The Breast 49 (2020) 267–273

Contents lists available at ScienceDirect

The Breast

journal homepage: www.elsevier.com/brst

Original article

Artificial intelligence in digital breast pathology: Techniques and applications

Asmaa Ibrahim [a, 1], Paul Gamble [b, 1], Ronnachai Jaroensri [b, 1], Mohammed M. Abdelsamea [c], Craig H. Mermel [b], Po-Hsuan Cameron Chen [b], Emad A. Rakha [a, *]

[a] *Department of Histopathology, Division of Cancer and Stem Cells, School of Medicine, The University of Nottingham and Nottingham University Hospitals NHS Trust, Nottingham City Hospital, Nottingham, NG5 1PB, UK*
[b] *Google Health, Google, Palo Alto, CA, USA*
[c] *School of Computing and Digital Technology, Birmingham City University, Birmingham, UK*

ARTICLE INFO

Article history:
Received 2 December 2019
Accepted 12 December 2019
Available online 19 December 2019

Keywords:
Breast cancer
AI
(Artificial intelligence)
ML
(Machine learning)
DL
(Deep learning)
WSI
(Whole slide image)
Digital
Pathology
Breast pathology
Applications

ABSTRACT

Breast cancer is the most common cancer and second leading cause of cancer-related death worldwide. The mainstay of breast cancer workup is histopathological diagnosis - which guides therapy and prognosis. However, emerging knowledge about the complex nature of cancer and the availability of tailored therapies have exposed opportunities for improvements in diagnostic precision. In parallel, advances in artificial intelligence (AI) along with the growing digitization of pathology slides for the primary diagnosis are a promising approach to meet the demand for more accurate detection, classification and prediction of behaviour of breast tumours. In this article, we cover the current and prospective uses of AI in digital pathology for breast cancer, review the basics of digital pathology and AI, and outline outstanding challenges in the field.

1. Introduction

Modern approaches to the treatment of breast cancer require careful diagnostic stratification of patients and prediction of survival for tailored therapy. This stratification is primarily based on manual interpretation of pathology slides - a time-consuming process with significant interobserver variability [1,2]. The trend towards digitization in pathology opens the door to computer-based image analysis solutions which have the potential to provide a more objective and quantitative slide reviews [3].

Over the last several decades, due to algorithmic advances, more accessible computing power, and the curation of large datasets, machine learning techniques have come to define the state-of-the-art in many computer vision tasks - including many healthcare applications [4,5]. Concurrently, digital pathology has emerged as a method for imaging and handling high magnification images of pathology slides - initially for research purposes but increasingly as a clinical tool. Recently, these two fields have intersected as computer scientists and pathologists have come together to apply the latest artificial intelligence (AI) techniques to the problem of analysing pathology slides for diagnostic, prognostic, predictive and other clinically relevant purposes in addition to other applications such as improving the efficiency of the diagnostic workflow.

Many problems in breast cancer pathology involve assessing morphological features of the tissue. However, this is often not straightforward and significant research has gone into improving reliability and reducing variability of the assessment [6–8]. The reliability and variability problem has the potential to be solved efficiently with computational methods. Once trained, the

[*] Corresponding author.
E-mail addresses: Emad.Rakha@nottingham.ac.uk, mzxai2@nottingham.ac.uk (E.A. Rakha).
[1] Denotes equal contribution.

https://doi.org/10.1016/j.breast.2019.12.007
0960-9776/

algorithms always give consistent results when the same input data is provided.

In this article, we explore AI applications in breast pathology. We start with an overview of digital pathology, a necessary prerequisite for the application of AI techniques. Then, we do a deep dive into the applications of machine learning to digital pathology for breast cancer, including both diagnostic and prognostic applications. Finally, we address outstanding challenges in the field and promising future directions.

1.1. Digital pathology

Digital pathology is the process of transforming histopathology slides into digital images using whole-slide scanners and subsequent analysis of these digitized images. In 1966, Prewitt [9–11] and Mendelsohn [12] first proposed a method to scan images from a microscopic field of a blood smear and use these scanned images to discern the presence of different cell types. In the mid 1990's, advances in microscopic imaging and software systems for storing, serving, and viewing large images (an average whole-slide image scanned at 40× magnification is greater than 1 GB) led to the development of whole-slide imaging (WSI) techniques. These techniques allow an entire slide (rather than individual fields-of-view) to be digitized and examined at a resolution comparable to light microscopy. Further developments in the following decades have brought digital pathology from a niche research topic to the edge of mainstream adoption in clinical practice.

An early large-scale comparison of diagnostic performance between digital pathology and conventional microscopy was performed by *Mukhopadhyay* et al. and included specimens from 1992 patients with various types of tumours read by 16 surgical pathologists.[13] The study showed that the diagnostic performance with digitized WSIs was nearly equal to that achieved with traditional microscopy-based methods (with a major discordance rate from the reference standard of 4.9% for WSI and 4.6% for microscopy). This study was used as the pivotal study for the FDA approval of Philips' digital pathology system [14]. Similarly, in pursuit of FDA approval for the Aperio AT2 DX WSI system, Leica Biosystems conducted a clinical trial across five study sites involving over 16,000 slides and found 97.9% intra-system concordance (i.e. agreement between reads from glass and digital images at any one site) [15].

Williams et al. performed a clinical validation of breast cancer diagnosis from digital slides and found total concordance between glass and digital slide reads in 98.8% of 694 cases by speciality breast pathologists who had received a short digital pathology training course [9]. The same group performed a systematic analysis of 8069 published comparisons between glass and digital reads - finding discordant diagnoses in 335 cases (4%) [10].

While equivalency of diagnostic accuracies from glass slides and digital images is well validated, studies report conflicting evidence on the potential impact that "going digital" can have on practicing pathologists. In 2016, the Granada University Hospital system adopted an entirely digital workflow for primary histopathology diagnosis. *Retamero* et al. showed a 21% average increase in annual per-pathologist case sign-outs [11]. In contrast, *Hanna* et al. report a 19% decrease in efficiency - defined as signout turnaround time - per case for approximately 200 cases across six anatomic pathology specializations at Memorial Sloan Kettering Cancer Centre. The authors note that their study does not evaluate any learning effect and that the participating pathologists had varied levels of experience with digital pathology. A similar study conducted by *Mills* et al. assessed diagnostic efficiency on 510 surgical pathology cases and found a median increase in diagnostic time of 4 s per case for digital reads compared to glass [16]. They observed significant inter-reader variability in the increased digital read times and noted a dramatic learning effect over the duration of the study ($\geq$ 6 weeks) which reduced the glass-digital difference to near zero by the end of the experimental series.

A consistent theme in evaluations of digital pathology, both for diagnostic accuracy and efficiency, is that successful implementation depends on proper training design and integration with existing workflows. Beyond direct effects on time-per-slide workflow efficiency, other potential benefits of digital pathology adoption include reduced risk of patient and slide misidentification, reduced risk of tissue loss or damage, better case tracking and workload allocation, streamlined retrieval of archival cases, and improved telepathology consultations in addition to facilitating cross coverage between hospitals for primary diagnosis, remote reporting centralisation of pathology laboratories [17]. However, one of the most important advantages of WSI for primary diagnosis is the ability apply various AI-based algorithms in the routine diagnostic workflow.

2. Machine learning basics

AI is a broad research field which aims at designing computer systems that simulate human intelligence. Machine learning (ML) is a subfield of AI that develops algorithms that allows computer to adapt to a new problem without being reprogrammed. That is, a machine learning system "learns" to solve a problem directly from data. This is done by applying statistical methods to recognize patterns from a set of provided data without human instruction. Most ML algorithms can be viewed as mathematical models that map a set of observed variables, known as 'features' or 'predictors', of a data point or sample, into a set of outcome variables, known as 'labels' or 'targets' [18,19]. The observed variable and the output labels can be simple scalars such as age, weight, and gender of a patient, to a more primitive observation such as images of a histopathology slides. However, as the relationship between the observed feature variables and the desired outcome labels becomes more complicated (such as mapping raw pixel values of images to its semantic label), the sophistication of the ML algorithms will have to grow to match that complexity of the relationship.

As the computational power grew, more complex algorithms become available. Deep learning techniques utilize millions of neuron-like units in order to learn complex relationship between image pixel values and its semantic labels, without the need for manual feature engineering–the features are learned automatically from data [20,21]. Deep learning algorithms can be subdivided into categories based on network architectures. Convolutional neural networks (CNN) with hierarchical layers of pattern detectors has shown great success for image recognition problems. CNN-based approaches have been used for image-based detection and segmentation tasks to identify and quantify cells [22–25] and histological features [26–29]. Finally, recurrent neural networks (RNN) use self-connecting pattern detectors for sequence processing.

ML methods have been widely explored for various histopathology predictions. Broadly speaking, the types of predictions can be categorized as learning from humans or enabling identification of unknown signals. For the former, a model learns from human annotated datasets with an aim to assist pathologists for their diagnostics task in the clinical workflow. For the latter, a model can be developed using the same input data with outcome-based labels. These models have the potential to provide more accurate prognosis predictions and identify unknown signals for drug discovery (Table 1). There have also been applications that are a mixture of these types, such as models identifying well known morphological features as building blocks for predicting novel outcomes. Here we present the main applications of ML and AI-based algorithms in

Table 1
Current applications of artificial intelligence in breast pathology.

Applications	Comments
Diagnostic applications	
Tumor detection	
Primary tumor detection:	AI-based algorithms have been developed to detect malignant tumours in the breast and to differentiate it from benign and normal structures. Osareh et al. introduced ML techniques to differentiate between malignant and benign tumours using digitalized images of fine-needle aspiration biopsy samples [73]. Also, algorithms have been developed to provide quantitative measurements of nuclear shape and size, which could be applied across different tumour subtypes [54].
Metastatic deposits detection in lymph nodes:	One of the most important application is detection of metastatic tumour deposits in the lymph nodes. Babak et al., 2017 detected lymph node metastasis in breast cancer patients with a higher diagnostic achievement over 11 pathologists [33]
Breast cancer grading	Several algorithms have been developed to assess breast cancer grade. *Coutre* et al. [50] have used image analysis with DL to predict breast cancer grade. Other algorithms were developed to allow objective enumeration of mitotic figures [26], measurements of nuclear shape and size, and with the automatic detection and segmentation of cell nuclei in histopathology images [74].
Breast cancer subtype	Breast cancer comprises more than 20 histotypes. Coutre et al., used image analysis with DL to detect breast cancer histologic subtypes [50].
Assessment of tumour heterogeneity and tumour microenvironment	AI-based assays to measure tumour intra-tumour and inter-tumour heterogeneity [26,56], identify and quantify non-epithelial cells such as fibroblast, neutrophils, lymphocytes and macrophages [77] ^{and} computerized image-based detection and grading of tumour infiltrating lymphocytic (TILs) in HER2+ breast cancer [78] have been developed
Receptor status and intrinsic subtype assessment	AI algorithms have been developed to provide quantitative measurements of immunohistochemically stained Ki-67 [52], ER [50], PR and Her2neu images [75]. *Xu* et al. proposed a novel GAN-based approach to provide a virtual immunohistochemistry staining pattern from the H&E stained WSIs that potentially obviates the need for IHC-based tissue testing [52,76] *Coutre* et al., used image analysis with DL to predict breast cancer intrinsic subtype [50].
Prognostic Applications	
Prognostic significance of tumour morphological features	Morphological features as nuclear shape, texture and architecture can predict risk of recurrence and overall survival. *Whitney* et al., [54] showed that quantitative features of nuclear shape, texture and architecture independently enable prediction risk of recurrence in patients with ER-positive breast tumours
Prognostic significance of different peri-tumoral elements	AI-based assays to measure the arrangement and architecture of different tissue elements such as TILs within the tumour have been developed and demonstrated their value in predicting survival [79] and that the spatial distribution of TILs among tumour cells expression profiling is associated with late recurrence in ER-positive breast cancer [57].
Applications related to predictive values and response to treatment.	ML approached can be used to correlate the expression of certain markers such as cell cycle and proliferation markers [80] or the presence of certain morphological features in the tumour to the response of specific therapy.

breast pathology with emphasis on the diagnostic and prognostic applications.

2.1. Diagnostic applications

A critical first step in the diagnostic workup of suspected breast cancer is the detection of invasive tumor cells, the characterization of tumor type, and the quantification of tumor extent. *Cruz-Roa* et al. [30] built CNN to classify images patches from breast cancer WSIs as either containing invasive ductal carcinoma or not. They used manually annotated region labels for 400 slides from multiple sites to train their model and validated its performance on 200 slides with similar annotations from The Cancer Genome Atlas. They report a pixel-level F1 score of 75.86%. *Han* et al. [31] use the BreaKHis dataset [32] to train a classifier which can distinguish between eight classes of benign and malignant breast tumours with 93.2% accuracy. Their model is pretrained on imagenet and they employ extensive data augmentation to prevent overfitting.

One of the prominent ML diagnostic applications for breast cancer is the diagnosis of lymph node metastasis. *Bejnori* et al. reported on the performance of seven DL algorithms developed as part of a challenge competition; the algorithms were found to outperform a panel of 11 pathologists in a simulated time-constrained diagnostic setting [33]. On the basis of training data that included 270 images from two centres with ($n = 110$) and without ($n = 160$) nodal metastases, and evaluation on an independent set of 129 images (49 with and 80 without metastases), the AUC of the best algorithm was 0.99, whereas the best performance by a pathologist achieved an AUC of 0.88. A similar study had 6 pathologists reviewed 70 digitized slides with and without ML assistance [34,35]. The review time was measured as a primary endpoint. The average review time was significantly shorter with assistance than without assistance for both micrometastases (1.9 times faster) and images without any metastases (1.2 times faster).

In addition to tumour identification, AI and ML methods were used to characterise invasive breast tumours and histologic grading of breast cancer was a prime candidate due to the inherent subjectivity with low concordance rates despite its important prognostic value. Various cell-level and tissue-level features can be identified to describe the morphological structure of objects to discriminate between histological components, e.g., tumor/epithelial cells (for tubular formation) and mitotic cells (for mitotic count). However, most work in this area focuses on mitosis detection, which is the most prognostic but also the most laborious task. In 2013, *Veta* et al. proposed a public mitosis detection challenge [36] with a dataset containing 12 training, 11 testing slides, and roughly one thousand mitotic figures annotated. The winner of the challenge used a 10-layer deep convolutional neural network to achieve 0.61 overall F-1 score against the consensus of pathologists, whereas individual pathologist achieves >0.75 overall F-1 score. In 2016, *Veta* et al. published a follow up challenge that focuses on slide-level mitotic score [37]. The winners of this challenge achieved a Cohen's kappa score of 0.56 with the pathologists' slide-level score, and 0.65 F-1 score on cell-level mitosis detection.

The labour-intensive nature of mitotic counting can lead to high degree of discordance. PHH3 stains, which detects mitosis at high sensitivity, is an immunohistochemistry method for resolving this issue. *Tellez* et al. uses aligned scans of PHH3 and H&E stains to generate annotations for CNN [38]. Because of PHH3, they collected over 22,000 annotations from less than 100 slides. Their CNN did not achieve the state-of-the-art on TUPAC16, which could be caused by annotation variability. Nonetheless, their subsequent work shows that using CNN mitosis detection as an assistant can help improve the level of concordance among human pathologists [39].

Tubular formation and nuclear grade are the other two important components in the histopathological grading of breast cancer. However, fully automated methods for these two tasks are still to be developed. Current published work focuses on analysing tissue structure that could be used for these tasks. *Romo-Bucheli* et al. train a CNN to detect tubule nuclei, and compute statistics about the nuclei to predict Oncotype DX risk categories [39,40]. *Veta* et al. proposed a series of non-CNN algorithms to segment and detect nuclei [41,42]. These segmentation are then used to detect nuclei for further morphological analysis [39].

Biomarker status determination is another important element of breast cancer diagnosis. Oestrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor (HER2), and ki67 are all commonly assessed and used to determine which therapy options are offered to the patient. Current clinical practice relies on immunohistochemical (IHC) staining and manual comparison of staining intensity against a set of biomarker-specific scoring systems [43,44]. IHC techniques use targeting molecules, usually antibodies, paired with enzymes which are used to visualize specific antigen sites in tissue. Although this approach is well-validated and widely used, there is significant inter-observer variability in manual biomarker status assessment from IHC [45,46] suggesting an opportunity for machine-learning based systems to assist pathologists with biomarker status assessments.

Automated quantification of IHC staining intensity was an early application of statistical learning methods in digital breast cancer pathology. *Wang* et al. [47] used an automated cellular imaging system to determine the fraction of breast cancer cells which stained at varying levels of HER2 intensity and found a higher correlation between the algorithmic assay and HER2 status as determined by FISH (the gold standard) than between FISH and manual assessment of the IHC slides. In a similar approach, *Rexhepaj* et al. [48] designed a nuclear detection algorithm and used it to quantify IHC staining for ER and PR expression and found a correlation of 0.9 between manual and algorithmic quantification. *Skaland* et al. [49] used open source digital image analysis software to analyse HER2 IHC stains from 60 breast cancer cases with scores of 2 + or 3 + and found 100% concordance between the algorithm's prediction and the consensus clinical status assessment.

More recently, efforts have been made to predict breast cancer biomarkers directly from H&E slides - potentially bypassing the need for immunohistochemical staining altogether. *Couture* et al. [50] implemented both feature-based and deep learning models to predict ER status (as well as intermediate vs high grade, basal-like vs non-basal-like, and ductal vs lobular), trained on 571 H&E tissue micro-array images and tested on 288 images with a final test accuracy of 84% for ER status. In a related study, *Shamai* et al. [51] built a deep learning system to predict the statuses of 19 biomarkers including ER and PR. For ER status prediction, they chose positive and negative prediction thresholds such that they were only able to predict the statuses for 51% of their validation set - but within this subgroup of high-confidence cases they obtained 92% accuracy.

Other groups have generated immunohistochemically stained Ki-67 images using conditional GANs. *Senaras* et al. [52] presented 15 real and 15 synthetic images of breast cancer pathology to five experts, and their mean accuracy in distinguishing them was 47.3%, suggesting that the synthetic images were indistinguishable from real images and supporting the application of GANs to boost the training sets used to optimize classification of prostate cancer.

In breast pathology, 10 cellular features introduced by *Osareh* et al. [53] into ML that were identified by an expert breast pathologist to differentiate between malignant and benign tumours using images of fine-needle aspiration biopsy samples [53].

2.2. Prognostic applications

Many morphological features of the tumour tissues are known to have prognosis values. *Veta* et al. [39] showed that, in a tissue microarray (TMA) of male breast cancers, features such as nuclear shape or texture can be used to predict a patient's prognosis. *Whitney* et al. [54] showed that quantitative features of nuclear shape, texture and architecture independently enable prediction of recurrence risk in patients with ER-positive breast tumours (on the basis of the 21-gene expression-based companion diagnostic test Oncotype DX).

Structure and organization of tumour-infiltrating lymphocytes (TILs) are found to be prognostic of the clinical outcome. A study used a CNN to detect and quantify the structure of TILs in images from The Cancer Genome Atlas and found that this feature was prognostic of outcome for 13 different cancer subtypes [55]. Yuan proposed a method to model and analyse the spatial distribution of lymphocytes among tumour cells on triple-negative breast cancer WSIs [56]. Using this model, they identified three different categories of lymphocytes according to their spatial proximity to cancer cells. The ratio of intratumoral lymphocytes to cancer cells was found to be independently prognostic of survival and correlated with the levels of cytotoxic T lymphocyte protein 4 (CTLA-4) expression determined by TMA gene expression profiling. These investigators further expanded this method and found that the spatial distribution of immune cells was also associated with late recurrence in ER-positive breast cancer [57]. *Ali* et al. used classical ML methods to count features on breast cancer biopsies, and were able to predict neoadjuvant response [58]. Lymphocyte density in the surrounding tissue was found to be the biggest predictor.

Recurrence risk is another important aspect of prognosis. In a recent study [59] we developed a novel ML pipeline to predict risk of ipsilateral recurrence of DCIS using digitized WSI and clinicopathologic long-term outcome data from a retrospectively collected cohort of DCIS patients ($n = 344$) treated with lumpectomy [59]. The sections from primary tumours were stained with H&E, then digitized and analysed by the pipeline. In the first step, a classifier was applied to WSI to annotate the areas of stroma, normal/benign ducts, cancer ducts, dense lymphocyte region, and blood vessels. In the second step, a recurrence risk classifier was trained on eight select architectural and spatial organization tissue features from the annotated areas to predict recurrence risk, the recurrence classifier significantly predicted the 10-year recurrence risk in an independent validation set with high accuracy (85%). This tool showed superior accuracy, specificity, positive predictive value, concordance, and hazard ratios relative to tested clinicopathological variables in predicting recurrences ($p < 0.0001$). Furthermore, it significantly identified patients that might benefit from additional therapy (validation cohort $p = 0.0006$) [59].

While most work primarily focused on analysing cells of epithelial origin within the tumour, several papers also considered tumour stroma for its prognostic pattern. *Beck* et al. [60] showed that the features relating to morphology as well as to spatial relationships and global image features of epithelial and stromal regions were extracted from digitized WSIs of specimens from patients with breast cancer. The features were used to train a prognostic model, and were found to be strongly associated with overall survival (OS) in cohorts of patients with breast cancer from two different institutions; features extracted from the stromal compartment had a stronger prognostic value ($P = 0.004$) than features extracted from the epithelial compartment ($P = 0.02$).

2.3. Challenges to AI application in breast pathology

With the advancement of AI applications in breast pathology, several challenges remain to be solved. For example the low penetration rate of digital pathology, despite rapid technological innovation and sustained enthusiasm for digital pathology among pathologists and researchers [61,62]. In this section, we discuss challenges that impede adoption and development of AI in breast pathology.

2.4. Data and image quality

AI-based approach is markedly dependent on the quantity and quality of the input data, for instance the lack of agreed file format for digitized slides and absence of information system integration [61]. Also the data used in training an AI algorithm should be clean, artefact free, and comprehensive to develop a model that has good predictive performance [22,62–64].

For image-based models, image quality can have a large impact on AI performance. If the scanner cannot produce images at high enough resolution, the model will not be able to distinguish details required to make assessment of tissues. Out of focusing is another challenging issue. Scans at high magnification tend to have limited focusing range, and scanned images cannot be refocused after the fact (unlike slides being observed under the microscope). This becomes problematic when small objects such as mitotic figures are under consideration as refocusing digital image will require a rescan. *Kohlberger* et al. developed a technique that uses synthetic data to train a CNN to detect out of focus area, which helps ensure image quality of the scanned images [65].

Furthermore, the high degree of variability in morphological features and biological structures might affect the performance of the algorithms. Several factors such as staining, orientation, and magnification of the biological sample contribute to the visual heterogeneity of the images. Moreover, illumination variations and the level of noise are affecting the morphological and architectural structure of the histological regions [62]. Furthermore, challenges such as foreground-background intensity overlaps, partial occlusion, touching objects, and weak boundaries are usually presented with histological images that make it hard to distinguish between different classes in the images. Another challenge problem is the computational efficiency of the analysis methodology and the sensitivity to the parameter settings.

Besides the quality of input data, the quality of annotation is equally, if not more, important. For an AI approach to segment biological structures, the performance is dependent on the fidelity of the annotations by expert pathologists in the learning set [30,66]. If the annotations contain high variability, the supervisory signal to the model will be inconsistent, and the model can expect to fail. Furthermore, the evaluation of model performance is often against the reference standard. The rigor of the reference standard decides the trustworthiness of the evaluation results. Situations such as those discussed above warrant the need for the creation of accurately annotated reference datasets by expert pathologists in order to standardize the evaluation of the performance of AI algorithms.

2.5. Algorithm validation

Once we have a developed ML-based tools for a specific task, it is important to also consider how the model can be integrated in the real world. This is challenging because proper evaluation can be very different for different use cases. Lack of proper evaluation can hinder trust from physicians and impede adoption of AI in breast pathology.

First, the validation should be appropriate for the anticipated use cases [5,67]. For example, in the pre-diagnosis and post-diagnosis use cases, ML-based tools need to be adequately validated using representative multi-institutional data to ensure generalization of the approaches and interoperability. Furthermore, retrospective evaluation dataset may contain unexpected biases that cause failure in the real world. Prospective studies can improve trust in the ML model as its performance is proven over time, but they are also much more challenging to implement because they require integration with real world clinical workflow.

On the other hand, in the peri-diagnosis use case, a model is used as an assisted tool to the pathologist during slide review. To validate an ML-based tool for peri-diagnosis, additional evaluation on the human-computer interface through a multi-case multi-reader study is required [68]. In this regard, interpretability of the model becomes an important topic [64,69,70]. In the medical community, the lack of interpretability could hinder trusts from physicians [70]. If the physician cannot understand why the algorithms makes the decision, they may be forced to ignore the algorithm's decision, limiting its usefulness. Displaying confidence level or limiting the amount of information shown to the physician may alleviate this issue [34]. Previous study [35] also explored interpretability methods [71,72] to understand what were the input features that triggers the model's activation in identifying tumor. Nonetheless, improving interpretability remains an active area of research both in the medical domain, and in the AI community at large.

2.6. Adoption of AI-based tools and future perspectives

As digital pathology continues to spread, larger and richer datasets will become available enabling the development of increasingly accurate models. Over time, we expect widespread adoption of digital pathology, but the question of whether or not this will extend to the acceptance of AI-based diagnostic tools is less clear. The first major hurdles are scientific and regulatory: researchers need to consistently demonstrate that their models can achieve clinically useful performance on relevant tasks using data from a diverse set of patients from a wide variety of institutes, scanners, and slide preparation processes. Commercial and academic entities will need to collaborate to bring the best performing approaches over the regulatory finish line.

Assuming that regulators can ultimately be convinced that ML models are as reliable as pathologists, a potentially more challenging question remains: will pathologists actually adopt these tools? The most compelling argument for adoption would be an improvement in diagnostic accuracy. Considering that the ground truth labels in most digital pathology ML experiments come from pathologist annotaters, the best performing models will in general only be able to match human performance rather than exceed it. There are two notable exceptions here. First, a model may more closely match specialist performance compared to general pathologists on a given task, driving adoption of the model by general pathologists but not necessarily specialists. Second, prognostic labels, such as disease-specific and overall survival, response to therapy, and other outcome variables, provide a modelling target that is not bound by human-generated labels. If a model can provide a better prediction of how long a given patient is likely to live than existing risk stratification systems, it is likely to find widespread use.

A second motivating factor in the adoption of ML tools for digital pathology is potential workflow improvements. Automated tools may make individual pathologists more efficient, particularly at certain laborious tasks like counting mitoses. As discussed above, the present evidence for the impact of digital pathology itself on workflow efficiency is mixed and it seems premature to speculate on whether even the best machine learning tools could compensate for a potential 20% increase in per-slide reading time. However, we can say with confidence that just as the impact of adopting digital pathology depends on the execution - training methods, software design, and integration with the existing toolset - so too will the success of machine learning in digital pathology depend on the details of its implementation. An algorithm is not a software tool and even the best model won't be useful if its predictions are not presented in an understandable way. All too often researchers adopt an “if we build it, they will come” mentality. It's our belief that for machine learning to actually impact breast cancer patients, leading researchers need to recognize and shoulder (at least in part) the responsibility for developing tools that people want to use.

An underlying question for the application of ML to digital breast pathology is: to what extent should ML tools align with or rely upon existing expert understanding of a given process? Consider prognostic predictions for new breast cancer diagnoses. One approach would be to attempt to predict known correlates of survival, such as histological grade, tumor size, biomarker status, tumor-infiltrating lymphocytes, and then to use these features to train a survival prediction model. An alternative approach is to “start from scratch”: rather than attempting to recapture what we already know about breast cancer survival; we attempt to model the relationship between the tissue and survival directly. This approach has the potential to capture subtle patterns and indicators of survival that have never occurred to us, but generally requires a much larger dataset. In practice, this doesn't have to be a binary decision. We can build in expert knowledge when we can, but allow models to pick up on unimagined signals, hopefully finding the balance which leads to the best performance given the data we have. This outlook mirrors a general trend in machine learning away from hand-engineered features towards models which capture the informational structure of the inputs but allow for extremely open-ended function approximation.

There is a strong correlation between morphology and underlying molecular features and this could be the basis of AI application in breast cancer to decipher and predict relevant molecular alterations. AI-based tools can be used to predict biomarker status of clinical relevance including ER, PR, Ki67 and HER2 status in addition to the intrinsic molecular subtypes. Combining the power of AI with autofluorescence or spectroscopic image technology can also provide a potential powerful tool to characterise breast cancer, differentiate between benign and malignant, in situ and invasive malignant lesions.

3. Conclusions

The widespread use of WSI technology for primary diagnosis of breast pathology will enable the adoption of AI-based tools. The applications of AI in the field of breast pathology is increasing and it is expected that will not only complement the work of breast pathologists, reduce their workload and improve their diagnostic accuracy but also provide information beyond that can be gain by eyeball assessment of morphological features with the potential to replace some of the expensive multigene assays to predict the outcome of breast cancer.

References

[1] Gomes DS, Porto SS, Balabram D, Gobbi H. Inter-observer variability between general pathologists and a specialist in breast pathology in the diagnosis of lobular neoplasia, columnar cell lesions, atypical ductal hyperplasia and ductal carcinoma in situ of the breast. Diagn Pathol 2014;9:121.

[2] Allison KH, Reisch LM, Carney PA, et al. Understanding diagnostic variability in breast pathology: lessons learned from an expert consensus review panel. Histopathology 2014;65(2):240–51.

[3] Robertson S, Azizpour H, Smith K, Hartman J. Digital image analysis in breast pathology-from image processing techniques to artificial intelligence. Transl Res 2018;194:19–35.

[4] Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. J Am Med Assoc 2016;316(22):2402–10.

[5] Liu Y, Chen P-HC, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. J Am Med Assoc 2019;322(18):1806–16.

[6] Rakha EA, Aleskandarani M, Toss MS, et al. Breast cancer histologic grading using digital microscopy: concordance and outcome association. J Clin Pathol 2018;71(8):680–6.

[7] Rakha EA, Bennett RL, Coleman D, Pinder SE, Ellis IO. UK national coordinating committee for breast pathology (EQA scheme steering committee). Review of the national external quality assessment (EQA) scheme for breast pathology in the UK. J Clin Pathol 2017;70(1):51–7.

[8] Rakha EA, Ahmed MA, Aleskandarany MA, et al. Diagnostic concordance of breast pathologists: lessons from the national health service breast screening programme pathology external quality assurance scheme. Histopathology 2017;70(4):632–42.

[9] Williams BJ, Hanby A, Millican-Slater R, Nijhawan A, Verghese E, Treanor D. Digital pathology for the primary diagnosis of breast histopathological specimens: an innovative validation and concordance study on digital pathology validation and training. Histopathology 2018;72(4):662–71. https://doi.org/10.1111/his.13403.

[10] Williams BJ, DaCosta P, Goacher E, Treanor D. A systematic analysis of discordant diagnoses in digital pathology compared with light microscopy. Arch Pathol Lab Med 2017;141(12):1712–8. https://doi.org/10.5858/arpa.2016-0494-oa.

[11] Retamero JA, Aneiros-Fernandez J, del Moral RG. Complete digital pathology for routine histopathology diagnosis in a multicenter hospital network. Archives of Pathology & Laboratory Medicine; 2019. https://doi.org/10.5858/arpa.2018-0541-oa.

[12] Prewitt JM, Mendelsohn ML. The analysis of cell images. Ann N Y Acad Sci 1966;128(3):1035–53.

[13] Mukhopadhyay S, Feldman MD, Abels E, et al. Whole slide imaging versus microscopy for primary diagnosis in surgical pathology. Am J Surg Pathol 2017;1. https://doi.org/10.1097/pas.0000000000000948.

[14] [No title], https://www.accessdata.fda.gov/cdrh_docs/reviews/K172174.pdf. [Accessed 20 November 2019].

[15] [No title], https://www.accessdata.fda.gov/cdrh_docs/pdf19/K190332.pdf. [Accessed 20 November 2019].

[16] Mills AM, Gradecki SE, Horton BJ, et al. Diagnostic efficiency in digital pathology: a comparison of optical versus digital assessment in 510 surgical pathology cases. Am J Surg Pathol 2018;42(1):53–9.

[17] Williams BJ, Bottoms D, Treanor D. Future-proofing pathology: the case for clinical adoption of digital pathology. J Clin Pathol 2017;70(12):1010–8.

[18] Bishop C. Pattern recognition and machine learning. first ed. New York: Springer-Verlag; 2006.

[19] Haykin S. Neural networks: a comprehensive foundation. first ed. Upper Saddle River, NJ, USA: Prentice Hall PTR; 1994.

[20] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. Advances in neural information processing systems 25. Curran Associates, Inc.; 2012. p. 1097–105.

[21] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44.

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

[23] Chen J, Srinivas C. Automatic lymphocyte detection in H&E images with deep neural networks. December 2016. arXiv [csCV], http://arxiv.org/abs/1612.03217.

[24] Garcia E, Hermoza R, Castanon CB, Cano L, Castillo M, Castanñeda C. Automatic lymphocyte detection on gastric cancer IHC images using deep learning. In: 2017 IEEE 30th international symposium on computer-based medical systems (CBMS). ; 2017:200-204.

[25] Basavanhally AN, Ganesan S, Agner S, et al. Computerized image-based detection and grading of lymphocytic infiltration in HER2+ breast cancer histopathology. IEEE Trans Biomed Eng 2010;57(3):642–53.

[26] Lu C, Xu H, Xu J, Gilmore H, Mandal M, Madabhushi A. Multi-pass adaptive voting for nuclei detection in histopathological images. Sci Rep 2016;6:33985.

[27] Sornapudi S, Stanley RJ, Stoecker WV, et al. Deep learning nuclei detection in digitized histology images by superpixels. J Pathol Inform 2018;9:5.

[28] Höfener H, Homeyer A, Weiss N, Molin J, Lundström CF, Hahn HK. Deep learning nuclei detection: a simple approach can deliver state-of-the-art results. Comput Med Imag Graph 2018;70:43–52.

[29] Xu J, Luo X, Wang G, Gilmore H, Madabhushi A. A Deep Convolutional Neural Network for segmenting and classifying epithelial and stromal regions in histopathological images. Neurocomputing 2016;191:214–23.

[30] Cruz-Roa A, Gilmore H, Basavanhally A, et al. Accurate and reproducible invasive breast cancer detection in whole-slide images: a Deep Learning approach for quantifying tumor extent. Sci Rep 2017;7:46450.

[31] Han Z, Wei B, Zheng Y, Yin Y, Li K, Li S. Breast cancer multi-classification from histopathological images with structured deep learning model. Sci Rep 2017;7(1):4172.

[32] Spanhol FA, Oliveira LS, Petitjean C, Heutte L. A dataset for breast cancer histopathological image classification. IEEE (Inst Electr Electron Eng) Trans Biomed Eng 2016;63(7):1455–62. https://doi.org/10.1109/tbme.2015.2496264.

[33] Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. J Am Med Assoc 2017;318(22):2199–210.

[34] Steiner DF, MacDonald R, Liu Y, et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. Am J Surg Pathol 2018. https://doi.org/10.1097/PAS.0000000000001151. October.

[35] Liu Y, Kohlberger T, Norouzi M, et al. Artificial intelligence–based breast cancer nodal metastasis detection: insights into the black box for pathologists. Arch Pathol Lab Med 2019;143(7):859–68. https://doi.org/10.5858/arpa.2018-0147-oa.

[36] Veta M, van Diest PJ, Willems SM, et al. Assessment of algorithms for mitosis detection in breast cancer histopathology images. Med Image Anal 2015;20(1):237–48.

[37] Veta M, Heng YJ, Stathonikos N, et al. Predicting breast tumor proliferation from whole-slide images: the TUPAC16 challenge. Med Image Anal 2019;54:111–21.

[38] Tellez D, Balkenhol M, Otte-Holler I, et al. 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 2018. https://doi.org/10.1109/TMI.2018.2820199. March.

[39] Veta M, Kornegoor R, Huisman A, et al. Prognostic value of automatically extracted nuclear morphometric features in whole slide images of male breast cancer. Mod Pathol 2012;25(12):1559–65.

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

[41] Veta M, Huisman A, Viergever MA, van Diest PJ, Pluim JPW. Marker-controlled watershed segmentation of nuclei in H&E stained breast cancer biopsy images. In: 2011 IEEE international symposium on biomedical imaging: from nano to macro; 2011. https://doi.org/10.1109/isbi.2011.5872483.

[42] Veta M, van Diest PJ, Kornegoor R, Huisman A, Viergever MA, Pluim JPW. Automatic nuclei segmentation in H&E stained breast cancer histopathology images. PLoS One 2013;8(7):e70221. https://doi.org/10.1371/journal.pone.0070221.

[43] Andre F, Ismaila N, Stearns V. Use of biomarkers to guide decisions on adjuvant systemic therapy for women with early-stage invasive breast cancer: ASCO clinical practice guideline update summary. J Oncol Pract 2019;15(9):495–7.

[44] Fitzgibbons PL, Dillon DA, Alsabeh R, et al. Template for reporting results of biomarker testing of specimens from patients with carcinoma of the breast. Arch Pathol Lab Med 2014;138(5):595–601.

[45] Thomson TA, Hayes MM, Spinelli JJ, et al. HER-2/neu in breast cancer: interobserver variability and performance of immunohistochemistry with 4 antibodies compared with fluorescent in situ hybridization. Mod Pathol 2001;14(11):1079–86.

[46] Mengel M, von Wasielewski R, Wiese B, Rüdiger T, Müller-Hermelink HK, Kreipe H. Inter-laboratory and inter-observer reproducibility of immunohistochemical assessment of the Ki-67 labelling index in a large multi-centre trial. J Pathol 2002;198(3):292–9. https://doi.org/10.1002/path.1218.

[47] Wang S, Hossein Saboorian M, Frenkel EP, et al. Assessment of HER-2/neu status in breast cancer. Am J Clin Pathol 2001;116(4):495–503. https://doi.org/10.1309/tmuw-g4wb-lxj2-fudn.

[48] Rexhepaj E, Brennan DJ, Holloway P, et al. Novel image analysis approach for quantifying expression of nuclear proteins assessed by immunohistochemistry: application to measurement of oestrogen and progesterone receptor levels in breast cancer. Breast Canc Res 2008;10(5). https://doi.org/10.1186/bcr2187.

[49] Skaland I, Ovestad I, Janssen EAM, et al. Comparing subjective and digital image analysis HER2/neu expression scores with conventional and modified FISH scores in breast cancer. J Clin Pathol 2007;61(1):68–71. https://doi.org/10.1136/jcp.2007.046763.

[50] Couture HD, Williams LA, Geradts J, et al. Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. NPJ Breast Cancer 2018;4:30.

[51] Shamai G, Binenbaum Y, Slossberg R, Duek I, Gil Z, Kimmel R. Artificial intelligence algorithms to assess hormonal status from tissue microarrays in patients with breast cancer. JAMA Network Open 2019;2(7):e197700. https://doi.org/10.1001/jamanetworkopen.2019.7700.

[52] Sahiner B, Tozbikian G, Lozanski G, Gurcan M, Senaras C. Creating synthetic digital slides using conditional generative adversarial networks: application to Ki67 staining. Med Imag 2018: Dig Pathol 2018. https://doi.org/10.1117/12.2294999.

[53] Osareh A, Shadgar B. Machine learning techniques to diagnose breast cancer. In: 2010 5th international symposium on health informatics and bioinformatics; 2010. p. 114–20.

[54] Whitney J, Corredor G, Janowczyk A, et al. Quantitative nuclear histomorphometry predicts oncotype DX risk categories for early stage ER+ breast cancer. BMC Canc 2018;18(1):610.

[55] Saltz J, Gupta R, Hou L, et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. Cell Rep 2018;23(1):181–93. e7.

[56] Website. the spatial heterogeneity and molecular correlates of lymphocytic infiltration in triple-negative breast cancer. J R Soc Interface 2015. https://doi.org/10.1098/rsif.2014.1153. . [Accessed 24 November 2019].

[57] Heindl A, Sestak I, Naidoo K, Cuzick J, Dowsett M, Yuan Y. Relevance of spatial heterogeneity of immune infiltration for predicting risk of recurrence after endocrine therapy of ER+ breast cancer. J Natl Cancer Inst 2018;110(2). https://doi.org/10.1093/jnci/djx137.

[58] Ali HR, Dariush A, Provenzano E, et al. Computational pathology of pre-treatment biopsies identifies lymphocyte density as a predictor of response to neoadjuvant chemotherapy in breast cancer. Breast Cancer Res 2016;18(1):21.

[59] Klimov S, Miligy IM, Gertych A, et al. A whole slide image-based machine learning approach to predict ductal carcinoma in situ (DCIS) recurrence risk. Breast Cancer Res 2019;21(1):83.

[60] Beck AH, Sangoi AR, Leung S, et al. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. Sci Transl Med 2011;3(108):108ra113.

[61] Hartman DJ, Pantanowitz L, McHugh JS, Piccoli AL, OLeary MJ, Lauro GR. Enterprise implementation of digital pathology: feasibility, challenges, and opportunities. J Digit Imaging 2017;30(5):555–60.

[62] Higgins C. Applications and challenges of digital pathology and whole slide imaging. Biotech Histochem 2015;90(5):341–7.

[63] Madabhushi A, Lee G. Image analysis and machine learning in digital pathology: challenges and opportunities. Med Image Anal 2016;33:170–5.

[64] Tizhoosh HR, Pantanowitz L. Artificial intelligence and digital pathology: challenges and opportunities. J Pathol Inform 2018;9:38.

[65] Kohlberger T, Liu Y, Moran M, et al. Whole-slide image focus quality: automatic assessment and impact on AI cancer detection. arXiv [csCV], http://arxiv.org/abs/1901.04619; January 2019.

[66] Doyle S, Feldman M, Tomaszewski J, Madabhushi A. A boosted Bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies. IEEE Trans Biomed Eng 2012;59(5):1205–18.

[67] Chen P-HC, Liu Y, Peng L. How to develop machine learning models for healthcare. Nat Mater 2019;18(5):410–4. https://doi.org/10.1038/s41563-019-0345-0.

[68] Gallas BD, Chan H-P, D'Orsi CJ, et al. Evaluating imaging and computer-aided detection and diagnosis devices at the FDA. Acad Radiol 2012;19(4):463–77.

[69] Khurd P, Bahlmann C, Maday P, et al. COMPUTER-AIDED gleason grading OF prostate cancer histopathological images using texton forests. In: Proc IEEE Int Symp Biomed Imaging, vol. 14–17; 2010. p. 636–9. April 2010.

[70] Ching T, Himmelstein DS, Beaulieu-Jones BK, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface 2018;15(141). https://doi.org/10.1098/rsif.2017.0387.

[71] Xu K, Ba J, Kiros R, et al. Show, attend and tell: neural image caption generation with visual attention. arXiv [csLG], http://arxiv.org/abs/1502.03044; February 2015.

[72] Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. SmoothGrad: removing noise by adding noise. arXiv [csLG], http://arxiv.org/abs/1706.03825; June 2017.

[73] Osareh A, Shadgar B. Microarray data analysis for cancer classification. In: 2010 5th international symposium on health informatics and bioinformatics; 2010. https://doi.org/10.1109/hibit.2010.5478893.

[74] Al-Kofahi Y, Lassoued W, Lee W, Roysam B. Improved automatic detection and segmentation of cell nuclei in histopathology images. IEEE (Inst Electr Electron Eng) Trans Biomed Eng 2010;57(4):841–52. https://doi.org/10.1109/tbme.2009.2035102.

[75] Hossain MS, Hanna MG, Uraoka N, et al. Automatic quantification of HER2 gene amplification in invasive breast cancer from chromogenic in situ hybridization whole slide images. J Med Imaging 2019;6:1. https://doi.org/10.1117/1.jmi.6.4.047501. 04.

[76] Website, Xu Z, Moro CF, Bozóky B, Zhang Q. GAN-based virtual re-staining: a promising solution for whole slide image analysis. ArXiv.org, https://arxiv.org/abs/1901.04059. [Accessed 28 November 2019].

[77] Website. Automatic lymphocyte detection in H&E images with deep neural networks. 2016. . [Accessed 28 November 2019]. https://arxiv.org/abs/1612.03217.

[78] Basavanhally AN, Ganesan S, Agner S, et al. Computerized image-based detection and grading of lymphocytic infiltration in HER2 breast cancer histopathology. IEEE (Inst Electr Electron Eng) Trans Biomed Eng 2010;57(3):642–53. https://doi.org/10.1109/tbme.2009.2035305.

[79] Website, Yuan Y. Modelling the spatial heterogeneity and molecular correlates of lymphocytic infiltration in triple-negative breast cancer. J R Soc Interface 2015. https://doi.org/10.1098/rsif.2014.1153. . [Accessed 28 November 2019].

[80] Tőkés T, Tőkés A-M, Szentmártoni G, et al. Expression of cell cycle markers is predictive of the response to primary systemic therapy of locally advanced breast cancer. Virchows Arch 2016;468(6):675–86.