unsupervised learning strategies. DBN have architectures composed by different inter-connected hidden layers without intra-layer connections between nodes. Supervised and unsupervised DBN are often used as

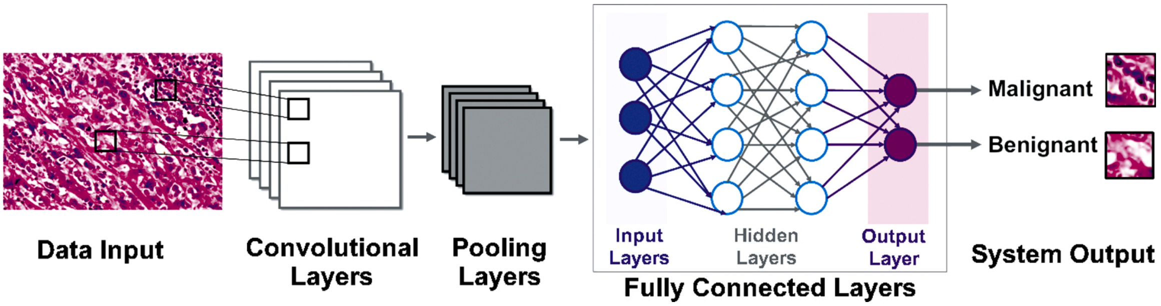


Fig. 2. Typical architecture of a DL NN. The classifier is composed by several layers of neurons (circles) arranged in different visible layers (such as the input and the output of the system) as well as in hidden layers, responsible for the bulk of data processing. The hidden part of the system is, in turn, divided into convolutional, pooling and fully connected layers.

Computer-Aided Diagnosis (CAD) systems [49,50].

2.3. Assembling DL architectures

Deep NNs are commonly designed by combining and rearranging the basic elementary 'bricks' summarized above. While a virtually endless number of configuration can be generated, a few archetypal arrangements can be highlighted, such as CifarNet, a CNN composed by three convolutional and pooling layers and one fully-connected layer [40]; AlexNet [51], another CNN composed by five convolutional layers, three pooling layers, and two fully-connected layers; VGG-net [52], a uniformly designed CNN typically consisting of 16 convolutional layers and 138 millions of parameters, which is often the network of choice for image classification when the size of the training dataset allows its training; GoogLeNet [53], a more complex architecture made of two convolutional layers and two pooling layers connected to nine so called 'inception' modules where the latter are composed by six convolutional layers and one pooling layer each; and the so-called Residual Neural Network (ResNet) [54], which is similar to VGG-net, but includes numerous skipped connections between groups of convolutional layers and is generally faster to train with respect to VGG-net despite the fact that it may be composed by 50 to hundreds of layers. These are prototypical examples of general-purpose image-to-decision machines that can be easily tailored to specific applications by simply retraining on an application-specific trainingset. This process – i.e. the specialization of a previously trained model to a specific dataset – usually goes under the name of 'transfer learning' (see [55] for an introduction to this topic).

3. DL in breast histology

3.1. Introduction to the problem

The preparation of the sample for Breast Histopathology Image Analysis (BHIA) (to be used by a human operator or by a computer software) usually occurs as follows: (i) tissue fixation (to prevent autolysis and putrefaction); (ii) specimen trimming and transfer to cassette; (iii) tissue processing, which involve dehydration, clearing, and embedding; (iv) sectioning and placement on the slide; and finally (v) staining, where the standard staining protocol is haematoxylin and eosin (H&E) staining. Immunohistochemical markers are also often used, e.g. for cancer-subtype classification and hence to support decisions about therapeutic strategies. Traditionally, at this point the whole slide can be directly observed by the histopathologist at a multi-headed microscope to formulate a diagnosis. However, with the advent of digital pathology, more and more often the glass slide is converted into a digital slide for WSI analysis, which can be performed by humans on-screen. Currently, the observation of either glass slides or digital slides by a board-certified human operator is the only recognized way to determine the 'histological truth'. It is well known however that the histological truth exhibits noticeable inter-individual variability, which may depend on differently effective search strategies, better developed eye movements, different cognitive processes, and other perceptual and cognitive factors [56]. It has been shown [57] that the three most influencing (and possibly non-independent) causes for diagnostic inter-individual variability are: (i) subtle differences in professional opinion regarding whether the features met the diagnostic criteria for a specific diagnosis; (ii) not noticing a focal finding; (iii) different diagnostic philosophies on whether purely morphological criteria should be used as opposed to incorporating additional clinical information and/or potential clinical impact of the diagnosis itself. Further factors such as fatigue, stress or variability in emotional states contribute to intra-individual inconsistency in detecting and in the grading of tumors. In view of the above, the possible introduction of AI-guided image processing pipelines, and ultimately of AI-powered diagnostic systems in BHIA, is typically perceived as a potential improvement of medical care efficiency in terms of financial costs and human resources. Maybe more importantly computer guidance would remove the human-related variability in assessing the 'histological truth'.

3.2. Color normalization

The spatial and chromatic distributions of H&E-stained slides depend on a large number of variables itself (such us staining providers, chemical concentration and reactivity, storage conditions, light transmission on tissue), and is compounded with further variability when the slide is digitized due to variability in mechanical and optical properties of the scanners. While the human eye is able to adapt seamlessly to small variation of tone and contrast, DL-based image analyses can be sensitive to color tone distribution shifts [58–60]. In this context, two main approaches have been proposed to color-standardize histological images: stain color deconvolution [61] and template matching [62], both of which perform an image-to-image translation task [63]. Stain color deconvolution makes use of prior knowledge of the color vector of every dye [61] accrued through a manual selection of pixels which represent a specific stain class. This semi-supervised approach has been further automated (see [64] for recent developments). Conversely, template matching attempts to normalize the color space of the source image with reference to the color space of an expertly picked reference template image. Drawbacks and improvements are discussed in [64,58]. Also, very recently more sophisticated color normalization techniques has been developed which make use of DL and NN methods, such as CNN [65] or the so-called self-attentive AN [66,67]. It is however likely that stain normalization will be much less critical when DL algorithms will be trained on massively multicentric data, hence acquiring the ability to extract relevant histomorphic information regardless of color variability. Indeed, it has been already demonstrate that artificial color augmentation during DL training (a procedure which can be though of as also mimicking multicentric-related varaibility) improves the generalization capability of CNN in some histological tasks such as mitotic count and cancer staging [68,69].

3.3. Public databases

The advent of digital pathology has facilitated the organization of computational pathology contests and grand challenges. In turn, this has highlighted the need for publicly available, curated and labeled datasets which can be employed across laboratories for algorithm development. In turn, this has fostered the creation and training of additional DL algorithms, both within or outside the scope of the challenges those databases have been created for. It is also important to point out that such contests and challenges are organized in such a way that the test data in not available to participants, and the evaluation is typically run by a centralized expert panel. This ensures methodological rigor and the absence statistically circular analyses, which would lead to inflated performances. Because of the pivotal role played by those databases in computational pathology research we briefly review past and present resources which are currently available in this realm.

- The MITOS-ICPR12 challenge (2012) [70] was focused on automated mitosis detection in breast tissue. The ground truth was provided by manual annotation by expert pathologists of all mitotic cells. In its first version, the challenge was based on a relatively small amount of data (5 slides in total, 10 annotated microscope high power fields per slide), it not accounted for the inter-subject variability in tissue appearance and staining, and regions from the same slides were included in both the training and testing sets. The dataset was later expanded to include a total of 226 mitoses annotated on 35 high power fields on glass slide at 400× magnification.
- AMIDA13 (2013) [71] was also a challenge in mitosis detection in breast tissue, based on a notably larger dataset as compared to MITOS. It consisted of 23 slides from invasive breast carcinoma patients (12 patients for training, 11 patients held out for testing).

- GLAS (2015) [72], was a gland segmentation challenge based on colon histology images, which was presented at the Medical Image Computing and Computer Assisted Intervention (MICCAI) conference in 2015. Although not directly related to breast cancer, it sparked the development of several CNN architectures, which generally proved to be directly applicable in problems in H&S BHIA. The dataset used in this challenge consists of 165 images derived from 16 H&E stained histological sections of stage T3 or T4 colorectal adenocarcinoma. Each section belongs to a different patient, and since they were processed at different times, the dataset exhibits inter-subject variability both in stain distribution and tissue architecture.
- MITOS-ATYPIA-14 (2014) [73] was a challenge composed of both a mitosis detection contest, and, separately, of an evaluation of nuclear atypia score (assessed through the Nottingham Grading System (NGS) by both senior and junior pathologists), and it was based on an update of the MITOS challenge. The training set contained 284 frames extracted from ×20 magnification and 1136 frames extracted from ×40 magnification.
- CAMELYON16 (2016) [74] was a challenge for automated detection of metastases in WSI of lymph node sections. With respect to previous challenges, the organization of CAMELYON16 marked a steep increment in the volume of data AND consisted of a total of 400 WSIs (including both training and testing data) and an equivalent amount of masks which defined the annotation of metastatic regions. The size of the CAMELYON16 challenge data-set allowed the training of very deep models such as a 22-layer GoogLeNet [75], 16-layer VGG-Net [52], and 101-layer ResNet [54].
- TUPAC16 (2016) [76] was a challenge to predict tumor proliferation from WSI. The challenge consisted of two separate tasks: (i) to predict mitotic scores based on manual labels generated by expert pathologist; (ii) to predict gene expression-based PAM50 (Prosigna-(r)) proliferation scores. The challenge dataset consisted of 500 training and 321 testing breast cancer histopathology WSIs, with a small overlap with the AMIDA13 dataset. Successively, the PATCHED-CAMELYON16 (2017) [77] database was generated by repackaging and streamlining CAMELYON16 data. It consists of 327.680 color images (96 × 96 pixel) where each image has a binary label indicating presence/absence of metastatic tissue. PATCHED-CAMELYON16 provides also a ready-to-train deep NN model which can be executed on a single GPU, and allows very fast algorithm deployment without the burden of SWI management. However, since it was not released in form of a challenge, no independent review of results obtained across the research community is available.
- CAMELYON17 (2017) [78] was a challenge for automatic detection and classification of breast cancer metastases in WSI of histological lymph node sections. With respect to CAMELYON16, which was centered on metastases recognition, CAMELYON17 is focused on the patient-level analysis and hence aims to provide a higher clinical relevance. The challenge consists in merging the information obtain by detection and classification of metastatic sites in multiple slides, each obtain from a surgically removed lymph node, and provide a single outcome in form of a pN-stage (which, according to the so called TNM staging system [79], evaluates whether the cancer has spread to the regional lymph nodes).
- BACH (2018) [80] was a challenge for breast carcinoma detection and labeling from H&E stained microscopy images. The challenge consisted of two separate tasks: (i) automatically labeling H images according to four classes (normal, benign, in situ carcinoma and invasive carcinoma); (ii) performing pixel-wise labeling of the same images (same four classes). Participants were provided with 100 images for each class (for a total of 400 images, each 2048 × 1536 px in size).

4. AI for assessing predetermined clinical criteria

4.1. Clinical tasks

The assessment of histological tumor grade is a key step in prognostic evaluation in oncology. Tumor grading is currently based on visual assessment of the morphological characteristics of tumor tissue. This assessment is semi-quantitative, i.e. the expert pathologist provides various indicators such as the approximate percentage of mitotic cells, or abundance and the staining intensity of tumor tissue, all of which, however, are generated by visual inspection. Various international scientific boards-such as the World Health Organization (WHO),[1] the American Joint Committee on Cancer (AJCC),[2] the European Union (EU), and the Royal College of Pathologists (UK RCPath)[3] – recommend the NGS [81] for tumor grading based on visual inspection of the histological sample. The NGS, estimates tumor grade on a scale from 1 to 3, and is based on the assessment of three morphological features: (i) degree of tubule or gland formation, (ii) mitotic count, and (iii) nuclear pleomorphism. The NGS has a number of advantages: it is simple, inexpensive and provides a proven prognostic value (see [82] for a review). Research targeted towards the intelligent automation of image-based tumour grading has therefore flourished.

4.2. Mitotic count

The MITOS-ICPR12 and the AMIDA13 challenges provided high quality, multiple-observer-labeled data. As such, it generated a research spurt in the field of creating automating mitotic count algorithms. Mitotic count is conventionally performed within an area of $2 \times 2\,\mathrm{mm}^2$ and provides a proxy for tumor aggressiveness. Mitotic cells appear as hyperchromatic objects without a nuclear membrane; further, they usually exhibit specific shapes which are rare in other cells. A major breakthrough in automated mitosis count was delivered by Ciresan and co-workers [83] who capitalizing on their previous experience [84,85] provided a major milestone in AI mitotic count detection. Their method was built by averaging the result of several independently trained CNN feed-forward architectures. Each architecture was composed by 10 to 12 layers. The main innovations were (a) the introduction of max-pooling layers which, at that time, were being investigated in the computer-vision community [86] (feed-forward CNN with max-pooling layers are now considered the *de-facto* standard when building a CNN architecture from scratch) and (b) the reduction of variance by averaging the probabilities of several independently trained CNNs. This approach led them to win both the MITOS-ICPR12 challenge and the AMIDA13 challenge [87]. The proposed method reduced the complexity of previously employed CNNs (both in terms of number of layers and of number of overall parameters), by introducing a combination of CNN and handcrafted features. The workflow consisted in applying extraction of handcrafted features on mitosis candidates via a random forests classifier and evaluating the same patch via a CNN; in case of non-concordance, a second-stage random forests classifier was employed on CNN-derived and handcrafted features; the final decision was made by numerical consensus across all three classifiers [87]. This work demonstrated that an integrated approach made of a classical ontology (distinctions in intensity, shape, texture) and a NN-derived ontology (unintelligible by humans) resulted in superior detection performances as well as in less computationally demanding workflow. While MITOS-ICPR12 and the AMIDA13 challenge are closed for submission, the availability of the dataset continue to spark interest in the development of automatic mitotic count, and the reported detection performances are steadily increasing, although at a lower incremental

[1] www.who.int.

[2] www.facs.org.

[3] www.rcpath.org.

rate. Virtually all of the most successful DL approaches are based on a combination of the two key ingredients: (i) the use of max-pooling CNN layers and (ii) the inclusion of classifiers which use handcrafted features (see for instance [88–93]). In [94], one can find an in-depth review of the most successful approaches presented since the introduction of the MITOS-ICPR2012 challenge. Also, the approach by [95] represents a notable exception. The authors employ a NN architecture called AggNet, which is designed to capitalize on information from “crowdsourced” data. Crowdsourcing is a practice for collecting data from participative online activity of individuals. While it was initially introduced as market research strategy, it can be exploited to recruit large crowds for tedious and time-consuming tasks, especially in the field of visual recognition and labeling. Authors of [95] attempt to understand whether the labels generated from non-expert users (which inevitably lead to noisy annotations) can be employed in a massive scale to train a deep CNN. Their results confirm that training from crowdsourced annotation for mitotic count is robust to noisy labels, opening new perspectives for future channels of information that NN architectures will be trained on.

4.3. Tubules analysis

The morphology of tubules is another proxy for cancer aggressiveness. With cancer progression, the tubules become less organized and deviations from a semi-circular section occurs. Analyzing the structure of the tubules can therefore improve the accuracy of cancer staging as well as of prognosis formulation. However, from an AI perspective, this problem of recognizing and segmenting a tubule is fairly complex. Once the tubule is correctly segmented, features like its shape, area and size can be employed by a downstream classifier. In [96], the authors use a DNN architecture to delineate the tubules by detecting both the margin of the lumen and the external tubule margin. Successively, another CNN detects and counts the nuclei between the two margins, hence extracting an index called tubule formation indicator. This index is then used to predict the oncotype DX test in a cohort of 174 patients. In [97] Janowczy and coauthors discuss and provide a tutorial for tubule detection with DL approaches. While tubule analysis with DL has not attained the status of an independent tool to improve diagnosis accuracy, it is being tackled with extremely deep NN as well as transfer learning [98], and it has been included in a more comprehensive framework for tumor detection and staging.

4.4. Nuclear pleomorphism

The term nuclear ‘pleomorphism’ comprises irregularities in nuclear shape, nuclear size, and changes in chromatin amount and distribution. The presence of large pleomorphisms is an indicator of cancer, and pathologists asses it through a dedicated score which contributes to the formulation of a diagnosis. Still, systematic differences between pathologists in scoring nuclear pleomorphism have been reported [99]. Interestingly, the analysis of nuclear shape deformity may overlap with the task of mitotic count. In detail, analyzing nuclear shape may contribute to mitigating a certain classification bias in mitotic and non-mitotic nuclei classification. For instance authors of [100], used global binary thresholding on blue ratio images to develop a two-phase CNN: phase-1 was used to discriminate between easy, normal, and hard non-mitoses; hard mitoses where then heavily augmented by flipping and rotations before being passed on to phase-2 classification to compensate for class imbalances. An example of an efficient deep NN architecture, based on stacked sparse autoencoder (SSAE) and dedicated to nuclear pleomorphism detection, has been proposed in [101]. The authors showed that deeper SSAE outperform “shallower” architectures in terms of nuclear detection accuracy. Other architectures proposed for nuclear pleomorphism include the ones proposed by authors of [102], which demonstrated that accurate measurements of individual nuclear area as well as regional statistics such as the mean nuclear area can be obtained directly via deep CNN models, hence bypassing the intermediate step of nuclei segmentation. Also, authors of [103] proposed the use of a ‘shape-preserving’ learning approach for automatic nucleus segmentation, where a CNN generates a ‘shape’ probability map which is iteratively improved and successively fed to segmentation algorithm which employ selection-based sparse shape models and local repulsive deformable models.

4.5. Himmunohistochemistry

Immunohistochemistry (IHC) is a staining method which employs antigen-specific antibodies, and it is routinely employed in breast cancer diagnosis. Many IHC staining targets are relevant in breast cancer. A non-exhaustive list includes e-cadherin (used to differentiate ductal from lobular carcinoma), 34βE12, CK8, CK5/6, p120-catenin and β-catenin, calponin, smooth muscle myosin heavy chain, p63, Ki-67, human epidermal growth factor receptor 2 (HER2), hormone receptors and lymph-vascular invasion markers including ERG, CD31, CD34, factor VIII and podoplanin. Also, since the monoclonal antibody trastuzumab (Herceptin) has become available, which provides an effective (albeit expensive) treatment for HER2-positive breast cancer, IHC staining for HER2 has entered common diagnostic practice [104]. For a brief review of current IHC trends in breast cancer see [105]. The scoring method for HER2 IHC is semiquantitative and is based on 4 classes, or scores, called 0, 1+, 2+ and 3+ respectively. Score of 0 and 1+ are considered HER2-negative and correspond to no staining to weak, incomplete, membrane staining or else weak and complete staining in less than 10% of invasive tumor cells. A score of 2+ is considered HER2-equivocal and corresponds to circumferential membrane staining that is incomplete and/or weak/moderate in more than 10% of the invasive tumor cells. A score of 3+ is considered HER2-positive and corresponds to circumferential membrane staining that is complete and intense in a homogeneous and contiguous population in more than 10% of invasive tumor cells [106]. Because of the clinical benefit of an anti-HER2 therapy, high accuracy in identifying HER2+ tumors is crucial. Still, current standards in HER2 scoring are affected by high variability [107–111], with an estimated false positive rate of 4% and a false negative rate as high as 18% [112].

One of the earliest attempt to introduce DL algorithms in HER2 assessment can be found in [113], where authors proposed a DL approach based on CNNs to automatically “score” HER2 (see Fig. 3). The algorithm consisted of 3 convolutional layers followed by fully connected layers, and it significantly outperformed other ML methods such as linear Support Vector Machine (SVM) and Random Forest (RF) models. Further, a blind and independent scoring of some previously scored cases combined with controlling for intra-operator variability by re-scoring a second time after a washout period, demonstrated the ability of DL methods to identify possible misdiagnosis in HER2 staining.

5. From criteria-based clinical assessment to morphological feature extraction, classification, grading and subtyping

The first applications of DL-based methods in pathology were mainly aimed to compute predictors and descriptors based on classical categories for tumor assessment (e.g. mitotic count, tubule analysis, nuclear pleomorphism). More recently however, research efforts have shifted towards a more direct approaches. Since the primary question remains tumor detection and its contouring on WSI, this task has become the natural endpoint the most recently proposed algorithms. In this new paradigm, the AI architecture may be trained to reproduce the evaluation of a panel of expert pathologists about the presence as well as subtyping of abnormal tissue. This means that classical descriptors used in tumor assessment (which are defined and used by human operators) are fully bypassed, and the DL algorithms is tasked with building its own (often inaccessible) descriptors as an integral part of the training process. An example of this approach is provided in [114], who trained several DL architectures using WSI data from 349 estrogen

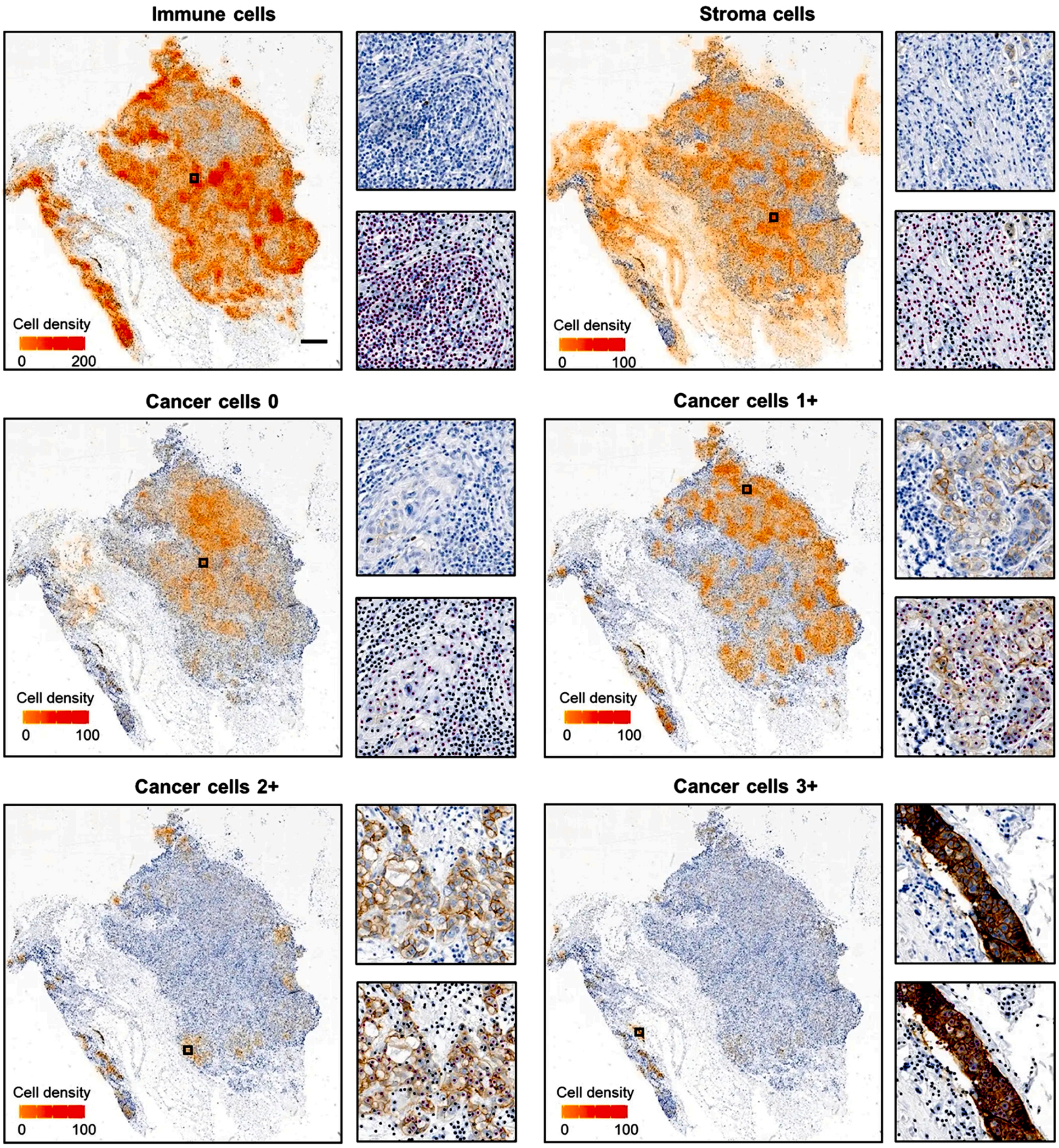


Fig. 3. The HER2 status of invasive breast carcinoma as determined automatically by DL. WSI analysis classifies the percentage of 3+, 2+, 1+ and 0 tumor cells present in the total population. Reproduced with permission from [113].

receptor-positive invasive breast cancer patients collected in multiple sites and digitized on different scanners. The data included annotations by expert pathologist who manual delineated invasive breast cancer regions. The authors demonstrated the ability of their DL method to automatically detect invasive breast cancer from whole slide histopathology images (see Fig. 4 for an example). Of note, the authors of [114] reported that out of three DL architectures (3-layer, 4-layer, and 6-layer ConvNet) models, the 6-layer ConvNet model performed marginally better that the 3-layer ConvNet model, which in turn performed better that the 4-layer model. This clearly indicates a non-trivial relationship between the depth of the DL architecture and its performance. In general, DL-based AI may be able to extract information from H&E stained histological samples that is not accessible to the human operator by visual inspection. Such information could potentially be related to biological variables with strong clinical relevance such as e.g. receptor status or intrinsic tumor subtype. Authors of [115] combined a color normalization technique with the training of a large CNN via transfer learning, demonstrating that it is possible to predict, tumor grade, ER status, PAM50 intrinsic subtype, histologic subtype, and risk of recurrence score from an H&E stained breast tumor tissue microarray. While prediction accuracy varied as a function of tumor grade, the overall message was that combining the “right” DL-architecture with sufficient amount of data for training allows to extract information that typically requires costly RNA-based, multi-gene molecular assays from H&E stained samples alone. While RNA-based genomic tests still provide superior accuracy, DL-based image analysis could therefore be employed for e.g. triaging candidates for genomic testing.

5.1. Weakly supervised systems and external validation

The classical DL approach in WSI analysis has been that of supervised learning: large amount of well-labeled data are employed to train the AI system to reproduce, with the highest possible accuracy, the output commonly produced by humans. In other words, the AI is constrained to

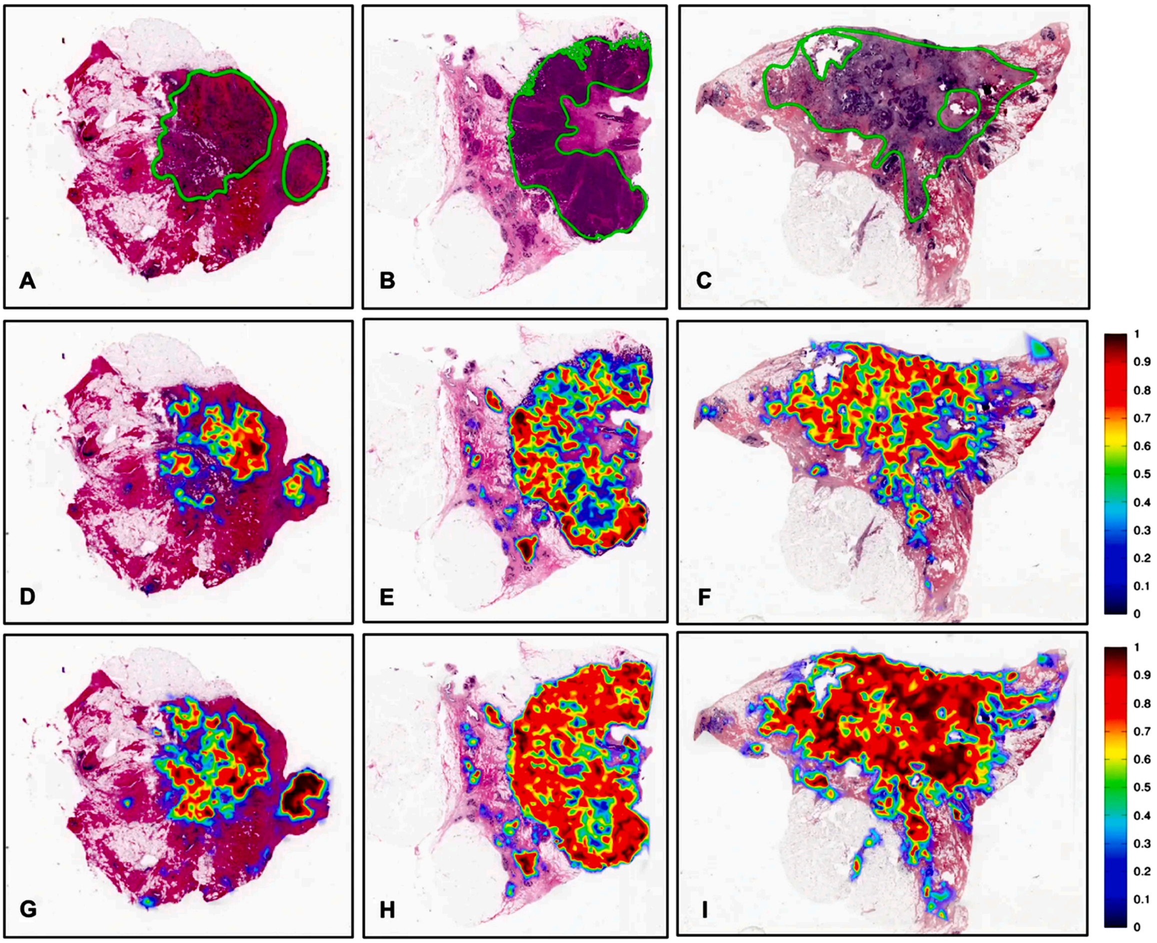


Fig. 4. (A–C) Example WSI and corresponding ground truth annotation (reproduced with permission from [114]. (D–F,G–I): probability maps generated by two different ConvNet classifiers as described in [114].

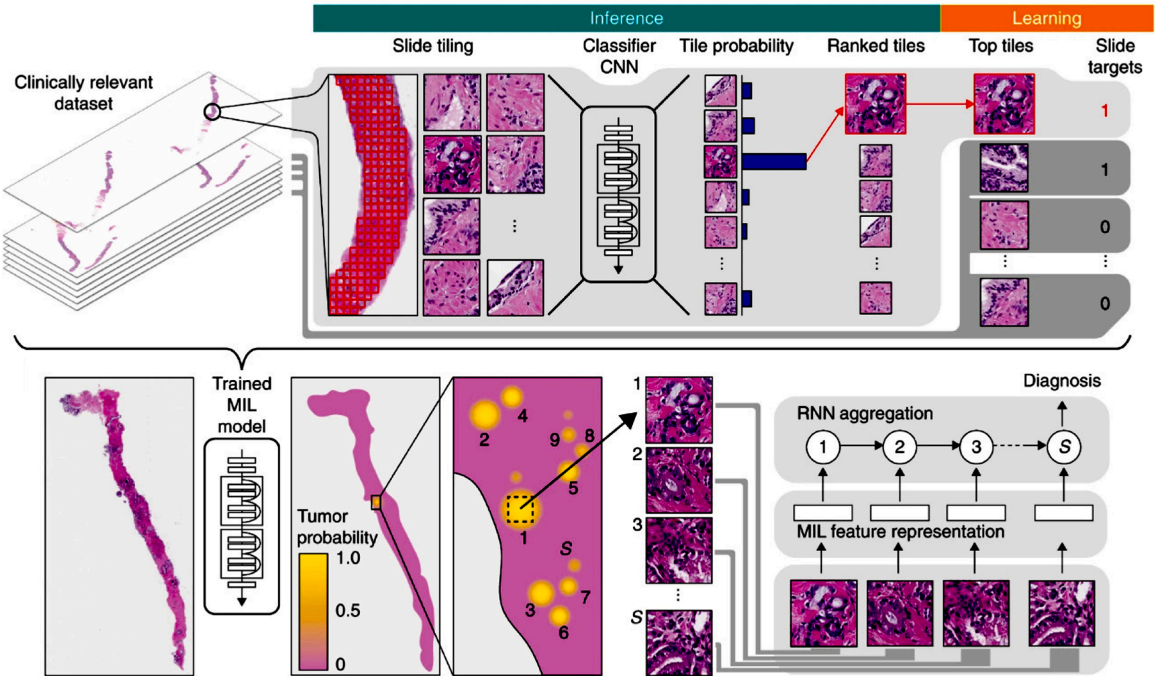


Fig. 5. The MIL training procedure presented in [119]: a deep NN is first trained to output a semantically rich tile-level feature representation. The highest scoring tile in terms of tumor probability is passed to a second stage recurrent NN to predict the final slide-level class. Adapted with permission from [119].

predicting functional categories which are pre-existing in the human mind. In a different approach, the AI can “learn” (i.e. define) its own categories, which would live in the space of its internal-state representation of the inputs, and employ those to predict clinically relevant output. Provided a large enough database of clinical cases is available, this would render tedious and labor intensive manual annotations at the pixel level unnecessary, since the output (e.g. therapeutic response) would be the natural end-point to be predicted by the AI system. Such strategies that circumvent the need of costly and time-consuming hand-labeled training sets are indicated with the umbrella term of ‘weakly supervised’ learning. Early proposers of this approach [116] demonstrated that weakly supervised histopathology segmentation is feasible even without patch-level annotation. The authors employed (for the first time in computational pathology) a learning method termed Multiple Instance Learning (MIL) (see [117] for a review). In the context of breast cancer, this type of DL model was trained to predict response to HER2-targeted therapy given the pre-treatment breast Magnetic Resonance Imaging (MRI) [118]. The same type of approach has been proposed by authors of [119] in the context of digital pathology (not limited to breast cancer). The authors assembled a large database consisting of (i) 24,859 slides from prostate core biopsies, (ii) 9962 slides from a skin lesion dataset and (iii) 9894 slides from breast metastases in lymph nodes. Since manual annotation was not feasible due to the size of the dataset, classical supervised learning could not be employed. Instead, the authors proposed a weakly supervised approach which leverages the readily available slide-level diagnosis (see Fig. 5). In a MIL framework, a deep NN was first trained to output a semantically rich tile-level feature representation. Then, given that a positively diagnosed slice contains at least one positive tile, the tile-level feature representation was used to train a secondary RNN to predict the final class. It should also be noted that, given the large amount of variability originally present in the data, no augmentation techniques were employed. This work [119] likely represents a milestone in computational pathology for several reasons: (1) it demonstrated a that pixel-level annotation can be bypassed, (2) the approach can be generalized to provide multiple diagnoses, inspiring the concept of multi-type and multi-subtype cancer diagnostics, or “pan-cancer” DL diagnosis (the latter term was recently introduced in [120], where 30,000 WSI from 25 primary anatomic sites and 32 cancer subtypes where used to train a ‘consensus’-based DL diagnostic tool), (3) the algorithm has broad tolerance to image quality levels, and it is robust to artifacts introduced during fixation, tissue-processing, slicing or staining; (4) it can also be generalized to extract, from WSI, biomarkers relevant to other clinical questions, such as response to a specific therapy or 10-years cancer-free survival probability.

6. Conclusions, challenges and future trends

It is likely that in the very near future DL-based algorithms will tend to be more and more general, substantially independent from anatomical site as well as scale-agnostic (two recent scale-agnostic DL-based algorithms have already been proposed [121,122]) and even perform cross-species. The authors of [123] demonstrate that cross-species histology transfer learning leads to a richer feature representation for performing DL on human tissue. This paves the way for pan-cancer, pan-tissue, cross-species general-purpose high-performing models for lesion classification. A futuristic, but not unlikely algorithm will be able to provide a patient-specific therapeutic strategy and evaluate patient-specific survival probability by analyzing WSI, possibly in conjunction with additional multi-domain clinical information available at diagnosis time.

Interestingly, the fast paced research in AI-assisted digital pathology, and the exponential output in terms of methods and research papers, has not been accompanied by the introduction of DL in clinical practice. This may partially be due to the non yet ubiquitous adoption of digital WSI (which is itself still under development) and dedicated processing hardware. Perhaps more importantly, internationally recognized and algorithmically approved workflows are still missing (for reference, see the recently proposed regulatory framework by Food and Drug Administration (FDA) for the approval of AI methods in clinical routine [124]). In turn, this crucial step would require the availability of very large, annotated multicentric databases.

As architectures becomes more and more complex, another important limitation may be the loss of interpretability of the model’s inner workings. The design of highly complex, deep NNs has created a new “black box” problem [125]. While this may not be crucial in a number of other applications, in medical disciplines the decision making process is closely tied to questions of accountability, as well as of regulatory and ethical nature. In this sense, the strive for model performance should not eclipse the attention for model transparency. As a result, some authors [125] advocate the exclusive use of high-level abstraction of those attributes which have been associated with prior knowledge by human experts, even if this comes with a detriment in terms of classification accuracy. An opposite strategy (which would not *per se* compromise in terms of classification accuracy) would be to focus on AI systems which are explicable by design. Such architectures are currently under development in the computer science community [126,127]. In computational pathology, inherently self-interpretable models may be based on techniques like tile-level captioning [128], saliency maps [129] or visual attention maps [130,131]. An additional caveat is represented by operator “deskilling”: some authors [132] argue that reliance on automated DL-based decisions could result in the gradual loss of human diagnostic skills, hence exposing the system to potential disruptions in case of technological failure [133]. While it is important to note that the history of medical innovation, and any technological innovation in general, has inevitably brought some degree of human deskilling, some authors also note that technological progress in medicine has a generally negative influence on patient-practitioner communication [134]. Further it should be kept in mind that the adoption of DL-based decisions making systems typically requires high-performance computing infrastructure to develop/implement AI algorithms. To-date, this may represent an economic barrier in the less economically developed countries. In this respect, cloud solutions, where the bulk of computation time is rented, may aid in faster and cheaper deployment.

In summary, there is little doubt that DL will grain significant ground in biomedical sciences and medical practice, and pathology is no exception. A recent paper [135], demonstrated statistically that the speed of adoption of DL technologies is driven by high rates of mortality of some types of cancer, suggesting an imminent shift of technological paradigm for diagnostic assessment. Given to the relatively high mortality rates in breast cancer, the adoption of DL is likely to advance quicker as compared to other contexts. Further socioeconomic factors, such as e.g. the lack of an adequate number of pathologists in developing countries [136], as well as the ever-increasing workload in the context of histological diagnoses (which may also lead to an emergent workforce crisis [137] even in developed countries), will likely drive a digitization of healthcare services, which in turn will further ease the introduction of computer-assisted decision making in medicine.

Conflict of interest

The authors declare no conflict of interest.

References

[1] H. Nguyen, L.-M. Kieu, T. Wen, C. Cai, Deep learning methods in transportation domain: review, IET Intell. Transp. Syst. 12 (November (9)) (2018) 998–1004.

[2] T.J. Sejnowski, The Deep Learning Revolution, MIT Press, 2018.

[3] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, M.S. Lew, Deep learning for visual understanding: a review, Neurocomputing 187 (April) (2016) 27–48.

[4] A.B. Nassif, I. Shahin, I. Attili, M. Azzeh, K. Shaalan, Speech recognition using deep neural networks: a systematic review, IEEE Access 7 (2019) 19143–19165.

[5] T. Young, D. Hazarika, S. Poria, E. Cambria, Recent trends in deep learning based natural language processing [review article], IEEE Comput. Intell. Mag. 13 (August (3)) (2018) 55–75.

[6] Z.-H. Ling, S.-Y. Kang, H. Zen, A. Senior, M. Schuster, X.-J. Qian, H.M. Meng, L. Deng, Deep learning for acoustic modeling in parametric speech generation: a systematic review of existing techniques and future trends, IEEE Signal Process. Mag. 32 (May (3)) (2015) 35–52.

[7] D.C. Elton, Z. Boukouvalas, M.D. Fuge, P.W. Chung, Deep learning for molecular design-a review of the state of the art, Mol. Syst. Des. Eng. 4 (4) (2019) 828–849.

[8] D. Grapov, J. Fahrmann, K. Wanichthanarak, S. Khoomrung, Rise of deep learning for genomic, proteomic, and metabolomic data integration in precision medicine, OMICS: J. Integr. Biol. 22 (October (10)) (2018) 630–636.

[9] S.M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, M.K. Khan, Medical image analysis using convolutional neural networks: a review, J. Med. Syst. 42 (October (11)) (2018).

[10] M. Bakator, D. Radosav, Deep learning and medical diagnosis: a review of literature, Multimodal Technol. Interact. 2 (August (3)) (2018) 47.

[11] L. Boldrini, J.-E. Bibault, C. Masciocchi, Y. Shen, M.-I. Bittner, Deep learning: a review for the radiation oncologist, Front. Oncol. 9 (2019) 977.

[12] C. Liew, The future of radiology augmented with artificial intelligence: a strategy for success, Eur. J. Radiol. 102 (May) (2018) 152–156.

[13] P. Lanillos, D. Oliva, A. Philippsen, Y. Yamashita, Y. Nagai, G. Cheng, A review on neural network models of schizophrenia and autism spectrum disorder, Neural Netw. 122 (2020) 338–363.

[14] S.E. Dilsizian, E.L. Siegel, Artificial intelligence in medicine and cardiac imaging: Harnessing big data and advanced computing to provide personalized medical diagnosis and treatment, Curr. Cardiol. Rep. 16 (December (1)) (2013).

[15] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, Y. Wang, Artificial intelligence in healthcare: past, present and future, Stroke Vasc. Neurol. 2 (June (4)) (2017) 230–243.

[16] A.B. Levine, C. Schlosser, J. Grewal, R. Coope, S.J.M. Jones, S. Yip, Rise of the machines: advances in deep learning for cancer diagnosis, Trends Cancer 5 (March (3)) (2019) 157–169.

[17] D. Ardila, A.P. Kiraly, S. Bharadwaj, B. Choi, J.J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, D.P. Naidich, S. Shetty, End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography, Nat. Med. 25 (May (6)) (2019) 954–961.

[18] P. Tschandl, N. Codella, B.N. Akay, G. Argenziano, R.P. Braun, H. Cabo, D. Gutman, A. Halpern, B. Helba, R. Hofmann-Wellenhof, A. Lallas, J. Lapins, C. Longo, J. Malvehy, M.A. Marchetti, A. Marghoob, S. Menzies, A. Oakley, J. Paoli, S. Puig, C. Rinner, C. Rosendahl, A. Scope, C. Sinz, H.P. Soyer, L. Thomas, I. Zalaudek, H. Kittler, Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study, Lancet Oncol. 20 (July (7)) (2019) 938–947.

[19] T.J. Brinker, A. Hekler, A.H. Enk, J. Klode, A. Hauschild, C. Berking, B. Schilling, S. Haferkamp, D. Schadendorf, T. Holland-Letz, J.S. Utikal, C. von Kalle, W. Ludwig-Peitsch, J. Sirokay, L. Heinzerling, M. Albrecht, K. Baratella, L. Bischof, E. Chorti, A. Dith, C. Drusio, N. Giese, E. Gratsias, K. Griewank, S. Hallasch, Z. Hanhart, S. Herz, K. Hohaus, P. Jansen, F. Jockenhöfer, T. Kanaki, S. Knispel, K. Leonhard, A. Martaki, L. Matei, J. Matull, A. Olischewski, M. Petri, J.-M. Placke, S. Raub, K. Salva, S. Schlott, E. Sody, N. Steingrube, I. Stoffels, S. Ugurel, A. Zaremba, C. Gebhardt, N. Booken, M. Christolouka, K. Buder-Bakhaya, T. Bokor-Billmann, A. Enk, P. Gholam, H. Hänßle, M. Salzmann, S. Schäfer, K. Schäkel, T. Schank, A.-S. Bohne, S. Deffaa, K. Drerup, F. Egberts, A.-S. Erkens, B. Ewald, S. Falkvoll, S. Gerdes, V. Harde, A. Hauschild, M. Jost, K. Kosova, L. Messinger, M. Metzner, K. Morrison, R. Motamedi, A. Pinczker, A. Rosenthal, N. Scheller, T. Schwarz, D. Stölzl, F. Thielking, E. Tomaschewski, U. Wehkamp, M. Weichenthal, O. Wiedow, C.M. Bär, S. Bender-Säbelkampf, M. Horbrügger, A. Karoglan, L. Kraas, J. Faulhaber, C. Geraud, Z. Guo, P. Koch, M. Linke, N. Maurier, V. Müller, B. Thomas, J.S. Utikal, A.S.M. Alamri, A. Baczako, C. Berking, M. Betke, C. Haas, D. Hartmann, M.V. Heppt, K. Kilian, S. Krammer, N.L. Lapczynski, S. Mastnik, S. Nasifoglu, C. Ruini, E. Sattler, M. Schlaak, H. Wolff, B. Achatz, A. Bergbreiter, K. Drexler, M. Ettinger, S. Haferkamp, A. Halupczok, M. Hegemann, V. Dinauer, M. Maagk, M. Mickler, B. Philipp, A. Wilm, C. Wittmann, A. Gesierich, V. Glutsch, K. Kahlert, A. Kerstan, B. Schilling, P. Schrüfer, Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task, Eur. J. Cancer 113 (may) (2019) 47–54.

[20] L. Pehrson, M. Nielsen, C.A. Lauridsen, Automatic pulmonary nodule detection applying deep learning or machine learning algorithms to the LIDC-IDRI database: a systematic review, Diagnostics 9 (March (1)) (2019) 29.

[21] W. Kuo, C. Häne, P. Mukherjee, J. Malik, E.L. Yuh, Expert-level detection of acute intracranial hemorrhage on head computed tomography using deep learning, Proc. Natl. Acad. Sci. USA 116 (October (45)) (2019) 22737–22745.

[22] A. Yala, C. Lehman, T. Schuster, T. Portnoi, R. Barzilay, A deep learning mammography-based model for improved breast cancer risk prediction, Radiology 292 (July (1)) (2019) 60–66.

[23] S. Mukhopadhyay, M.D. Feldman, E. Abels, R. Ashfaq, S. Beltaifa, N. G. Cacciabeve, H.P. Cathro, L. Cheng, K. Cooper, G.E. Dickey, R.M. Gill, R. P. Heaton, R. Kerstens, G.M. Lindberg, R.K. Malhotra, J.W. Mandell, E. D. Manlucu, A.M. Mills, S.E. Mills, C.A. Moskaluk, M. Nelis, D.T. Patil, C. G. Przybycin, J.P. Reynolds, B.P. Rubin, M.H. Saboorian, M. Salicru, M.A. Samols, C.D. Sturgis, K.O. Turner, M.R. Wick, J.Y. Yoon, P. Zhao, C.R. Taylor, Whole slide imaging versus microscopy for primary diagnosis in surgical pathology, Am. J. Surg. Pathol. (September) (2017) 1.

[24] F. Aeffner, M.D. Zarella, N. Buchbinder, M. Bui, M.R. Goodman, D.J. Hartman, G. M. Lujan, M.A. Molani, A.V. Parwani, K. Lillard, O.C. Turner, V.N.P. Vemuri, A. G. Yuil-Valdes, D. Bowman, Introduction to digital image analysis in whole-slide imaging: a white paper from the digital pathology association, J. Pathol. Inform. 10 (1) (2019) 9.

[25] M. Recht, R. Nick, Bryan, Artificial intelligence: threat or boon to radiologists? J. Am. Coll. Radiol. 14 (November (11)) (2017) 1476–1480.

[26] S. Jha, E.J. Topol, Adapting to artificial intelligence, JAMA 316 (December (22)) (2016) 2353.

[27] F. Pesapane, M. Codari, F. Sardanelli, Artificial intelligence in medical imaging: threat or opportunity?. radiologists again at the forefront of innovation in medicine, Eur. Radiol. Exp. 2 (October (1)) (2018).

[28] A.L. Beam, I.S. Kohane, Translating artificial intelligence into clinical care, JAMA 316 (December (22)) (2016) 2368.

[29] A.L. Fogel, J.C. Kvedar, Artificial intelligence powers digital medicine, NPJ Digit. Med. 1 (March (1)) (2018).

[30] K.-H. Yu, A.L. Beam, I.S. Kohane, Artificial intelligence in healthcare, Nat. Biomed. Eng. 2 (October (10)) (2018) 719–731.

[31] D. Chen, S. Liu, P. Kingsbury, S. Sohn, C.B. Storlie, E.B. Habermann, J. M. Naessens, D.W. Larson, H. Liu, Deep learning and alternative learning strategies for retrospective real-world clinical data, NPJ Digit. Med. 2 (May (1)) (2019) 1–5.

[32] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, 2002.

[33] W.H. Guss, R. Salakhutdinov, On Characterizing the Capacity of Neural Networks Using Algebraic Topology, 2018.

[34] D.T. Jones, Setting the standards for machine learning in biology, Nat. Rev. Mol. Cell Biol. 20 (September (11)) (2019) 659–660.

[35] M. Wainberg, D. Merico, A. Delong, B.J. Frey, Deep learning in biomedicine, Nat. Biotechnol. 36 (October (9)) (2018) 829–838.

[36] Y. Jiang, C. Yang, J. Na, G. Li, Y. Li, J. Zhong, A brief review of neural networks based learning and control and their applications for robots, Complexity 2017 (2017) 1–14.

[37] Y. LeCun, K. Kavukcuoglu, C. Farabet, Convolutional networks and applications in vision, in: Proceedings of 2010 IEEE International Symposium on Circuits and Systems, IEEE, 2010, pp. 253–256.

[38] J. Schmidhuber, Deep learning in neural networks: an overview, Neural Netw. 61 (January) (2015) 85–117.

[39] R. Yamashita, M. Nishio, R.K.G. Do, K. Togashi, Convolutional neural networks: an overview and application in radiology, Insights Imaging 9 (4) (2018) 611–629.

[40] A. Krizhevsky, G. Hinton, et al., Learning Multiple Layers of Features From Tiny Images, Technical Report, Citeseer, 2009.

[41] K. Gnana Sheela, S.N. Deepa, Review on methods to fix number of hidden neurons in neural networks, Math. Probl. Eng. 2013 (2013) 1–11.

[42] J. Fan, C. Ma, Y. Zhong, A Selective Overview of Deep Learning, 2019. arXiv preprint arXiv:1904.05526.

[43] S. Sengupta, S. Basak, P. Saikia, S. Paul, V. Tsalavoutis, F. Atiah, V. Ravi, A. Peters, A Review of Deep Learning With Special Emphasis on Architectures, Applications and Recent Trends, 2019. arXiv preprint arXiv:1905.13294.

[44] H. Al-Askar, N. Radi, Á. MacDermott, Recurrent neural networks in medical data analysis and classifications. Applied Computing in Medicine and Health, Elsevier, 2016, pp. 147–165.

[45] D. Shen, G. Wu, H.-I. Suk, Deep learning in medical image analysis, Annu. Rev. Biomed. Eng. 19 (2017) 221–248.

[46] J.M. Tomczak, Application of classification restricted Boltzmann machine to medical domains, World Appl. Sci. J. 31 (2014) 69–75.

[47] K. Raza, N.K. Singh. A tour of unsupervised deep learning for medical image analysis.

[48] M. Chen, X. Shi, Y. Zhang, D. Wu, M. Guizani, Deep Features Learning for Medical Image Analysis With Convolutional Autoencoder Neural Network, 2017, p. 1.

[49] A.M. Abdel-Zaher, A.M. Eldeib, Breast cancer classification using deep belief networks, Expert Syst. Appl. 46 (March) (2016) 139–144.

[50] G. Altan, Diagnosis of Coronary Artery Disease Using Deep Belief Networks, vol. 2 (1, 2017, pp. 29–36.

[51] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems (2012) 1097–1105.

[52] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, 2014.

[53] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, June, 2015.

[54] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, 2015.

[55] W.M. Kouw, M. Loog, An Introduction to Domain Adaptation and Transfer Learning, 2018.

[56] E.A. Krupinski, A.A. Tillack, L. Richter, J.T. Henderson, A.K. Bhattacharyya, K. M. Scott, A.R. Graham, M.R. Descour, J.R. Davis, R.S. Weinstein, Eye-movement study and human performance using telepathology virtual slides. Implications for medical education and differences with experience, Hum. Pathol. 37 (December (12)) (2006) 1543–1556.

[57] K.H. Allison, L.M. Reisch, P.A. Carney, D.L. Weaver, S.J. Schnitt, F.P.O. Malley, B. M. Geller, J.G. Elmore, Understanding diagnostic variability in breast pathology: lessons learned from an expert consensus review panel, Histopathology 65 (April (2)) (2014) 240–251.

[58] B.E. Bejnordi, G. Litjens, N. Timofeeva, I. Otte-Holler, A. Homeyer, N. Karssemeijer, J.A.W.M. van der Laak, Stain specific standardization of whole-slide histopathological images, IEEE Trans. Med. Imaging 35 (February (2)) (2016) 404–415.

[59] A. Sethi, L. Sha, A. Vahadane, R.J. Deaton, N. Kumar, V. Macias, P.H. Gann, Empirical comparison of color normalization methods for epithelial-stromal classification in h and e images, J. Pathol. Inform. 7 (1) (2016) 17.

[60] F. Ciompi, O. Geessink, B.E. Bejnordi, G.S. de Souza, A. Baidoshvili, G. Litjens, B. van Ginneken, I. Nagtegaal, J. van der Laak, The importance of stain normalization in colorectal tissue classification with convolutional networks, in: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), IEEE, 2017. April.

[61] A.C. Ruifrok, D.A. Johnston, Quantification of histochemical staining by color deconvolution, Anal. Quant. Cytol. Histol. Int. Acad. Cytol. Am. Soc. Cytol. 23 (4) (2001) 291–299.

[62] E. Reinhard, M. Adhikhmin, B. Gooch, P. Shirley, Color transfer between images, IEEE Comput. Graphics Appl. 21 (4) (2001) 34–41.

[63] Y. Liu, K. Gadepalli, M. Norouzi, G.E. Dahl, T. Kohlberger, A. Boyko, S. Venugopalan, A. Timofeev, P.Q. Nelson, G.S. Corrado, J.D. Hipp, L. Peng, M. C. Stumpe, Detecting Cancer Metastases on Gigapixel Pathology Images, 2017.

[64] A.M. Khan, N. Rajpoot, D. Treanor, D. Magee, A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution, IEEE Trans. Biomed. Eng. 61 (June (6)) (2014) 1729–1738.

[65] A. Anghel, M. Stanisavljevic, S. Andani, N. Papandreou, J.H. Rüschoff, P. Wild, M. Gabrani, H. Pozidis, A high-performance system for robust stain normalization of whole-slide images in histopathology, Front. Med. (September) (2019) 6.

[66] A. Bentaieb, G. Hamarneh, Adversarial stain transfer for histopathology image analysis, IEEE Trans. Med. Imaging 37 (March (3)) (2018) 792–802.

[67] A. Shrivastava, W. Adorno, L. Ehsan, S.A. Ali, S.R. Moore, B.C. Amadi, P. Kelly, D. E. Brown, S. Syed, Self-Attentive Adversarial Stain Normalization, 2019.

[68] M. Balkenhol, N. Karssemeijer, G.J.S. Litjens, J. van der Laak, F. Ciompi, D. Tellez, H&e stain augmentation improves generalization of convolutional networks for histopathological mitosis detection, in: M.N. Gurcan, J. E. Tomaszewski (Eds.), Medical Imaging 2018: Digital Pathology, SPIE, 2018. March.

[69] D. Tellez, M. Balkenhol, I. Otte-Holler, R. van de Loo, R. Vogels, P. Bult, C. Wauters, W. Vreuls, S. Mol, N. Karssemeijer, G. Litjens, J. van der Laak, F. Ciompi, Whole-slide mitosis detection in h&e breast histology using PHH3 as a reference to train distilled stain-invariant convolutional networks, IEEE Trans. Med. Imaging 37 (September (9)) (2018) 2126–2136.

[70] L. Roux, D. Racoceanu, N. Loménie, M. Kulikova, H. Irshad, J. Klossa, F. Capron, C. Genestie, G. Naour, M.N. Gurcan, Mitosis detection in breast cancer histological images an ICPR 2012 contest, J. Pathol. Inform. 4 (1) (2013) 8.

[71] M. Veta, P.J. van Diest, S.M. Willems, H. Wang, A. Madabhushi, A. Cruz-Roa, F. Gonzalez, A.B.L. Larsen, J.S. Vestergaard, A.B. Dahl, D.C. Cireşan, J. Schmidhuber, A. Giusti, L.M. Gambardella, F.B. Tek, T. Walter, C.-W. Wang, S. Kondo, B.J. Matuszewski, F. Precioso, V. Snell, J. Kittler, T.E. de Campos, A. M. Khan, N.M. Rajpoot, E. Arkoumani, M.M. Lacle, M.A. Viergever, J.P.W. Pluim, Assessment of algorithms for mitosis detection in breast cancer histopathology images, Med. Image Anal. 20 (February (1)) (2015) 237–248.

[72] K. Sirinukunwattana, J.P.W. Pluim, H. Chen, X. Qi, P.-A. Heng, Y.B. Guo, L. Y. Wang, B.J. Matuszewski, E. Bruni, U. Sanchez, A. Böhm, O. Ronneberger, B. B. Cheikh, D. Racoceanu, P. Kainz, M. Pfeiffer, M. Urschler, D.R.J. Snead, N. M. Rajpoot, Gland segmentation in colon histology images: the glas challenge contest, Med. Image Anal. 35 (jan) (2017) 489–502.

[73] https://mitos-atypia-14.grand-challenge.org/.

[74] B.E. Bejnordi, M. Veta, P.J. van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, J.A.W.M. van der Laak, M. Hermsen, Q.F. Manson, M. Balkenhol, O. Geessink, N. Stathonikos, M.C.R.F. van Dijk, P. Bult, F. Beca, A.H. Beck, D. Wang, A. Khosla, R. Gargeya, H. Irshad, A. Zhong, Q. Dou, Q. Li, H. Chen, H.-J. Lin, P.-A. Heng, C. Haß, E. Bruni, Q. Wong, U. Halici, M.Ü. Öner, R. Cetin-Atalay, M. Berseth, V. Khvatkov, A. Vylegzhanin, O. Kraus, M. Shaban, N. Rajpoot, R. Awan, K. Sirinukunwattana, T. Qaiser, Y.-W. Tsang, D. Tellez, J. Annuscheit, P. Hufnagl, M. Valkonen, K. Kartasalo, L. Latonen, P. Ruusuvuori, K. Liimatainen, S. Albarqouni, B. Mungal, A. George, S. Demirci, N. Navab, S. Watanabe, S. Seno, Y. Takenaka, H. Matsuda, H.A. Phoulady, V. Kovalev, A. Kalinovsky, V. Liauchuk, G. Bueno, M.M. Fernandez-Carrobles, I. Serrano, O. Deniz, D. Racoceanu, R. Venâncio, Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer, JAMA 318 (December (22)) (2017) 2199.

[75] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going Deeper With Convolutions, 2014.

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

[77] B.S. Veeling, J. Linmans, J. Winkens, T. Cohen, M. Welling, Rotation Equivariant CNNS for Digital Pathology, 2018.

[78] P. Bandi, O. Geessink, Q. Manson, M. Van Dijk, M. Balkenhol, M. Hermsen, B. E. Bejnordi, B. Lee, K. Paeng, A. Zhong, Q. Li, F.G. Zanjani, S. Zinger, K. Fukuta, D. Komura, V. Ovtcharov, S. Cheng, S. Zeng, J. Thagaard, A.B. Dahl, H. Lin, H. Chen, L. Jacobsson, M. Hedlund, M. cetin, E. Halici, H. Jackson, R. Chen, F. Both, J. Franke, H. Kusters-Vandevelde, W. Vreuls, P. Bult, B. van Ginneken, J. van der Laak, G. Litjens, From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge, IEEE Trans. Med. Imaging 38 (February (2)) (2019) 550–560.

[79] T.L. Tio, The TNM staging system, Gastrointest. Endosc. 43 (January (2)) (1996) S19–S24.

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

[81] C.W. ELSTON, ELLI. pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up, Histopathology 19 (November (5)) (1991) 403–410.

[82] E.A. Rakha, J.S. Reis-Filho, F. Baehner, D.J. Dabbs, T. Decker, V. Eusebi, S.B. Fox, S. Ichihara, J. Jacquemier, S.R. Lakhani, J. Palacios, A.L. Richardson, S.J. Schnitt, F.C. Schmitt, P.-H. Tan, G.M. Tse, S. Badve, I.O. Ellis, Breast cancer prognostic classification in the molecular era: the role of histological grade, Breast Cancer Res. 12 (July (4)) (2010).

[83] D.C. Cireşan, A. Giusti, L.M. Gambardella, J. Schmidhuber, Mitosis detection in breast cancer histology images with deep neural networks, in: In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013, Springer Berlin Heidelberg, 2013, pp. 411–418.

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

[85] D. Cireşan, U. Meier, J. Schmidhuber, Multi-column Deep Neural Networks for Image Classification, 2012.

[86] D. Scherer, A. Müller, S. Behnke, Evaluation of pooling operations in convolutional architectures for object recognition, in: Artificial Neural Networks – ICANN 2010, Springer Berlin Heidelberg, 2010, pp. 92–101.

[87] H. Wang, A. Cruz-Roa, A. Basavanhally, H. Gilmore, N. Shih, M. Feldman, J. Tomaszewski, F. Gonzalez, A. Madabhushi, Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features, J. Med. Imaging 1 (October (3)) (2014) 034003.

[88] C.D. Malon, E. Cosatto, Classification of mitotic figures with convolutional neural networks and seeded blob features, J. Pathol. Inform. 4 (1) (2013) 9.

[89] F. Pourakpour, H. Ghassemian, Automated mitosis detection based on combination of effective textural and morphological features from breast cancer histology slide images, in: 2015 22nd Iranian Conference on Biomedical Engineering (ICBME), IEEE, 2015. November.

[90] H. Chen, Q. Dou, X. Wang, J. Qin, P.A. Heng, Mitosis detection in breast cancer histology images via deep cascaded networks, Thirtieth AAAI Conference on Artificial Intelligence (2016).

[91] H. Chen, X. Wang, P.A. Heng, Automated mitosis detection with deep regression networks, in: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), IEEE, 2016. April.

[92] C. Li, X. Wang, W. Liu, L.J. Latecki, DeepMitosis: mitosis detection via deep detection, verification and segmentation networks, Med. Image Anal. 45 (April) (2018) 121–133.

[93] M.C.A. Balkenhol, D. Tellez, W. Vreuls, P.C. Clahsen, H. Pinckaers, F. Ciompi, P. Bult, J.A.W.M. van der Laak, Deep learning assisted mitotic counting for breast cancer, Lab. Investig. 99 (June (11)) (2019) 1596–1606.

[94] M. Saha, C. Chakraborty, D. Racoceanu, Efficient deep learning model for mitosis detection using breast histopathology images, Comput. Med. Imaging Graphics 64 (March) (2018) 29–40.

[95] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, N. Navab, AggNet: deep learning from crowds for mitosis detection in breast cancer histology images, IEEE Trans. Med. Imaging 35 (May (5)) (2016) 1313–1321.

[96] D. Romo-Bucheli, A. Janowczyk, H. Gilmore, E. Romero, A. Madabhushi, Automated tubule nuclei quantification and correlation with oncotype DX risk categories in ER+ breast cancer whole slide images, Sci. Rep. 6 (September (1)) (2016) 32706.

[97] A. Janowczyk, A. Madabhushi, Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases, J. Pathol. Inform. 7 (1) (2016) 29.

[98] M.Z. Alom, T. Aspiras, T.M. Taha, V.K. Asari, T.J. Bowen, D. Billiter, S. Arkell, Advanced Deep Convolutional Neural Network Approaches for Digital Pathology Image Analysis: A Comprehensive Evaluation With Different Use Cases, 2019.

[99] B. Dunne, J.J. Going, Scoring nuclear pleomorphism in breast cancer, Histopathology 39 (September (3)) (2001) 259–265.

[100] N. Wahab, A. Khan, Y.S. Lee, Two-phase deep convolutional neural network for reducing class skewness in histopathological images based breast cancer detection, Comput. Biol. Med. 85 (June) (2017) 86–97.

[101] J. Xu, L. Xiang, Q. Liu, H. Gilmore, J. Wu, J. Tang, A. Madabhushi, Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images, IEEE Trans. Med. Imaging 35 (January (1)) (2016) 119–130.

[102] M. Veta, P.J. van Diest, J.P.W. Pluim, Cutting out the middleman: Measuring nuclear area in histopathology slides without segmentation, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, Springer International Publishing, 2016, pp. 632–639.

[103] F. Xing, Y. Xie, L. Yang, An automatic learning-based framework for robust nucleus segmentation, IEEE Trans. Med. Imaging 35 (February (2)) (2016) 550–566.

[104] N. Dendukuri, K. Khetani, M. McIsaac, J. Brophy, Testing for HER2-positive breast cancer: a systematic review and cost-effectiveness analysis, Can. Med. Assoc. J. 176 (May (10)) (2007) 1429–1434.

[105] T. Bonacho, F. Rodrigues, J. Liberal, Immunohistochemistry for diagnosis and prognosis of breast cancer: a review, Biotech. Histochem. (2019) 1–21. September.

[106] A.C. Wolff, M. Elizabeth, H. Hammond, D.G. Hicks, M. Dowsett, L.M. McShane, K. H. Allison, D.C. Allred, J.M.S. Bartlett, M. Bilous, P. Fitzgibbons, W. Hanna, R. B. Jenkins, P.B. Mangu, S. Paik, E.A. Perez, M.F. Press, P.A. Spears, G.H. Vance, G. Viale, D.F. Hayes, Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American society of clinical oncology/college of american pathologists clinical practice guideline update, J. Clin. Oncol. 31 (November (31)) (2013) 3997–4013.

[107] C.L. Vogel, K. Bloom, H. Burris, J.R. Gralow, M. Mayer, M. Pegram, H.S. Rugo, S. M. Swain, D.A. Yardley, M. Chau, D. Lalla, M.G. Brammer, P.A. Kaufman, P1-07-02: discordance between central and local laboratory HER2 testing from a large HER2-negative population in VIRGO, a metastatic breast cancer registry, in: In Poster Session Abstracts, American Association for Cancer Research, 2011. December.

[108] P.C. Roche, V.J. Suman, R.B. Jenkins, N.E. Davidson, S. Martino, P.A. Kaufman, F. K. Addo, B. Murphy, J.N. Ingle, E.A. Perez, Concordance between local and central laboratory HER2 testing in the breast intergroup trial n9831, JNCI J. Natl. Cancer Inst. 94 (June (11)) (2002) 855–857.

[109] E.A. Perez, V.J. Suman, N.E. Davidson, S. Martino, P.A. Kaufman, W.L. Lingle, P. J. Flynn, J.N. Ingle, D. Visscher, R.B. Jenkins, HER2 testing by local, central, and reference laboratories in specimens from the north central cancer treatment group n9831 intergroup adjuvant trial, J. Clin. Oncol. 24 (July (19)) (2006) 3032–3038.

[110] J.M. Bueno de Mesquita, D.S.A. Nuyten, J. Wesseling, H. van Tinteren, S.C. Linn, M.J. van de Vijver, The impact of inter-observer variation in pathological assessment of node-negative breast cancer on clinical risk assessment and patient selection for adjuvant systemic treatment, Ann. Oncol. 21 (July (1)) (2009) 40–47.

[111] K. Bloom, D. Harrington, Enhanced accuracy and reliability of HER-2/neuImmunohistochemical scoring using digital microscopy, Am. J. Clin. Pathol. 121 (May (5)) (2004) 620–630.

[112] P.A. Kaufman, K.J. Bloom, H. Burris, J.R. Gralow, M. Mayer, M. Pegram, H. S. Rugo, S.M. Swain, D.A. Yardley, M. Chau, D. Lalla, B. Yoo, M.G. Brammer, C. L. Vogel, Assessing the discordance rate between local and central HER2 testing in women with locally determined HER2-negative breast cancer, Cancer 120 (17 jun) (2014) 2657–2664.

[113] M.E. Vandenberghe, M.L.J. Scott, P.W. Scorer, M. Söderberg, D. Balcerzak, C. Barker, Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer, Sci. Rep. 7 (April (1)) (2017).

[114] A. Cruz-Roa, H. Gilmore, A. Basavanhally, M. Feldman, S. Ganesan, N.N.C. Shih, J. Tomaszewski, F.A. González, A. Madabhushi, Accurate and reproducible invasive breast cancer detection in whole-slide images: a deep learning approach for quantifying tumor extent, Sci. Rep. 7 (April (1)) (2017).

[115] H.D. Couture, L.A. Williams, J. Geradts, S.J. Nyante, E.N. Butler, J.S. Marron, C. M. Perou, M.A. Troester, M. Niethammer, Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype, NPJ Breast Cancer 4 (September (1)) (2018).

[116] Y. Xu, J.-Y. Zhu, E.I.-C. Chang, M. Lai, Z. Tu, Weakly supervised histopathology cancer image segmentation and classification, Med. Image Anal. 18 (April (3)) (2014) 591–604.

[117] M.-A. Carbonneau, V. Cheplygina, E. Granger, G. Gagnon, Multiple instance learning: a survey of problem characteristics and applications, Pattern Recognit. 77 (May) (2018) 329–353.

[118] M. Vulchi, M.E.l. Adoui, N. Braman, P. Turk, M. Etesami, S. Drisis, D. Plecha, M. Benjelloun, A. Madabhushi, J. Abraham, Development and external validation of a deep learning model for predicting response to HER2-targeted neoadjuvant therapy from pretreatment breast MRI, J. Clin. Oncol. 37 (May (15_suppl)) (2019) 593.

[119] G. Campanella, M.G. Hanna, L. Geneslaw, A. Miraflor, V.W.K. Silva, K.J. Busam, E. Brogi, V.E. Reuter, D.S. Klimstra, T.J. Fuchs, Clinical-grade computational pathology using weakly supervised deep learning on whole slide images, Nat. Med. 25 (July (8)) (2019) 1301–1309.

[120] S. Kalra, H.R. Tizhoosh, S. Shah, C. Choi, S. Damaskinos, A. Safarpoor, S. Shafiei, M. Babaie, P. Diamandis, C.J.V. Campbell, L. Pantanowitz, Pan-Cancer Diagnostic Consensus Through Searching Archival Histopathology Images Using Artificial Intelligence, 2019.

[121] P. Bándi, M. Balkenhol, B. van Ginneken, J. van der Laak, G. Litjens, Resolution-agnostic tissue segmentation in whole-slide histopathology images with convolutional neural networks, PeerJ 7 (December) (2019) e8242.

[122] N. Hashimoto, D. Fukushima, R. Koga, Y. Takagi, K. Ko, K. Kohno, M. Nakaguro, S. Nakamura, H. Hontani, I. Takeuchi, Multi-Scale Domain-Adversarial Multiple-Instance CNN for Cancer Subtype Classification With Non-Annotated Histopathological Images, 2020.

[123] T. Sing, H. Hoefling, I. Hossain, J. Boisclair, A. Doelemeyer, T. Flandre, A. Piaia, V. Romanet, G. Santarossa, C. Saravanan, E. Sutter, O. Turner, K. Wuersch, P. Moulin, A Deep Learning-Based Model of Normal Histology, 2019. November.

[124] US FDA, Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (ai/ml)-Based Software as a Medical Device (samd), 2019.

[125] Y. Hayashi, The right direction needed to develop white-box deep learning in radiology, pathology, and ophthalmology: a short review, Front. Robot. AI (2019) 6. April.

[126] W. Samek, T. Wiegand, K.-R. Müller, Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models, 2017.

[127] C. Chen, O. Li, C. Tao, A.J. Barnett, J. Su, C. Rudin, This Looks Like That: Deep Learning for Interpretable Image Recognition, 2018.

[128] Z. Zhang, P. Chen, M. McGough, F. Xing, C. Wang, M. Bui, Y. Xie, M. Sapkota, L. Cui, J. Dhillon, N. Ahmad, F.K. Khalil, S.I. Dickinson, X. Shi, F. Liu, H. Su, J. Cai, L. Yang, Pathologist-level interpretable whole-slide cancer diagnosis with deep learning, Nat. Mach. Intell. 1 (May (5)) (2019) 236–245.

[129] D. Tellez, G. Litjens, J. van der Laak, F. Ciompi, Neural image compression for gigapixel histopathology image analysis, IEEE Trans. Pattern Anal. Mach. Intell. (2019) 1.

[130] A. BenTaieb, G. Hamarneh, Predicting cancer with a recurrent visual attention model for histopathology images, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Springer International Publishing, 2018, pp. 129–137.

[131] Y. Huang, A.C.S. Chung, Celnet: Evidence Localization for Pathology Images Using Weakly Supervised Learning, 2019.

[132] F. Cabitza, R. Rasoini, G.F. Gensini, Unintended consequences of machine learning in medicine, JAMA 318 (August (6)) (2017) 517.

[133] J. Lu, Will medical technology deskill doctors? Int. Educ. Stud. 9 (June (7)) (2016) 130.

[134] A.M. Johnston, Deskilling and return to practice on low-tempo contingency operations, J. R. Army Med. Corps 165 (May (5)) (2019) 310–311.

[135] M. Coccia, Deep learning technology for improving cancer care in society: new directions in cancer imaging driven by artificial intelligence, Technol. Soc. 60 (February) (2020) 101198.

[136] A.M. Nelson, M. Hale, M.I. Jean-Marie Diomande, Q. Eichbaum, Y. Iliyasu, R. M. Kalengayi, B. Rugwizangoga, S. Sayed, Training the next generation of african pathologists, Clin. Lab. Med. 38 (March (1)) (2018) 37–51.

[137] B.J. Williams, D. Bottoms, D. Treanor, Future-proofing pathology: the case for clinical adoption of digital pathology, J. Clin. Pathol. 70 (August (12)) (2017) 1010–1018.