Medical Image Analysis 89 (2023) 102891

Contents lists available at ScienceDirect

Medical Image Analysis

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

Minimum resolution requirements of digital pathology images for accurate classification

Lydia Neary-Zajiczek [a,b,*], Linas Beresna [b,1], Benjamin Razavi [c], Vijay Pawar [b,2], Michael Shaw [a,b,d], Danail Stoyanov [a,b]

[a] Wellcome/EPSRC Centre for Interventional and Surgical Sciences, Charles Bell House, 43-45 Foley Street, Fitzrovia, London, W1W 7TS, United Kingdom
[b] Department of Computer Science, University College London, Gower Street, London, WC1E 6BT, United Kingdom
[c] University College London Medical School, 74 Huntley Street, London, WC1E 6BT, United Kingdom
[d] National Physical Laboratory, Hampton Road, Teddington, TW11 0LW, United Kingdom

ARTICLE INFO

Dataset link: https://iciar2018-challenge.grand-challenge.org/, https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/, https://www.kaggle.com/c/histopathologic-cancer-detection

Keywords:
Digital pathology
Image quality
Deep learning
Automated diagnostics

ABSTRACT

Digitization of pathology has been proposed as an essential mitigation strategy for the severe staffing crisis facing most pathology departments. Despite its benefits, several barriers have prevented widespread adoption of digital workflows, including cost and pathologist reluctance due to subjective image quality concerns. In this work, we quantitatively determine the minimum image quality requirements for binary classification of histopathology images of breast tissue in terms of spatial and sampling resolution. We train an ensemble of deep learning classifier models on publicly available datasets to obtain a baseline accuracy and computationally degrade these images according to our derived theoretical model to identify the minimum resolution necessary for acceptable diagnostic accuracy. Our results show that images can be degraded significantly below the resolution of most commercial whole-slide imaging systems while maintaining reasonable accuracy, demonstrating that macroscopic features are sufficient for binary classification of stained breast tissue. A rapid low-cost imaging system capable of identifying healthy tissue not requiring human assessment could serve as a triage system for reducing caseloads and alleviating the significant strain on the current workforce.

1. Introduction

Pathology services are a critical component of integrated health care systems and underpin many patient pathways, particularly the diagnosis and treatment of cancer. The need for increased pathology capacity has become more urgent in recent years; according to Cancer Research U.K. (2016), cellular pathology requests have increased by 4.5% on average each year and are becoming more complex as health services target early diagnosis (Williams et al., 2017). Despite this increased demand, maintaining adequate staffing levels remains an ongoing challenge. The current vacancy rate for pathologists in England is 12.5%, while only 3% of surveyed UK pathology departments report having enough staff. Further compounding the difficulty is low uptake of training places to replace those due to retire imminently, as 25% of all histopathologists are aged 55 and older (The Royal College of Pathologists, 2018). Similar workforce trends have been seen in other high-income countries including Spain (Retamero et al., 2020), Canada and the United States (Metter et al., 2019), and shortages are even more severe in low- and middle-income countries (Wilson et al., 2018; Mudenda et al., 2020). Planned expansion of existing screening programs such as those in place for breast cancer, for example, would further exacerbate an already serious problem.

Most professional membership organizations for clinical pathologists have identified digitization of pathology workflows as a key mitigation strategy to address this crisis, where physical samples are scanned using a whole-slide imager (WSI[3]) to create digital or "virtual" slides, flexible file formats that are assessed on a computer workstation. Digitization provides opportunities for increased efficiency in assigning, managing and auditing cases (Williams et al., 2017), allows for the provision of remote diagnostics for under-served areas (Pare et al., 2016) and generates invaluable resources for teaching and research (Hamilton et al., 2012). The eventual availability of automated diagnostic tools is the ultimate goal of digitization, and would significantly alleviate

* Correspondence to: LUMA Vision Ltd., Block C, Parkview House, Beech Hill Road, Dublin, D04 K5D0, Ireland.
E-mail address: lydia.zajiczek.17@ucl.ac.uk (L. Neary-Zajiczek).
[1] Current affiliation: School of Computing Science, Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada.
[2] Current affiliation: Bartlett School of Sustainable Construction, University College London, 1-19 Torrington Place, London, WC1E 7HB, United Kingdom.
[3] WSI is also used as an acronym for "whole-slide image".

https://doi.org/10.1016/j.media.2023.102891
Received 19 April 2022; Received in revised form 22 May 2023; Accepted 6 July 2023
Available online 13 July 2023

staffing pressures on pathology departments worldwide (Chang et al., 2019).

Despite these clear benefits, routine clinical use of WSIs for diagnosis is limited. In a 2018 survey by Williams et al., 60% of participating UK pathology departments had access to a slide scanner, but only 31% of departments reported using digital slides for primary diagnosis, with the majority of use cases being for education. This is in stark contrast to radiology departments, with some hospital trusts in the United Kingdom becoming fully digitized as early as 2007 (Barham and Madden, 2007). Two major contributing factors unique to pathology have been cost (Griffin and Treanor, 2017) and pathologist perception of poor image quality (Flotte and Bell, 2018). These two issues are closely linked: demands for increased image quality have spurred incredible technological development of slide scanners, requiring a large capital outlay for digitization.

Modern ultra-fast slide scanners image samples in colour at very high magnification (usually up to 40× for histopathology). The technical challenges of capturing high-quality, high-resolution and in-focus images at the speed required to process the volume of cases generated by a typical pathology department[4] have resulted in scanners that can cost up to $500,000 USD. In addition to the large capital outlay, scanning at such volumes incurs significant operating costs including "companion software, data storage, and personnel costs" (Liu and Pantanowitz, 2019). Data storage is a particular challenge, as a single digital slide is on the order of 1 GB (Zarella et al., 2019).

The demand for scanning images at such high resolution was driven in large part by pathologist assessment of inadequate digital image quality. A number of studies have investigated the impact of image quality on diagnostic concordance and confidence (Krupinski et al., 1999; Williams et al., 2003; Krupinski et al., 2012a,b; Shrestha et al., 2015, 2016). Most concluded, however, that optimal image quality was not necessarily crucial for an accurate diagnosis: (Krupinski et al., 2012a) for example found that "images may be compressible to relatively high levels before impacting WSI interpretation performance". Validation studies comparing diagnostic concordance between WSI and conventional light microscopy (CLM) have found little evidence that accuracy is reduced when using WSI (Mukhopadhyay et al., 2018), however confidence is generally lower for WSI (Goacher et al., 2017), suggesting that the issue is perception of inadequate image quality rather than the absolute quality itself. In a 2015 study, Shrestha et al. identified the most important image quality metrics required by pathologists for an accurate and confident diagnosis as (in order): sharpness, contrast, brightness, uniform illumination and colour separation. This study however relied on a subjective measure of image quality assigned on a numeric scale by pathologists when assessing images with degradation of one of the five characteristics. Dodge and Karam (2016) investigated the effects of similar degradations on deep learning algorithms deployed on the ImageNet object classification dataset (Russakovsky et al., 2015), finding the trained networks most sensitive to image blur and noise, with blurred images obtained by applying a Gaussian kernel of increasing width. Noise is not generally an issue in brightfield microscopy, and the blur of low-resolution imaging systems is an Airy disk kernel rather than a Gaussian. More importantly, however, the kernel sizes were not related to physical imaging parameters such as macroscopic lens f-number, for example.

In this work we quantitatively determine the minimum image quality requirements necessary for an accurate preliminary diagnosis of the malignancy of breast tissue, focusing on the first two most important quality metrics as identified by Shrestha et al. (2015): sharpness and contrast, which are both dependent on the spatial resolution of the imaging system. Out-of-focus errors were also the most common contributor to discordant cases when comparing WSI and CLM-based diagnoses (Gilbertson et al., 2006; Snead et al., 2016; Araújo et al., 2019). We provide a theoretical description of image resolution both in terms of the spatial frequency support of the optical system used to generate the magnified image and the digital sampling frequency of the camera sensor used to capture it. We train a binary classification algorithm using publicly available datasets of breast histology images to achieve a baseline accuracy comparable to the current state of the art, and deploy the model on test images that have been computationally degraded using our derived framework to identify the size of learned features needed for robust diagnostics. We also re-train the classification algorithms on degraded images to determine if larger macroscopic features are sufficient for binary classification. Our results show that lower resolution imaging feasible for this task. In the context of high-volume national screening programs such as those in place for breast cancer, pairing a low-resolution imaging system with a sufficiently sensitive automated classifier could potentially reduce the number of cases needing to be scanned at full resolution and assessed by a human pathologist.

2. Theory and background

The spatial resolution of a shift-invariant imaging system can be quantified through its impulse response or point spread function (PSF), which is defined as the output of the system with a point source as its input (Goodman, 2005). The image amplitude $A(x, y)$ captured by the system is the object amplitude $O(x, y)$ convolved with the system's PSF $h(x, y)$, or

$$A(x, y) = h(x, y) * O(x, y). \tag{1}$$

For coherent illumination[5], Eq. (1) describes the linearity of an optical system's impulse response to the object's complex amplitude. Brightfield microscopy generally uses an incandescent lamp as an extended illumination source and is thus an incoherent imaging modality. Consequently, the phase of the incident electromagnetic field varies in an uncorrelated way, and the impulse response at all points must be computed on an intensity basis. Eq. (1) then becomes for image intensity I:

$$I(x, y) = |h(x, y)|^2 * |O(x, y)|^2. \tag{2}$$

The amplitude PSF h is equal to the Fraunhofer diffraction pattern of the exit pupil function P, which for a rotationally symmetric imaging system consisting of an aberration-free objective lens of numerical aperture NA and a matched tube lens resulting in a system magnification M is equal to unity with radius

$$r_{\text{pupil}} = f_{\text{obj}} \times \text{NA}_{\text{obj}}, \quad f_{\text{obj}} = \frac{f_{\text{tube}}}{M}. \tag{3}$$

The NA, $n \sin \alpha$, is defined by the maximum acceptance angle α of the objective lens and the refractive index n of the objective immersion medium. The unaberrated pupil function P is a top-hat function with a radius r_{pupil}:

$$P(r, \theta) = \begin{cases} 1, & r \le r_{\text{pupil}} \\ 0, & r > r_{\text{pupil}} \end{cases} \tag{4}$$

The intensity PSF is therefore equal to

$$|h(x, y)|^2 = |\mathcal{F}\{P(u, v)\}|^2, \tag{5}$$

which for a top-hat pupil function is an Airy disk (Airy, 1835). Eq. (2) can be simplified using the convolution theorem (Bracewell, 1999):

$$\mathcal{F}\{I(x, y)\} = \mathcal{F}\{|h(x, y)|^2\} \times \mathcal{F}\{|O(x, y)|^2\}. \tag{6}$$

[4] Retamero et al. (2020) found that the Philips IntelliSite Pathology Solution, for example, scans an average glass slide in 114 s and estimated the slide volume generated in Grenada University Hospital's central pathology laboratory at 700 slides per day.

[5] Goodman (2005) gives a simple definition of spatially coherent illumination as originating from a point source.

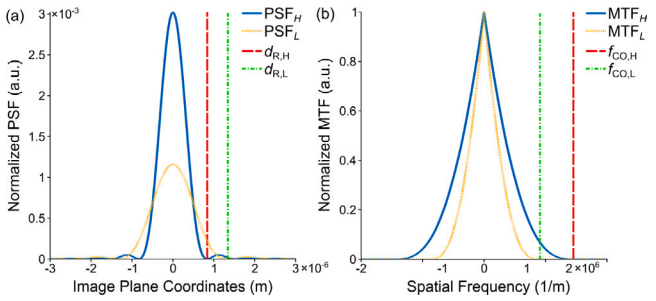

Fig. 1. (a) Intensity PSF $|h(x,y)|^2$ and (b) MTF $|H(f_x,f_y)|$ of $\mathrm{NA}_H = 0.4$ (solid blue line) and $\mathrm{NA}_L = 0.25$ (dotted yellow line) objectives, respectively, plotted across centre of image plane. Overlaid vertical lines are (a) Rayleigh criteria (Eq. (9)) and (b) incoherent cutoff frequencies (Eq. (10)) for NA_H (red dashed line) and NA_L (green dashed–dotted line) objectives, respectively. Magnification is 10× and wavelength is 550 nm.

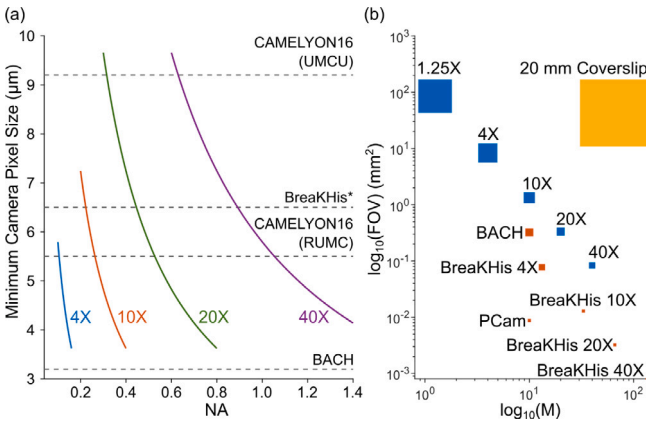

Fig. 2. (a) Minimum pixel size d_{pixel} required for diffraction-limited imaging ($f_s > f_N$) at different magnifications and NAs. Actual d_{pixel} values used to capture each dataset are overlaid as dashed grey lines for comparison. (b) Field of view (FOV) of typical sCMOS sensor 2048 × 1536 pixels in size with 6.5 µm wide pixels for different magnifications (blue boxes). Orange boxes show FOV of imaging systems used to capture datasets used in this work. Yellow box shows 20 mm coverslip to illustrate typical imaging area of a microscope slide. We note that the BreaKHis dataset was captured with a 3.3× relay lens to avoid the issue of undersampling at lower magnifications, hence the reduced FOV.

The Fourier transform of the intensity PSF $|h(x,y)|^2$ is defined as the optical transfer function (OTF) H of an incoherent imaging system, or

$$H(f_x,f_y) = \mathcal{F}\{|h(x,y)|^2\}. \tag{7}$$

The OTF $H(f_x,f_y)$ "specifies the complex weighting factor applied by the system to the frequency component at f_x [and] f_y, relative to the weighting factor applied to the zero-frequency component" (Goodman, 2005). Another important property of the OTF is that

$$|H(f_x,f_y)| \leq |H(0,0)|. \tag{8}$$

The modulus of the OTF $|H|$ is referred to as the modulation transfer function (MTF).

We now have sufficient information to quantify the aberration-free frequency response of an incoherent imaging system. To compare two well-corrected objective lenses with identical magnifications but different NAs, we calculate the radius of the pupil function given $M = 10$ and $f_{\mathrm{tube}} = 180$ mm (standard for Olympus microscopes) using Eqs. (3) and (4). An Olympus extended apochromat 10× objective lens, for example, has an NA of 0.4, whereas a Olympus plan achromat 10× objective has an NA of 0.25. The "high" NA 10× objective (0.4) is assigned the subscript H and the "low" NA 10× objective (0.25) is assigned the subscript L.

Cross sections of the squared and normalized[6] intensity PSFs and MTFs calculated using Eqs. (5) and (7) are plotted for both objectives in Fig. 1 at a wavelength of $\lambda = 550$ nm, corresponding to the green channel of a typical RGB camera sensor. Two resolution criteria that are useful as comparison metrics are overlaid on these plots for each of the lenses. The Rayleigh criterion d_R defines the minimum resolvable distance between two point objects for an objective lens with a given NA at a wavelength λ (Born and Wolf, 1999):

$$d_R = 0.61\frac{\lambda}{\mathrm{NA}}. \tag{9}$$

This value corresponds to the first minimum of the PSF of the system as shown in Fig. 1a. The incoherent cutoff frequency f_{CO} defines the bandwidth of our imaging system, and is calculated as (Goodman, 2005):

$$f_{\mathrm{CO}} = 2\frac{\mathrm{NA}}{\lambda}. \tag{10}$$

The imaging system cannot resolve any spatial frequencies beyond this cutoff frequency, and the contrast of the system drops to zero as illustrated in Fig. 1b.

As the output of any slide scanning system is a digital representation of a real image, we must also consider the sampling resolution of the camera sensor. The Shannon–Nyquist sampling theorem defines the minimum sampling frequency necessary to reconstruct a band-limited signal, or diffraction-limited image in this case, as twice the highest spatial frequency present in the image (Marks, 2009). This frequency is referred to as the Nyquist rate and is defined as

$$f_N = 2 \times f_{\mathrm{CO}}. \tag{11}$$

The maximum monochrome sampling frequency of the sensor f_s is set by the effective pixel size d_{eff}, or

$$f_s = \frac{1}{d_{\mathrm{eff}}}, \quad d_{\mathrm{eff}} = \frac{d_{\mathrm{pixel}}}{M}, \tag{12}$$

where d_{pixel} is the physical size of the sensor pixels and M is the system magnification. Our system is therefore limited in resolution both by the frequency response of the optical components and by the sampling resolution of the camera sensor; if the system is sufficiently sampled, *i.e.* if $f_s > f_N$, it will be considered to be diffraction limited or oversampled, whereas if $f_s < f_N$ the system will be considered to be Nyquist limited or undersampled. The consequences of undersampling manifest as aliasing artefacts, where shifted copies of the original image's frequency spectrum occurring at integer multiples of the sampling frequency overlap, resulting in information loss and distortion (Oppenheim et al., 1999). Lower magnification images are often undersampled due to the linear relationship between f_s and M and the physical limitations of camera pixel sizes, as illustrated in Fig. 2a. Smaller pixel sizes result in a higher sampling frequency but at the expense of less light incident on each pixel, which would be prohibitive for imaging modalities such as darkfield or fluorescence microscopy. Increasing the magnification reduces the field of view (FOV) of each image as shown in Fig. 2b, requiring significantly more images to capture an entire physical sample that may be tens of millimetres in size.

The sampling frequency of most systems will in reality be lower than f_s as defined in Eq. (12) due to the wavelength dependence of f_N and the type of camera sensor used. Most RGB cameras use a colour filter array (CFA) to produce a mosaiced image (Swirski, 2009); in the common Bayer CFA pattern, the green channel is sampled twice as often as the red and blue channels, and "the final system resolution is dominated by the green array" (Palum, 2001). Most methods for demosaicing single-channel images produced by a CFA camera into RGB images involve linear interpolation (Malvar et al., 2004). This interpolation

[6] The PSF is normalized such that $\sum_x \sum_y \mathrm{PSF}(x,y) = 1$, and the MTF is normalized such that $|\mathrm{MTF}(0,0)| = 1$.

combined with overlap in the passbands of the array filters means that the channel-specific sampling frequencies are not strictly limited by the geometry of the CFA, *e.g.* the sampling frequency of the red channel is not exactly half that of the green channel. The monochrome sampling frequency f_s will therefore be used in the absence of information about the colour camera sensors for a given imaging system.

Given an intensity image recorded by a high NA imaging system I_H of a real physical object O in the presence of noise ϵ, Eq. (2) becomes (removing the spatial coordinates for brevity)

$$I_H = |h_H|^2 * |O|^2 + \epsilon_H, \tag{13}$$

and the equivalent image recorded by an identical but lower NA imaging system I_L is

$$I_L = |h_L|^2 * |O|^2 + \epsilon_L. \tag{14}$$

The major sources of noise in digital imaging are read noise, or voltage fluctuations during the process of analog to digital conversion (Horowitz and Hill, 2015), shot noise, or a fundamental variability in the number of photons arriving at a sensor due to quantum fluctuations of the incident electromagnetic field (Mandel and Wolf, 1995), thermal noise, pixel response non-uniformity (PRNU) and pattern noise or spatial variation in the noise properties of individual pixels across the sensor (Gonzalez and Woods, 2007). When comparing images from two nominally identical imaging systems with different NAs, thermal noise, PRNU and pattern noise would be effectively uniform across both images. In the case of low photon number N, for example in fluorescence microscopy, the noise signal is often a combination of read and shot noise. In brightfield microscopy, N is generally large, the signal-to-noise ratio (SNR) is equal to $\sqrt{N}$ and ϵ_H and ϵ_L will be dominated by Poisson shot noise, which for large N can be modelled as additive white Gaussian noise (AWGN) having uniform power across the frequency spectrum.

We can computationally degrade the high-NA images I_H of the publicly available datasets to an equivalent image I_L that would be generated by a lower NA system by moving into frequency space and using Eq. (6). The frequency spectrum of the low NA image I_L can be approximated by the following expression, provided $(f_x, f_y) < f_{\mathrm{CO},H}$:

$$\mathcal{F}\{I_L\}(f_x, f_y) \approx \frac{|\mathcal{F}\{|h_L|^2\}(f_x, f_y)|}{|\mathcal{F}\{|h_H|^2\}(f_x, f_y)|} \times \mathcal{F}\{I_H\}(f_x, f_y).$$

While this expression is not exactly equivalent given Eqs. (13) and (14) for I_H and I_L respectively, the general property of the MTF given in Eq. (8) means that the ratio of the MTFs will always be less than unity for any nonzero frequency, and high-frequency AWGN noise ϵ_H present in I_H will be suppressed. The degraded image I_L is computationally approximated as (again removing the spatial coordinates for brevity):

$$I_L = \mathcal{F}^{-1}\left\{\frac{|H_L|}{|H_H|} \times \mathcal{F}\{I_H\}\right\}. \tag{15}$$

This equation defines the degradation function that was applied to the images in each of the histopathology datasets used in this work, using estimates of the parameters of the imaging system used to capture them.

3. Material and methods

We computationally degraded high-quality labelled histopathology images using the degradation model defined by Eq. (15) to systematically determine the minimum image resolution requirements for accurate automated classification using deep learning. These networks were first trained on the original images and deployed on degraded test sets for classification, then re-trained on degraded images and again deployed on similarly degraded images. To generate a realistic degradation model, the original NA of the imaging system (NA_H) was obtained or estimated, as well as the magnification M and the effective pixel size d_{eff}. The following section will outline the datasets used in this work, the methods used to estimate their relevant imaging parameters, the network architecture used for automated binary classification and details about training and testing.

3.1. Datasets

The publicly available datasets used in this study are listed in Table 1. Datasets were chosen based on their ubiquity in the computational pathology research community and the availability of state-of-the-art classification algorithms. This work focused on binary classification as this task is particularly well-suited to automated diagnostics; one such application of an imaging system optimized for cost and accuracy would be to reduce the volume of cases needing to be assessed by a human pathologist as a potential mitigation strategy for the current staffing crisis. Tasks such as pixel-wise segmentation or tissue grading are not considered, as the full replication of human competence is a longer-term goal of digital pathology and is outside the scope of this work.

The majority of labelled histopathology datasets consist of images of breast tissue due to the existence of screening programs in most high-income countries, with tissue sampling followed by pathological assessment forming part of the "triple-test" of breast diagnosis (Ginsburg et al., 2020). The first significant datasets to be released were BreaKHis (Spanhol et al., 2016) and the Bioimaging 2015 challenge dataset (Araújo et al., 2017), the latter of which was expanded into the Grand Challenge on BreAst Cancer Histology images, or BACH (Aresta et al., 2019). Another benchmark dataset, the Cancer Metastases in Lymph Nodes Challenge 2016 (CAMELYON16, Ehteshami Bejnordi et al.) was followed by PatchCamelyon or PCam (Veeling et al., 2018) consisting of labelled patches extracted from CAMELYON16.

The BACH dataset contains both annotated WSIs and microscopy images of breast tissue with four different classification labels. In this work we use only the microscopy images, grouping those with "Normal" and "Benign" labels into a single benign class and those with "In-situ" and "Invasive" carcinoma labels into a single malignant class. There are a total of 400 images in the training set with an additional 100 unlabelled images in the test set. BreaKHis consists of 7,909 labelled images of breast tissue, broadly divided into "Benign" and "Malignant" categories. Finer grading labels are also provided but these are also not used in this work. Images are provided at magnifications of 40×, 100×, 200× and 400×, however the actual objectives used for imaging had magnifications of 4×, 10×, 20× and 40×. This notation is consistent with terminology from CLM-based pathology, where the eyepiece adds an additional ocular magnification of 10×. The images are evenly divided between the four magnifications (approximately 2,000 total images for each) with a ratio of malignant to benign images of 2:1. Finally, PCam consists of 277,483 image patches extracted from the CAMELYON16 dataset at 10× magnification, and the version used in this work is the one hosted on Kaggle with duplicate images removed. Images are divided into "Normal" and "Tumour" classes, where "[a] positive label indicates that the centre 32 × 32 [pixel] region of a patch contains at least one pixel of tumour tissue" (Veeling et al., 2018).

3.2. Imaging system parameters

Images were degraded computationally in MATLAB (MathWorks Inc.) using Eq. (15), where the MTFs H_H and H_L were computed using Eq. (3), (4), (5) and (7) for two different values NA_H and NA_L with M, f_{obj}, d_{eff} being equal. The ratio of the MTFs was applied to the Fourier transform of each colour channel separately over the frequency range $|f_x, f_y| \leq f_{\mathrm{CO},H}$, with $f_{\mathrm{CO},H}$ defined in Eq. (10) for a given NA_H and wavelength λ corresponding to the centre of the spectral passband of a typical filter[7] for that channel, or $\{\lambda_R, \lambda_G, \lambda_B\} = \{625, 550, 475\}$ nm. Scaling of the frequency spectrum was followed by

[7] See, for example, the datasheet of the pco edge 5.5c colour sCMOS camera (pco).

Table 1
Publicly available histopathology datasets used in this study with relevant metadata.

Dataset	Tissue type	Classification labels	Number of images and splits	Citation
BACH (ICIAR2018)	Breast	Normal Benign In-situ carcinoma Invasive carcinoma **500 total**	100/25 train/test 100/25 train/test 100/25 train/test 100/25 train/test	Aresta et al. (2019)
BreaKHis (all)	Breast	Benign* Malignant**	2,480 5,429 **7,909 total**	Spanhol et al. (2016)
PatchCamelyon (PCam)	Breast	Normal Tumour Unlabelled (test set)	130,908 89,117 57,458 **277,483 total**	Veeling et al. (2018)

Grades: *adenosis, fibroadenoma, phyllodes tumour, tubular adenoma **ductal/lobular/mucinous/papillar carcinoma.

Table 2
Relevant imaging parameters of datasets listed in Table 1. Entries in bold denote parameters not explicitly provided in the description of the dataset that were either estimated or acquired from imaging system technical specifications. Entries with † denote provided parameters that are not exactly consistent with other parameters provided in the description and/or obtained from the technical specification. Shaded values of f_s indicate Nyquist-limited/undersampled imaging.

Dataset	M_{obj}	M_{relay}	d_{pixel} (μm)	d_{eff} (μm)	NA	f_s (μm^{-1})	f_N (μm^{-1})	Image size
BACH	20	**0.5**	**3.2**	**0.32**†	**0.30**	3.125	2.526	2048 × 1536
BreaKHis (4×)	4	3.3	6.5	0.49	**0.16**	6.250	1.346	700 × 460
BreaKHis (10×)	10	3.3	6.5	0.20	**0.40**	5.000	3.368	700 × 460
BreaKHis (20×)	20	3.3	6.5	0.10	**0.80**	10.00	6.737	700 × 460
BreaKHis (40×)	40	3.3	6.5	0.05	**1.40**	20.00	11.79	700 × 460
CAMELYON16 (RUMC)[a]	20	-	**5.5**	0.243†	**0.80**	4.115	6.737	WSIs (variable)
CAMELYON16 (UCMU)[b]	**20**	**2**	**9.2**	0.23	**0.75**	4.348	6.316	WSIs (variable)
PatchCamelyon (PCam)	10	-	-	0.972	**0.13**	1.029	1.095	96 × 96

[a]Radboud University Medical Center.
[b]University Medical Center Utrecht.

an inverse Fourier transform and histogram matching using the original image I_H as a reference via the `imhistmatch` function in MATLAB to maintain consistent colour balance. Computing Eq. (15) for each dataset required identifying NA_H, M and d_{eff} for the imaging systems used to generate these datasets and are summarized in Table 2; entries in **bold** denote parameters that were not explicitly provided in the dataset description and were either estimated or determined from the technical specification of the imaging system used. Note that we do not use the CAMELYON16 dataset in this work, but refer to the imaging parameters specified by Ehteshami Bejnordi et al. (2017) to estimate the parameters for PCam.

All dataset descriptions provided the manufacturer and model name of the imaging system, the effective pixel size d_{eff}, and in most cases, the image magnification M. No dataset descriptions specified the imaging system NA, and it was estimated by applying a low-pass filter at a steadily decreasing frequency to a set of test images and measuring the change in structural similarity (SSIM) between the filtered and original images (Wang et al., 2004). A significant change meant the filter frequency had exceeded f_{CO} as defined in Eq. (10); a full description of the method used to estimate NA is given in Supplementary Information. Technical detail regarding the type of RGB camera sensor or demosaicing algorithm used was not provided for any of the datasets used in this work, thus the monochrome sampling frequency f_s was used. We note that there was a discrepancy in the provided value of d_{eff} for the BACH dataset, which is also described in more detail in Supplementary Information. Fig. 3 shows the results of systematically degrading a single training image from the PCam dataset.

3.3. Binary classification architecture

The classification architecture chosen to measure diagnostic accuracy in a repeatable and deterministic manner was based on the method described by Kassani et al. (2020) which used an ensemble of three convolutional neural networks (CNNs) to achieve good classification accuracy across each of the datasets used in this work. The classifiers described here are used to establish a baseline accuracy, and we define a significant drop in accuracy as exceeding 10% and therefore an unacceptable degradation in image quality[8]. The focus in training the classifier is therefore not to achieve the highest accuracy possible in this task, but accuracy in line with the state of the art.

Ensemble networks have an advantage over individual networks in that architecture-specific limitations of any one network can be mitigated against; ensemble networks tend to achieve higher accuracies on datasets than their constituent architectures. In the work of Kassani et al. (2020), the three network architectures in the ensemble were VGG19 (Simonyan and Zisserman, 2015), MobileNetV2 (Howard et al., 2017) and DenseNet201 (Huang et al., 2017), and the ensemble achieved a classification accuracy of 98.13% on the BreaKHis dataset while an individual VGG16 network achieved only 93.54%. The same ensemble also achieved accuracies of 94.64% and 95.00% for PCam and BACH, respectively. The accuracy of a human pathologist for high level diagnostics *i.e.* binary classification of breast tissue is comparable; (Rakha et al., 2017) examined 240 breast lesions from routine practice across the UK's National Health Service, finding only 35 cases (14.6%) with a diagnostic concordance of less than 95% with each case being assessed by 600 participants on average. 13 discordant cases (5.4%) were due to pathologist misinterpretation.

All training and testing was carried out in Keras using a Tensorflow backend (both versions 2.3.0) with the CUDA 10.1 toolkit in Python 3.6.9 on a NVIDIA DGX running the Ubuntu 18.04.6 LTS operating system. The GPU used for training was a single NVIDIA Tesla V100-DGXS with 32 GB of onboard memory. Given the discrepancy in the size of the datasets shown in Table 1, the ensemble network was first trained on PCam, then re-deployed on the BreaKHis and BACH datasets as part

[8] We note that a 10% reduction in diagnostic accuracy would not necessarily be acceptable in routine clinical practice; see discussion in Section 5.

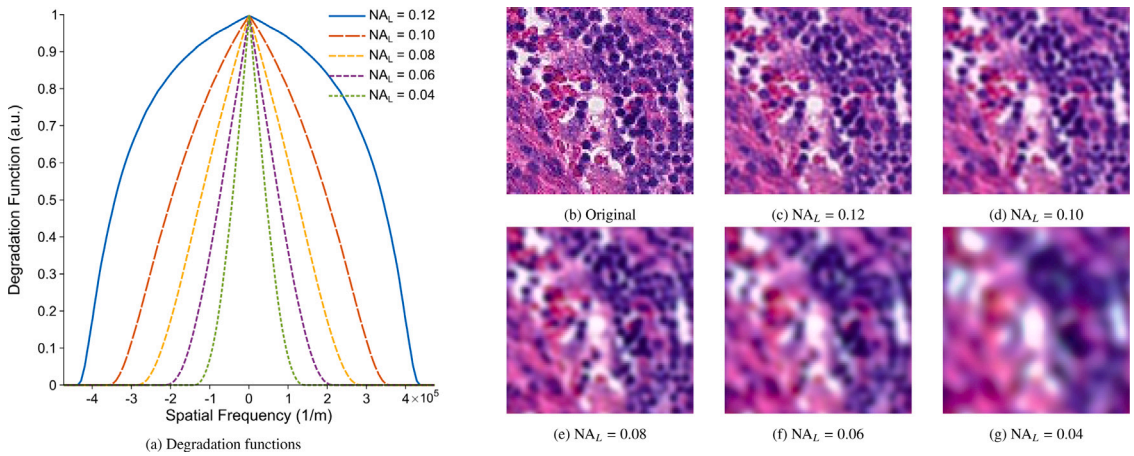

Fig. 3. (a) Degradation functions as defined by Eq. (15) applied to $\mathcal{F}\{I_H\}$ given $\mathrm{NA}_H = 0.13$, M = 10 and $d_{\mathrm{eff}} = 0.972$ μm for different values of NA_L. (b) Original image I_H from training set of PCam showing normal breast tissue. (c)–(g) Resulting degraded images I_L for different values of NA_L.

of a patch-based classifier utilizing transfer learning. This approach also worked well due to the small size of the images in the PCam dataset. Test sets were provided for PCam and BACH, and the training set was randomly divided into training and validation sets in a 80/20 split using the `ImageDataGenerator` class in Keras. BreaKHis was divided randomly into a 75/15/10 train/validation/test split.

Patch classifier

Each of the three CNNs (VGG19, MobileNetV2 and DenseNet201) was loaded in Keras pre-trained on ImageNet without the top classifier layers. The outputs were pooled using the `GlobalAveragePooling2D` layer and concatenated into a single feature layer. Bottleneck features were extracted from the training and validation sets to train the top classifier block, which consisted of a fully connected layer with 128 neurons, a dropout layer (rate of 0.5), a batch normalization layer and finally a dense binary classification layer with a sigmoid activation function. The classifier block was trained for a maximum of 1,000 epochs, keeping all other layers in the ensemble network frozen. Callback functions for early stopping and learning rate reduction on validation loss plateaus meant that training the top classifier usually finished after approximately 30–40 epochs. Training the top classifier block took approximately 23 min.

The final convolutional blocks of each classifier were then unfrozen (layers 16, 97 and 480 onwards in VGG16, MobileNetV2 and DenseNet201, respectively) and fine-tuned for an additional maximum 100 epochs. Identical callbacks were used to avoid overfitting. Fine-tuning also usually finished between 30–40 epochs, taking approximately 6 h. Data augmentation was used during all training stages, including up to 90 degree rotations, horizontal and vertical mirroring and shifts of up to 20% of image height and width). Batch size was set to 32 with an initial learning rate of 10^{-4}. An exponentially decaying learning rate was used, with 10^5 decay steps and a decay rate of 0.96. The models were compiled using the Adam optimizer ($\beta_1 = 0.6$, $\beta_2 = 0.8$) with a binary crossentropy loss function and optimizing for accuracy. The model with the highest validation accuracy was chosen as the best model.

Whole image classifier

The small size of the BreaKHis and BACH datasets necessitated the use of transfer learning, and images were downsampled to approximately match the effective pixel size d_{eff} of PCam (0.972 μm) and contain an integer number of 96 × 96 patches, the input tensor size of the patch-based classifier. The patch-based model was loaded with the same final blocks in each of the three ensemble CNNs set as trainable for fine-tuning. A single 96 × 96 patch was randomly selected from each image in the training and validation sets for each epoch, and the network was re-trained in the same manner as PCam. Validation accuracy as computed during training was only calculated for a single patch extracted from each image in the validation set, thus a custom metric combining training and validation accuracy was used to select the best model, and training loss was monitored for learning rate reduction and early stopping. To compute a more accurate final validation and testing score, the best model was deployed on whole images divided into patches and average voting was used to compute the entire image classification score. Training on both BreaKHis and BACH took approximately 25 min.

Testing and re-training on degraded datasets

Baseline models were trained for BACH, PCam and BreaKHis on the original image sets $\{I_H\}_{\mathrm{Train}}$ and $\{I_H\}_{\mathrm{Validation}}$ to establish a baseline accuracy. Each of the four magnifications in BreaKHis were treated as separate datasets. These models were then deployed on degraded test sets $\{I_L\}_{\mathrm{Test}}$ for each dataset at steadily decreasing NA_L to identify the relevant feature sizes learned by the baseline model for accurate classification. The model was then retrained on the degraded image sets $\{I_L\}_{\mathrm{Train}}$ and $\{I_L\}_{\mathrm{Validation}}$ to determine if lower resolution features could be learned while maintaining acceptable accuracy on the degraded test set $\{I_L\}_{\mathrm{Test}}$ for a given value of NA_L. For BACH and BreaKHis, where NA_L was decreased below the baseline NA_H value for PCam (0.13), the model of PCam was loaded that had been trained on similarly degraded images.

4. Results

Figs. 4 to 7 show training, validation and test results for original and degraded versions of each dataset, and Table 3 summarizes the relevant degradation thresholds in terms of degraded system NA, or NA_L. For consistency, we calculated the absolute accuracy using a benign/malignant label threshold of 0.5, and validation accuracy influenced the choice of the best model during training. Benchmark scores provided in the literature for these and other image classification datasets generally include accuracy (percentage of correctly classified images), area under the receiver operating characteristic curve (AUC, or diagnostic ability with variable threshold), precision/specificity (true negative rate), recall/sensitivity (true positive rate) and F_β (a weighted average of precision and recall). F_1 scores are most commonly presented, which equally weight the precision and recall into a single score. In the clinical context of identifying healthy samples with high confidence and without human assessment, the typical combination of binary accuracy/AUC and equally weighted F_1 score should be supplemented with the area under the precision recall curve (AUPRC). The specificity itself is particularly important for the type of application

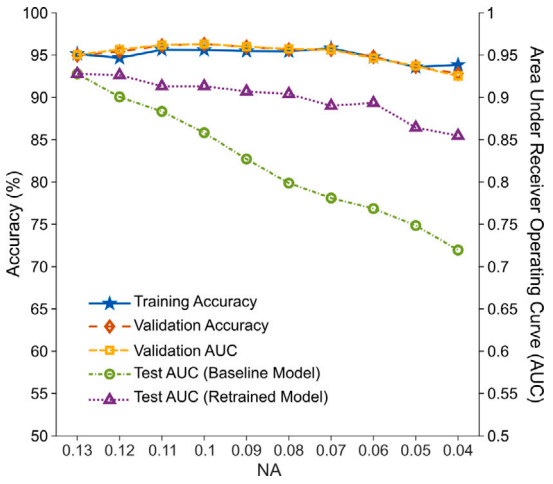

Fig. 4. Training and validation accuracy as well as validation and test AUC of automated classifier when presented with degraded images of the PCam dataset. Datapoints at NA = 0.13 correspond to original images. Solid blue line with pentagram markers shows training accuracy, while dashed lines show validation accuracy (diamond markers) and validation AUC (square markers) illustrating that accuracy and AUC are very similar. Dash-dotted green line with circle markers shows test AUC using original baseline model, and dotted purple line with triangle markers shows test AUC with models re-trained on equivalent degraded dataset.

considered here as false negatives would be much more serious, while false positives could still be rectified in the normal assessment pathway. The AUPRC is also more descriptive for unbalanced datasets (Davis and Goadrich, 2006), as is the case with both BreaKHis and PCam.

Full test labels were available for BreaKHis and thus training, validation and test accuracy are presented for baseline and degraded models in addition to AUPRC and specificity. Test set validation for PCam was obtained using the submission scoring system on Kaggle, and the area under the receiver operating characteristic curve (AUC) was provided as the evaluation metric. Validation accuracy (with a benign/malignant label threshold of 0.5) and validation AUC were very similar, thus test AUC was considered to be equivalent to test accuracy. We note that we did not have access to the test labels for the BACH dataset, as the challenge hosted at https://iciar2018-challenge.grand-challenge.org was no longer accepting legacy submissions for scoring at the time of this work and the organizers of the challenge did not respond to our requests for the test labels. We instead used our baseline model predictions as pseudo-ground truth test labels and measured the accuracy, AUPRC and specificity of the degraded test sets relative to these pseudo-labels, again using a benign/malignant label threshold of 0.5. The test set was balanced evenly between classes, and the baseline predicted 52 images out of 100 as malignant, thus our predicted ground-truth labels were similarly balanced between classes. Finally, the most routinely misclassified images for BreaKHis and BACH are identified, along with some additional statistics for misclassifications in general.

4.1. PCam

Fig. 4 shows the results for the PCam dataset. We achieved a baseline training accuracy of 95.5% and validation accuracy/AUC of 94.3%/0.95 for PCam, with a baseline test AUC of 0.93 on the original images. Test AUC steadily decreases for increasingly degraded images, with the previously defined threshold for unacceptable loss in accuracy (10%, or AUC = 0.83) occurring at an equivalent NA_L of 0.09. For models trained on degraded image sets, images can be degraded to NA_L = 0.04 while maintaining a test AUC of 0.85.

4.2. BreaKHis

Figs. 5 and 6 show the results for each of the magnifications provided in the BreaKHis dataset. The baseline training, validation and test accuracies achieved for BreaKHis 4× were 98.5%, 99.3% and 99.3%, respectively. Test accuracy for the baseline 4× model reached the threshold of accuracy loss (91.5%) at NA_L = 0.11 from the original NA_H of 0.16. For models trained on degraded image sets, images can be degraded to NA_L = 0.04 while still achieving a test accuracy of 97.0%. The baseline train/validation/test accuracies for 10× were 97.5%, 92.8% and 94.3%, respectively. Test accuracy for the baseline 10× model reached the threshold of accuracy loss (87.3%) at a much higher degradation of NA_L = 0.25 from the original NA_H of 0.40 than the baseline 4× model. Retrained models maintained acceptable accuracy for degradation at NH_L = 0.05 of 96.2%. Baseline train/validation/test accuracies for 20× were 99.3%, 98.0% and 98.5%, respectively, and test accuracy for the original model maintained virtually unchanged diagnostic accuracy to NA_L = 0.35, reaching the threshold (88.9%) at NA_L = 0.25, which is similar to the baseline 10× model. Retrained 20× models maintained performance over an even larger degradation range (NA_L = 0.10), only dropping to 92.0% at full degradation of NA_L = 0.05. Finally, baseline train/validation/test accuracies for BreaKHis 40× were 97.4%, 97.1% and 94.5%, respectively. The original model maintained acceptable diagnostic accuracy across the entire spectrum of degradation from NA_H = 1.40 to NA_L = 0.05, dropping to 84.5%. Re-trained models were very similar to the baseline, achieving 89.5% accuracy at NA_L = 0.05.

Using an equivalent reduction in AUPRC of 0.1 from the baseline value as a cutoff for unacceptable diagnostic ability, the baseline model trained on original images from the 4× dataset (baseline of 1.0) reaches the loss threshold at a resolution of NA_L = 0.06 (AUPRC = 0.898), while the retrained model maintained an AUPRC of at least 0.985 across the entire degradation range. The baseline model trained on the 10× dataset achieved a baseline AUPRC of 0.996, and never dropped below 0.933, even for the test set degraded to NA_L = 0.05 which was the worst performing model in most cases. The model retrained on NA_L = 0.05 achieved the worst performance of 0.933, still within the acceptable range, and slightly worse than the baseline model. For 20× images, a baseline AUPRC of 0.991 was reached, dropping to 0.897 for the NA_L = 0.05 degraded test set but still acceptable according to our definition. Retraining produced more variable performance than was expected, and in some cases did not improve relative to the baseline model; unusually, the retrained model at NA_L = 0.05 did not meet the acceptability criteria at all (0.813) while the baseline model was just within the cutoff (0.897). For the final set of images at 40×, training on the original images yielded a baseline AUPRC of 0.980, and the performance across the degraded test sets remained within the acceptable range in all cases (minimum 0.887 for baseline at NA_L = 0.05), and as with 20×, retrained performance was variable, with particularly poor performance at NA_L = 0.20 (0.905) but still acceptable. The specificity results appear to indicate that re-training is critical to maintain acceptable performance in all cases, however we note that a 10% reduction in specificity would almost certainly be unacceptable in clinical practice, and these values are calculated strictly for a 0.5 binary/malignant label threshold which is arbitrarily chosen. In reality, false negative errors would be more heavily penalized at all stages, *i.e.* a lower threshold would be used to flag potential malignancy and models would be optimized for specificity rather than accuracy.

In terms of misclassification statistics, all four datasets contained 1–2 images that were misclassified by the majority of trained models. At 4× magnification, 84 of the 199 test images were misclassified at least once. Two images of ductal carcinoma from the same slide (14-3909) were misclassified as benign by all models except the baseline when predicting on the original resolution version of the image. The next three images that were misclassified by the majority of models were also of ductal carcinoma and from the same slide (14-11031).

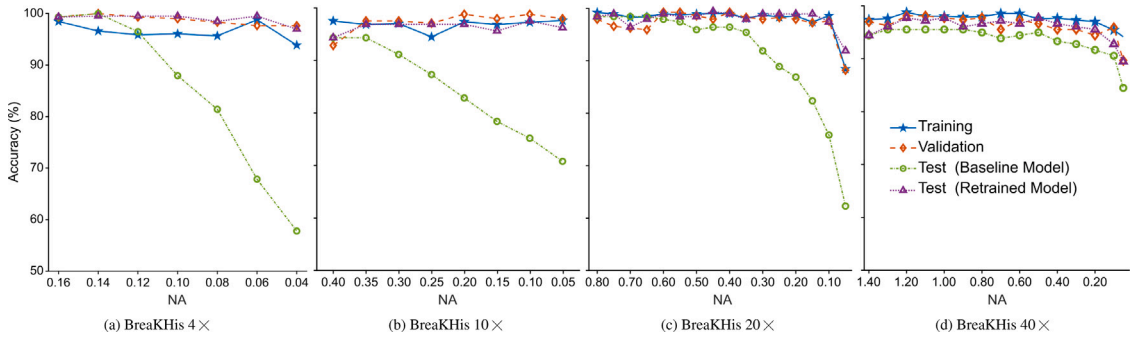

Fig. 5. Training, validation and test accuracy of automated classifier when presented with degraded images for each magnification present in the BreaKHis dataset. Datapoints at (a) NA = 0.16, (b) NA = 0.40, (c) NA = 0.80, and (d) NA = 1.40 correspond to original images. Solid blue line with pentagram markers shows training accuracy, while dashed orange line with diamond marker shows validation accuracy. Dash-dot green line with circle markers shows test accuracy using original baseline model and dotted purple line with triangle markers shows test accuracy with models re-trained on equivalent degraded training dataset.

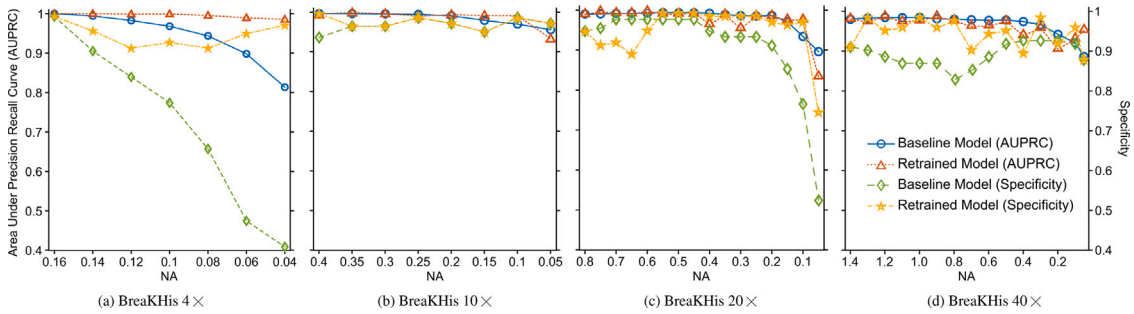

Fig. 6. Area under precision recall curve (AUPRC) and specificity of automated classifier when presented with degraded test images for each magnification present in the BreaKHis dataset. Datapoints at (a) NA = 0.16, (b) NA = 0.40, (c) NA = 0.80, and (d) NA = 1.40 correspond to original images. Solid blue line with circle markers shows test AUPRC and green dashed line with diamond markers shows test specificity for the original baseline model, while dotted red line with triangle markers shows test AUPRC and yellow dash-dot line with pentagram markers shows test specificity for models retrained on equivalent degraded training dataset.

Two images of benign fibroadenoma were misclassified in half of the cases (7/14) and also came from the same slide (14-25197). At 10× magnification, 62 out of 157 test images were labelled incorrectly in at least one case. A slide containing mucinous carcinoma (14-10147) contributed two images that were misclassified in all cases by all models, and a third image from this slide was misclassified in the majority of cases, the exception being the models retrained on NA_L = 0.25 and lower. A single image of ductal carcinoma proved difficult to identify in the majority of cases, as well as two more images of mucinous carcinoma from different slides. Only two benign images (one of fibroadenoma and one of phyllodes tumour) were misclassified as malignant in multiple cases (3/15), both by models retrained on lower resolution ($NA_L \leq 0.25$). At 20× magnification, 134 of 199 total test images were wrongly identified in one or more cases. A single benign image of adenosis tissue from slide 14-22549CD was misclassified in all cases except for the baseline model predicting on the image degraded to NA_L = 0.75. Two additional images from this slide were misclassified but only in single cases. Two images of benign tubular adenoma from slide 14-19854C were incorrectly classified as malignant by all models predicting on the image when it was degraded to NA_L = 0.4 and below. A single image of benign fibroadenoma was also misclassified by most models, but notably not for the baseline model when predicting on the image when it was degraded to below NA_L = 0.5. Only two malignant images of mucinous and ductal carcinoma were incorrectly predicted as benign in half of the cases or more. At 40×, 89 of 181 test images were mislabelled at least once. An image of tissue containing mucinous carcinoma was misclassified as benign by all models except one (retrained on NA_L = 0.10), while a slide (14-25197) containing benign fibroadenoma contributed three images that were misclassified in multiple instances, with one image being incorrectly labelled by all models except the model retrained on NA_L = 0.80. The next three images that were mislabelled in at least 16 of 29 cases were one of papillar and two of mucinous carcinoma.

4.3. BACH

Training the patch-based classifier on the BACH dataset was particularly difficult due to its small size (400 training and 100 test images). A good baseline was achieved on the original images in line with the state of the art with a patch-based training accuracy of 98.1%, and a full-image validation accuracy of 92.5%, however the results were quite variable for re-training on the degraded images. As a consequence, Fig. 7 shows only the test discrepancy for each of the degraded test sets. The model maintains acceptable relative accuracy for degradation at NA_L = 0.12 of 90%, or ten images classified differently to the original set of predictions. Test discrepancy significantly increases for NA_L of 0.11 and lower, which is largely consistent with the results for PCam and BreaKHis baseline models. The AUPRC metric implies reasonably strong performance across the range of degradation, dropping to 0.949 at NA_L = 0.05, with the caveat that all metrics were calculated with pseudo-ground truth labels and only show that the model did not generate significantly different labels for degraded images relative to the baseline images. Specificity however quickly drops below any sort of acceptable threshold at NA_L = 0.15. 40 of the 100 test images were labelled differently than the baseline in at least one instance, and 8 images were identified differently in at least 5 out of 9 cases (NA_L = 0.15 and below, and all 8 were misidentified below NA_L = 0.11).

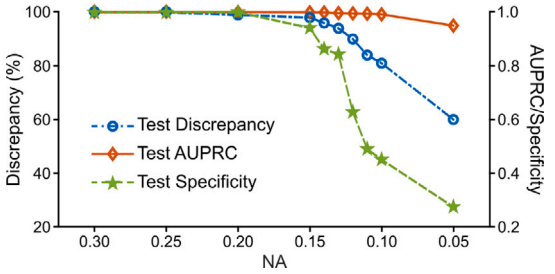

Fig. 7. Degradation of BACH dataset. As test labels were not available, predicted labels generated by original baseline model when presented with each degraded test set were compared to predicted labels generated from original dataset.

Table 3
Baseline test accuracies and threshold NA_L values that maintain acceptable binary diagnostic accuracy (≤10% relative to baseline) for original and retrained models. BH refers to BreaKHis. †Baseline test accuracy of BACH is 100% due to lack of labels for test dataset, thus predicted labels using the baseline model were taken as ground truth. Test accuracy for degraded datasets was calculated relative to these labels. ‡Test AUC is provided for PCam instead of accuracy as this was the metric provided by automated scoring on Kaggle.

Dataset	NA_H Baseline	Acc (%)	NA_L Original	Acc (%)	NA_L Retrained	Acc (%)
BACH	0.30	100†	0.12	90	–	–
BH4×	0.16	99.3	0.11	91.5	0.04	97.0
BH10×	0.40	94.3	0.25	87.3	0.05	96.2
BH20×	0.80	98.5	0.25	88.9	0.05	92.0
BH40×	1.40	94.5	0.05	84.5	0.05	89.5
PCam	0.13	0.927‡	0.09	0.827‡	0.04	0.855‡

5. Discussion

The results outlined in the previous section demonstrate that reasonable binary diagnostic accuracy can be maintained despite a significant reduction in the spatial resolution of breast tissue images relative to the original system resolution, but additional optimization would be required before such a system could be realistically deployed in a clinical setting. The best performance under the constraint of reduced spatial resolution was achieved with the BreaKHis 4× dataset, which is consistent with concept of the space-bandwidth product (SBWP), a parameter that describes the fixed relationship between optical resolution and FOV and quantifies the maximum information that can be encoded by an optical system in a single image (Lohmann et al., 1996). While more high-frequency information is gained by increasing the resolution, fewer macroscopic features of the object are visible in the reduced FOV.

It should be noted that due to the use of a 3.3× relay lens in the imaging system used to generate the BreaKHis images, the 4× dataset has a system magnification similar to PCam (10×). Despite the similarity in magnification, the models trained on PCam were by all metrics the worst performing, which is almost certainly due to the small size of the image patches and resulting lack of macroscopic information visible in the FOV, further illustrating the constraint of the SBWP. Its deployment on other datasets as part of a patch-based classifier however yielded good results, and the model was clearly still able to learn some relevant macroscopic structures.

The BreaKHis 4× images were also oversampled, while the PCam images were not; the WSIs in the original dataset that PCam was extracted from (CAMELYON16) were undersampled even at the monochrome sampling frequency defined in Eq. (12) and provided in Table 2, and the undersampling in reality due to the use of a colour camera was likely even more severe. It is possible that the effects of information loss due to aliasing artefacts obscured some of the macroscopic structure visible in the PCam images, affecting the classification performance beyond the FOV limitation. It is therefore reasonable to conclude that while high spatial resolution is not strictly necessary, sufficient digital sampling likely is.

In several cases, a minor decrease of 0.01 or 0.02 relative to the baseline NA_H value resulted in slightly improved training, validation and test accuracies/AUPRC values; the degradation function defined in Eq. (15) acted as a low-pass filter, removing high-frequency noise as well as compression artefacts that were visually present in some images, particularly BreaKHis as the images were provided in `.PNG` format. The images in the BACH dataset were provided as `.TIF`s and visually appeared to be the highest quality. They were also captured with the largest image sensor of all datasets, but despite the relatively large FOVs, results were inconsistent due to the very small size of the training dataset (400 images). It was difficult to confidently validate each model due to the lack of ground-truth labels for the test set. Overfitting was also an issue, with the model often reaching very high training accuracy but poor validation accuracy.

It is not clear why the baseline models trained on BreaKHis 10× and 20× were more sensitive to degradation than the baseline model trained on 40×, reaching the threshold of unacceptable accuracy at $NA_L = 0.25$ for both 10× and 20× compared to $NA_L = 0.05$ for 40×. Further investigation would involve identifying the types of structures that are most relevant to the output label of the classifier. It is possible that 4× images contained sufficient macroscopic structural information and 40× images contained more granular features of the cell nuclei and cytoplasm, and 10× and 20× images did not contain sufficient detail in either domain. The 40×-trained classifier may also have been overfit on spurious features such as brightness or stain variations however, as its performance was remarkably consistent across a very large degradation range, and fine nuclear details would not have been resolvable in the most heavily degraded images. The results for the retrained models on all datasets did follow the expected trend of reduced performance for increased magnification given the previously discussed constraint of reduced spatial resolution and consequently reduced information content.

The BACH dataset allowed for a straightforward interrogation of the consistency of each of the individual classifiers as well as the ensemble as a whole due to its small size; the 10% reduction in relative accuracy seen at $NA_L = 0.12$ was largely due to "flipping" of predicted labels for original images I_H images that were assigned a classification score of between 0.4 and 0.6 by the baseline model; for binary classification, labels in this range indicate a "borderline" diagnosis. Each of the ensemble networks were individually more inconsistent when presented with these borderline cases, and the 10% relative threshold was crossed at slightly higher NA values for each. These types of results are typical for ensemble networks in general and demonstrate the reasoning behind choosing this type of architecture for this specific task.

The 10%/0.1 metric reduction threshold was chosen for comparison purposes and would likely not be acceptable in routine clinical practice. The aim of this experiment was to identify if high spatial frequency information is strictly necessary to make an accurate binary classification, and the results suggest that macroscopic tissue structure is sufficient for high-level classification, *i.e.* benign or malignant. It is clear that imaging systems with lower magnification and spatial resolution are well suited for this task, which would require fewer images to fully capture a sample and generate WSIs with much smaller file sizes. Tasks such as segmentation or tissue grading require more granular analysis, and high-resolution scanning would of course still be necessary for these types of tasks.

A feasible application for the type of rapid low-resolution binary diagnostic system described here would be in the context of a high-volume screening program *e.g.* to identify normal tissue samples with high diagnostic confidence, minimizing (or in an ideal case, eliminating entirely) the number of false negative (FN) errors at the expense of absolute classification accuracy. Such a system would still serve to reduce the volume of samples needing to be scanned and stored at

full resolution for later assessment. One potential modification to the ensemble architecture presented here to achieve this would be to use maximum voting rather than average voting as a pooling layer when computing the whole-image score, amplifying any suspected malignancy and "flagging" the section for more detailed assessment using the standard pipeline. Optimizing for high specificity or an F_β score weighted more towards specificity rather than absolute accuracy during training could also improve the performance of an automated classifier for this specific task. Specific imaging system parameters would need to be carefully chosen to balance the reduction in FN errors while still successfully classifying enough normal samples and removing them from the pipeline to be useful. Optimizing the entire system (hardware and software) for specificity rather than absolute accuracy could also increase the system's flexibility, for example in extending its use for other tissue types than have been investigated here. System parameters would likely have to be carefully tailored to the type of tissue being assessed however, which could be a limitation in its wider applicability; it would likely find use only in particularly high-volume screening programs such as breast or colon tissue.

The threshold values listed in Table 3 can be assigned to equivalent microscope objectives: an NA of 0.04 corresponds to an Olympus apochromat 1.25× objective for macroscopic observation, and an NA of 0.10 to an Olympus achromat 4× objective. A 1.25× objective has a monochrome Nyquist sampling frequency corresponding to a camera pixel size of 3.7 μm using Eqs. (10) and (11), however in practice these images would likely still be undersampled if a CFA camera sensor was used, as discussed in Section 2 and demonstrated by the performance of the PCam-trained models. The spatial resolution of a 1.25× lens appears to be the limit beyond which binary classification is no longer possible, and any information loss due to undersampling would likely be prohibitive.

It is possible that future technological development of camera sensors with higher quantum efficiencies could allow pixel sizes to be reduced below the current limit of approximately 2.5 μm, in which case the use of such macroscopic observation lenses would allow for extremely rapid binary classification. A relay lens with less than unity magnification could also be used to optimize between spatial resolution and macroscopic structural information. Despite this current technical limitation, the use of a 4× achromat objective with an NA of 0.1 would still result in a significant increase in FOV compared with current WSI systems while leaving room for additional resolution loss due to undersampling. Imaging at 4× magnification would require a factor of 25 fewer images to fully capture a sample compared with 20× as is standard with most slide scanners, with a commensurate reduction in scanning time and WSI file size.

Another significant issue with WSI systems is ensuring the sample remains in focus. An advantage of using lower NA lenses is increased axial depth of field, which is inversely dependent on the square of the NA (Oldenbourg and Shrikak, 2009). The impact of defocus can also be modelled as a wavefront error in the pupil function P, which has the effect of widening the PSF and reducing the spatial frequency support of the MTF in a similar but not identical manner to a reduction in NA. The results presented here suggest that binary classification algorithms may be able to withstand fairly significant defocus, as well as other pupil aberrations such as spherical aberration. Preliminary work suggests that this is the case; see Supplementary Information for further detail. Another type of aberration common to lower-magnification microscope objectives are field-dependent aberrations such as field curvature and distortion, where the pupil wavefront error varies across the FOV. A future experiment could introduce variable degradation of the type presented here during training as a data augmentation strategy, which would force a classifier to learn distorted or blurred features as they may occur at the edges of the FOV and could increase robustness of patch-based WSI classifiers.

6. Conclusions and future work

The key result of this work is that robust automated binary classification is possible at low spatial resolution, which is consistent with expert pathologists' abilities to quickly make a high-level tissue assessment at low magnification. High-resolution images are costly to store and cumbersome to work with, both for human and automated assessment. Imaging at lower magnification results in shorter scanning times and reduced susceptibility to focus errors due to a larger depth of field. Given the current challenges facing histopathology departments, the introduction of a rapid, low-cost system capable of accurately identifying healthy or normal tissue without the need for human intervention could function as a pathological triage system to mitigate currently unmanageable caseloads. Even partially reducing caseloads would provide pathologists with additional time to devote to research and contribute their considerable expertise to the development of automated diagnostic tools. Reducing the number of samples needing to be scanned at full resolution (such as for segmentation or tissue grading) would also lower the operating costs of a digital pathology department in terms of data storage and technical support staff, which has been identified as a barrier to digitization. In terms of real-world applications, the type of system proposed in this work would most likely find use as an additional component of the digitization workflow rather than as a replacement for the high-quality scanners that many departments already have access to.

Future work would involve identifying specific low-resolution features that are most relevant for classification and correlating these with known anatomical and pathological structures. An investigation of the effects of undersampling and wavelength-dependent sampling resolution for different types of RGB cameras would also yield insight into the importance of these system parameters to accurate binary classification. The limited size of available datasets in digital pathology as well as variability in stain appearance often leads to overfitting and brittle diagnostic models; there is evidence that stain normalization and augmentation can be powerful tools for addressing these issues (Janowczyk et al., 2017; Salvi et al., 2020). An expansion of this work would investigate which of these strategies are relevant for accurate binary classification of low-resolution images of stained tissue. The techniques presented here for systematic image degradation could also be used as a real-time data augmentation strategy to improve the robustness of patch-based classifiers to regions of the tissue that are out of focus or distorted. As with any deep learning model, including additional datasets as they become available would of course improve the performance; the BRACS dataset (Brancati et al., 2022) in particular would be a major focus of future work, but would require adapting the training architecture to allow for input images of variable size. Finally, the most important extension of the work described here would be to determine if the results obtained for images of breast tissue are repeated with other tissue types that are generated through high-volume screening programs *e.g.* colon tissue (Sirinukunwattana et al., 2017; Kather et al., 2016; Graham et al., 2019) or more general datasets such as the Atlas of Digital Pathology (Hosseini et al., 2019).

CRediT authorship contribution statement

Lydia Neary-Zajiczek: Writing – Original draft, Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation. **Linas Beresna:** Software, Investigation. **Benjamin Razavi:** Software, Formal analysis, Investigation. **Vijay Pawar:** Supervision. **Michael Shaw:** Writing – review & editing, Supervision. **Danail Stoyanov:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The datasets used in this work are available at:

https://iciar2018-challenge.grand-challenge.org/ (BACH)
https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/ (BreaKHis)
https://www.kaggle.com/c/histopathologic-cancer-detection (PCam).

Acknowledgements

We thank Anita Rau and Sophia Bano for helpful discussions and guidance on network training. This research was funded in whole by the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) [203145/Z/16/Z]; the Engineering and Physical Sciences Research Council (EPSRC) [EP/P027938/1, EP/R004080/1, EP/P01284 1/1]; and the Royal Academy of Engineering Chair in Emerging Technologies Scheme. For the purpose of open access, the authors have applied a CC BY public copyright licence to any author accepted manuscript version arising from this submission.

Code availability

The MATLAB code for dataset NA estimation and image degradation is available at https://github.com/lydiazajiczek/Image-Degradation. The Python code for automated classification is available at https://github.com/lydiazajiczek/Image_Resolution_Patch_Ensemble. Both repositories are licensed under an MIT open source license.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2023.102891.

References

Cancer Research U.K., 2016. Testing times to come? An evaluation of pathology capacity across the UK. Technical Report, Cancer Research UK, London, pp. 1–60, URL https://www.cancerresearchuk.org/sites/default/files/testing_times_to_come_nov_16_cruk.pdf.

Williams, B.J., Bottoms, D., Treanor, D., 2017. Future-proofing pathology: the case for clinical adoption of digital pathology. J. Clin. Pathol. 70 (12), 1010–1018. http://dx.doi.org/10.1136/jclinpath-2017-204644, URL http://www.ncbi.nlm.nih.gov/pubmed/28780514.

The Royal College of Pathologists, 2018. Meeting pathology demand: Histopathology workforce census. Technical Report, The Royal College of Pathologists, London, UK, URL https://www.rcpath.org/uploads/assets/uploaded/aff26c51-8b62-463f-98625b1d3f6174b6.pdf.

Retamero, J.A., Aneiros-Fernandez, J., del Moral, R.G., 2020. Complete digital pathology for routine histopathology diagnosis in a multicenter hospital network. Arch. Pathol. Lab. Med. 144 (2), 221–228. http://dx.doi.org/10.5858/arpa.2018-0541-OA.

Metter, D.M., Colgan, T.J., Leung, S.T., Timmons, C.F., Park, J.Y., 2019. Trends in the US and Canadian pathologist workforces from 2007 to 2017. JAMA Netw. Open 2 (5), e194337. http://dx.doi.org/10.1001/JAMANETWORKOPEN.2019.4337, URL https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2734800.

Wilson, M.L., Fleming, K.A., Kuti, M.A., Looi, L.M., Lago, N., Ru, K., 2018. Access to pathology and laboratory medicine services: a crucial gap. Lancet 391 (10133), 1927–1938. http://dx.doi.org/10.1016/S0140-6736(18)30458-6.

Mudenda, V., Malyangu, E., Sayed, S., Fleming, K., 2020. Addressing the shortage of pathologists in Africa: Creation of a MMed Programme in Pathology in Zambia. Afr. J. Lab. Med. 9 (1), 1–7. http://dx.doi.org/10.4102/AJLM.V9I1.974, URL http://www.scielo.org.za/scielo.php?script=sci_arttext&pid=S2225-20102020000100004&lng=en&nrm=iso&tlng=es.

Pare, G., Meyer, J., Trudel, M.-C., Tetu, B., 2016. Impacts of a large decentralized telepathology network in Canada. Telemed. e-Health 22 (3), 246–250. http://dx.doi.org/10.1089/tmj.2015.0083, URL https://www.liebertpub.com/doi/10.1089/tmj.2015.0083.

Hamilton, P.W., Wang, Y., McCullough, S.J., 2012. Virtual microscopy and digital pathology in training and education. APMIS 120 (4), 305–315. http://dx.doi.org/10.1111/j.1600-0463.2011.02869.x, URL http://doi.wiley.com/10.1111/j.1600-0463.2011.02869.x.

Chang, H.Y., Jung, C.K., Woo, J.I., Lee, S., Cho, J., Kim, S.W., Kwak, T.-Y., 2019. Artificial intelligence in pathology. J. Pathol. Transl. Med. 53 (1), 1–12. http://dx.doi.org/10.4132/jptm.2018.12.16, http://www.ncbi.nlm.nih.gov/pubmed/30599506, http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6344799.

Williams, B.J., Lee, J., Oien, K.A., Treanor, D., 2018. Digital pathology access and usage in the UK: results from a national survey on behalf of the National Cancer Research Institute's CM-Path initiative. J. Clin. Pathol. 71 (5), 463–466. http://dx.doi.org/10.1136/jclinpath-2017-204808, http://www.ncbi.nlm.nih.gov/pubmed/29317516, http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5916098.

Barham, C., Madden, A.P., 2007. The national programme for information technology in the NHS. Anaesth. Intensive Care Med. 8 (12), 504–505. http://dx.doi.org/10.1016/j.mpaic.2007.09.009.

Griffin, J., Treanor, D., 2017. Digital pathology in clinical use: where are we now and what is holding us back? Histopathology 70 (1), 134–145. http://dx.doi.org/10.1111/his.12993.

Flotte, T.J., Bell, D.A., 2018. Anatomical pathology is at a crossroads. Pathology 50 (4), 373–374. http://dx.doi.org/10.1016/j.pathol.2018.01.003, URL http://www.ncbi.nlm.nih.gov/pubmed/29665965.

Liu, Y., Pantanowitz, L., 2019. Digital pathology: Review of current opportunities and challenges for oral pathologists. J. Oral Pathol. Med. 48 (4), 263–269. http://dx.doi.org/10.1111/jop.12825, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jop.12825.

Zarella, M.D., Bowman, D., Aeffner, F., Farahani, N., Xthona, A., Absar, S.F., Parwani, A.V., Bui, M.M., Hartman, D.J., 2019. A practical guide to whole slide imaging: A white paper from the digital pathology association. Arch. Pathol. Lab. Med. 143 (2), 222–234. http://dx.doi.org/10.5858/arpa.2018-0343-RA, URL http://www.archivesofpathology.org/doi/10.5858/arpa.2018-0343-RA.

Krupinski, E.A., LeSueur, B., Ellsworth, L., Levine, N., Hansen, R., Silvis, N., Sarantopoulos, P., Hite, P., Wurzel, J., Weinstein, R.S., Lopez, A.M., 1999. Diagnostic accuracy and image quality using a digital camera for teledermatology. Telemedicine J. 5 (3), 257–263. http://dx.doi.org/10.1089/107830299312005, URL https://www.liebertpub.com/doi/10.1089/107830299312005.

Williams, B.H., Hong, I.S., Mullick, F.G., Butler, D.R., Herring, R.F., O'Leary, T.J., 2003. Image quality issues in a static image-based telepathology consultation practice. Human Pathol. 34 (12), 1228–1234. http://dx.doi.org/10.1016/S0046-8177(03)00429-5, URL https://www.sciencedirect.com/science/article/pii/S0046817703004295.

Krupinski, E.A., Johnson, J.P., Jaw, S., Graham, A.R., Weinstein, R.S., 2012a. Compressing pathology whole-slide images using a human and model observer evaluation. J. Pathol. Inform. 3 (17), 1–14. http://dx.doi.org/10.4103/2153-3539.95129, http://www.ncbi.nlm.nih.gov/pubmed/22616029, http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3352607.

Krupinski, E.A., Silverstein, L.D., Hashmi, S.F., Graham, A.R., Weinstein, R.S., Roehrig, H., 2012b. Observer performance using virtual pathology slides: impact of LCD color reproduction accuracy. J. Digit. Imaging 25 (6), 738–743. http://dx.doi.org/10.1007/s10278-012-9479-1, URL http://link.springer.com/10.1007/s10278-012-9479-1.

Shrestha, P., Kneepkens, R., van Elswijk, G., Vrijnsen, J., Ion, R., Verhagen, D., Abels, E., Vossen, D., Bas Hulsken, A., 2015. Objective and subjective assessment of digital pathology image quality. AIMS Med. Sci. 2 (1), 65–78. http://dx.doi.org/10.3934/medsci.2015.1.65, URL http://www.aimspress.com/article/10.3934/medsci.2015.1.65.

Shrestha, P., Kneepkens, R., Vrijnsen, J., Vossen, D., Abels, E., Hulsken, B., 2016. A quantitative approach to evaluate image quality of whole slide imaging scanners. J. Pathol. Inform. 7, 56. http://dx.doi.org/10.4103/2153-3539.197205, URL http://www.ncbi.nlm.nih.gov/pubmed/28197359.

Mukhopadhyay, S., Feldman, M.D., Abels, E., Ashfaq, R., Beltaifa, S., Cacciabeve, N.G., Cathro, H.P., Cheng, L., Cooper, K., Dickey, G.E., Gill, R.M., Heaton, R.P., Kerstens, R., Lindberg, G.M., Malhotra, R.K., Mandell, J.W., Manlucu, E.D., Mills, A.M., Mills, S.E., Moskaluk, C.A., Nelis, M., Patil, D.T., Przybycin, C.G., Reynolds, J.P., Rubin, B.P., Saboorian, M.H., Salicru, M., Samols, M.A., Sturgis, C.D., Turner, K.O., Wick, M.R., Yoon, J.Y., Zhao, P., Taylor, C.R., 2018. Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: a multicenter blinded randomized noninferiority study of 1992 cases (pivotal study). Am. J. Surg. Pathol. 42 (1), 39–52. http://dx.doi.org/10.1097/PAS.0000000000000948, URL http://www.ncbi.nlm.nih.gov/pubmed/28961557.

Goacher, E., Randell, R., Williams, B.J., Treanor, D., 2017. The diagnostic concordance of whole slide imaging and light microscopy: a systematic review. Arch. Pathol. Lab. Med. 141 (1), 151–161. http://dx.doi.org/10.5858/arpa.2016-0025-RA, URL http://www.archivesofpathology.org/doi/10.5858/arpa.2016-0025-RA.

Dodge, S., Karam, L.J., 2016. Understanding how image quality affects deep neural networks. In: 2016 8th International Conference on Quality of Multimedia Experience, QoMEX 2016. IEEE, pp. 1–6. http://dx.doi.org/10.1109/QoMEX.2016.7498955, URL.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. 115 (3), 211–252. http://dx.doi.org/10.1007/S11263-015-0816-Y/FIGURES/16, arXiv:1409.0575, URL https://link.springer.com/article/10.1007/s11263-015-0816-y.

Gilbertson, J.R., Ho, J., Anthony, L., Jukic, D.M., Yagi, Y., Parwani, A.V., 2006. Primary histologic diagnosis using automated whole slide imaging: a validation study. BMC Clin. Pathol. 6 (1), 4. http://dx.doi.org/10.1186/1472-6890-6-4, URL http://bmcclinpathol.biomedcentral.com/articles/10.1186/1472-6890-6-4.

Snead, D.R.J., Tsang, Y.-W., Meskiri, A., Kimani, P.K., Crossman, R., Rajpoot, N.M., Blessing, E., Chen, K., Gopalakrishnan, K., Matthews, P., Momtahan, N., Read-Jones, S., Sah, S., Simmons, E., Sinha, B., Suortamo, S., Yeo, Y., El Daly, H., Cree, I.A., 2016. Validation of digital pathology imaging for primary histopathological diagnosis. Histopathology 68 (7), 1063–1072. http://dx.doi.org/10.1111/his.12879.

Araújo, A.L.D., Arboleda, L.P.A., Palmier, N.R., Fonsêca, J.M., de Pauli Paglioni, M., Gomes-Silva, W., Ribeiro, A.C.P., Brandão, T.B., Simonato, L.E., Speight, P.M., Fonseca, F.P., Lopes, M.A., de Almeida, O.P., Vargas, P.A., Madrid Troconis, C.C., Santos-Silva, A.R., 2019. The performance of digital microscopy for primary diagnosis in human pathology: a systematic review. Virchows Archiv 474 (3), 269–287. http://dx.doi.org/10.1007/s00428-018-02519-z, URL http://link.springer.com/10.1007/s00428-018-02519-z.

Goodman, J.W., 2005. Introduction to Fourier Optics. Roberts & Co, p. 491.

Airy, G.B., 1835. On the diffraction of an object-glass with circular aperture. Trans. Camb. Philos. Soc. 5, 283.

Bracewell, R.N., 1999. The Fourier Transform and Its Applications, third ed. McGraw Hill, pp. 108–112.

Born, M., Wolf, E., 1999. Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light, 7th ed. Cambridge University Press, p. 952.

Marks, R.J., 2009. Handbook of Fourier Analysis & Its Applications. Oxford University Press, New York, NY, USA, p. 772.

Oppenheim, A.V., Schafer, R.W., Buck, J.R., 1999. Discrete-Time Signal Processing, second ed. Prentice Hall, Upper Saddle River, NJ, p. 870.

Swirski, L., 2009. CFA interpolation detection. In: Topics in Security: Forensic Signal Analysis. University of Cambridge, Cambridge, UK.

Palum, R., 2001. Image sampling with the Bayer color filter array. In: Image Processing, Image Quality, Image Capture Systems Conference. The Society for Imaging Science and Technology, Montreal, QC, Canada, pp. 239–245, URL https://www.imaging.org/site/PDFS/Papers/2001/PICS-0-251/4631.pdf.

Malvar, H.S., He, L.W., Cutler, R., 2004. High-quality linear interpolation for demosaicing of Bayer-patterned color images. In: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing. Vol. 3, pp. iii–485. http://dx.doi.org/10.1109/ICASSP.2004.1326587.

Horowitz, P., Hill, W., 2015. The Art of Electronics, third ed. Cambridge University Press, New York, NY, USA, p. 1125.

Mandel, L., Wolf, E., 1995. Optical Coherence and Quantum Optics, first ed. Cambridge University Press, New York, NY, USA, p. 1194.

Gonzalez, R.C., Woods, R.E., 2007. Digital Image Processing, third ed. Pearson, Upper Saddle River, NJ, USA, p. 976, URL http://dl.acm.org/citation.cfm?id=1076432.

Aresta, G., Araújo, T., Kwok, S., Chennamsetty, S.S., Safwan, M., Alex, V., Marami, B., Prastawa, M., Chan, M., Donovan, M., Fernandez, G., Zeineh, J., Kohl, M., Walz, C., Ludwig, F., Braunewell, S., Baust, M., Vu, Q.D., To, M.N.N., Kim, E., Kwak, J.T., Galal, S., Sanchez-Freire, V., Brancati, N., Frucci, M., Riccio, D., Wang, Y., Sun, L., Ma, K., Fang, J., Kone, I., Boulmane, L., Campilho, A., Eloy, C., Polónia, A., Aguiar, P., 2019. BACH: Grand challenge on breast cancer histology images. Med. Image Anal. 56, 122–139. http://dx.doi.org/10.1016/j.media.2019.05.010, arXiv:1808.04277, URL https://linkinghub.elsevier.com/retrieve/pii/S1361841518307941.

Spanhol, F.A., Oliveira, L.S., Petitjean, C., Heutte, L., 2016. A dataset for breast cancer histopathological image classification. IEEE Trans. Biomed. Eng. 63 (7), 1455–1462. http://dx.doi.org/10.1109/TBME.2015.2496264, http://web.inf.ufpr.br/vri/breast-cancer-database http://ieeexplore.ieee.org/document/7312934/.

Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M., 2018. Rotation equivariant CNNs for digital pathology. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11071 LNCS, Springer Verlag, pp. 210–218. http://dx.doi.org/10.1007/978-3-030-00934-2_24, arXiv:1806.03962.

Ginsburg, O., Yip, C.-H., Brooks, A., Cabanes, A., Caleffi, M., Antonio Dunstan Yataco, J., Gyawali, B., McCormack, V., McLaughlin de Anderson, M., Mehrotra, R., Mohar, A., Murillo, R., Pace, L.E., Paskett, E.D., Romanoff, A., Rositch, A.F., Scheel, J.R., Schneidman, M., Unger-Saldaña, K., Vanderpuye, V., Wu, T.-Y., Yuma, S., Dvaladze, A., Duggan, C., Anderson, B.O., 2020. Breast cancer early detection: A phased approach to implementation. Cancer 126, 2379–2393. http://dx.doi.org/10.1002/cncr.32887.

Araújo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., Polónia, A., Campilho, A., 2017. Classification of breast cancer histology images using Convolutional Neural Networks. In: Sapino, A. (Ed.), PLOS ONE 12 (6), e0177544. http://dx.doi.org/10.1371/journal.pone.0177544, URL https://dx.plos.org/10.1371/journal.pone.0177544.

Ehteshami Bejnordi, B., Veta, M., van Diest, P.J., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J.A.W.M., Hermsen, M., Manson, Q.F., Balkenhol, M., Geessink, O., Stathonikos, N., van Dijk, M.C., Bult, P., Beca, F., Beck, A.H., Wang, D., Khosla, A., Gargeya, R., Irshad, H., Zhong, A., Dou, Q., Li, Q., Chen, H., Lin, H.-J., Heng, P.-A., Haß, C., Bruni, E., Wong, Q., Halici, U., Öner, M.U., Cetin-Atalay, R., Berseth, M., Khvatkov, V., Vylegzhanin, A., Kraus, O., Shaban, M.T., Rajpoot, N.M., Awan, R., Sirinukunwattana, K., Qaiser, T., Tsang, Y.-W., Tellez, D., Annuscheit, J., Hufnagl, P., Valkonen, M., Kartasalo, K., Latonen, L., Ruusuvuori, P., Liimatainen, K., Albarqouni, S., Mungal, B., George, A., Demirci, S., Navab, N., Watanabe, S., Seno, S., Takenaka, Y., Matsuda, H., Ahmady Phoulady, H., Kovalev, V., Kalinovsky, A., Liauchuk, V., Bueno, G., Fernandez-Carrobles, M.M., Serrano, I., Deniz, O., Racoceanu, D., Venâncio, R., 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 318 (22), 2199–2210. http://dx.doi.org/10.1001/jama.2017.14585, URL http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2017.14585.

Excelitas PCO GmbH, pco.edge 5.5 Datasheet, Technical Report, Kelheim, Germany, URL https://www.pco.de/fileadmin/user_upload/pco-product_sheets/DS_PCOEDGE55_V204.pdf.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: From error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612. http://dx.doi.org/10.1109/TIP.2003.819861.

Kassani, S.H., Kassani, P.H., Wesolowski, M.J., Schneider, K.A., Deters, R., 2020. Classification of histopathological biopsy images using ensemble of deep learning networks. In: CASCON 2019 Proceedings - Conference of the Centre for Advanced Studies on Collaborative Research - Proceedings of the 29th Annual International Conference on Computer Science and Software Engineering. Center for Advanced Studies on Collaborative Research, pp. 92–99. http://dx.doi.org/10.1145/3306307.3328180, arXiv:1909.11870.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: International Conference On Learning Representations. San Diego, CA, pp. 1–14, arXiv:1409.1556, URL http://arxiv.org/abs/1409.1556.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. MobileNets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, URL https://arxiv.org/abs/1704.04861v1.

Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Rakha, E.A., Ahmed, M.A., Aleskandarany, M.A., Hodi, Z., S Lee, A.H., Pinder, S.E., Ellis, I.O., Rakha, E.A., Ahmed, M.A., Aleskandarany, M.A., Lee, A.H.S., Pinder, S.E., Ellis, I.O., 2017. Diagnostic concordance of breast pathologists: lessons from the national health service breast screening programme pathology external quality assurance scheme. Histopathology 70 (4), 632–642. http://dx.doi.org/10.1111/HIS.13117, URL https://onlinelibrary.wiley.com/doi/full/10.1111/his.13117.

Davis, J., Goadrich, M., 2006. The relationship between precision-recall and ROC curves. In: ACM International Conference Proceeding Series. Vol. 148, pp. 233–240. http://dx.doi.org/10.1145/1143844.1143874, URL https://dl.acm.org/doi/10.1145/1143844.1143874.

Lohmann, A.W., Dorsch, R.G., Mendlovic, D., Ferreira, C., Zalevsky, Z., 1996. Space–bandwidth product of optical signals and systems. J. Opt. Soc. Amer. A 13 (3), 470. http://dx.doi.org/10.1364/JOSAA.13.000470, URL https://www.osapublishing.org/abstract.cfm?URI=josaa-13-3-470.

Oldenbourg, R., Shrikak, M., 2009. Microscopes. In: Bass, M., DeCusatis, C., Enoch, J., Lakshminarayanan, V., Li, G., Macdonald, C., Mahajan, V., Van Stryland, E. (Eds.), Handbook of Optics, Third Edition Volume I: Geometrical and Physical Optics, Polarized Light, Components and Instruments, third ed. McGraw-Hill, Inc., New York, NY, USA, p. 1248.

Janowczyk, A., Basavanhally, A., Madabhushi, A., 2017. Stain Normalization using Sparse AutoEncoders (StaNoSA): Application to digital pathology. Comput. Med. Imaging Graph. 57, 50–61. http://dx.doi.org/10.1016/J.COMPMEDIMAG.2016.05.003, URL https://www.sciencedirect.com/science/article/pii/S0895611116300404.

Salvi, M., Michielli, N., Molinari, F., 2020. Stain Color Adaptive Normalization (SCAN) algorithm: Separation and standardization of histological stains in digital pathology. Comput. Methods Programs Biomed. 193, 105506. http://dx.doi.org/10.1016/J.CMPB.2020.105506.

Brancati, N., Anniciello, A.M., Pati, P., Riccio, D., Scognamiglio, G., Jaume, G., Pietro, G.D., Bonito, M.D., Foncubierta, A., Botti, G., Gabrani, M., Feroce, F., Frucci, M., 2022. BRACS: a dataset for BReAst carcinoma subtyping in H&E histology images. Database 2022, http://dx.doi.org/10.1093/DATABASE/BAAC093, URL https://academic.oup.com/database/article/doi/10.1093/database/baac093/6762252.

Sirinukunwattana, K., Pluim, J.P., Chen, H., Qi, X., Heng, P.A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., Böhm, A., Ronneberger, O., Cheikh, B.B., Racoceanu, D., Kainz, P., Pfeiffer, M., Urschler, M., Snead, D.R., Rajpoot, N.M., 2017. Gland segmentation in colon histology images: The GlaS challenge contest. Med. Image Anal. 35, 489–502. http://dx.doi.org/10.1016/J.MEDIA.2016.08.008, arXiv:1603.00275.

Kather, J.N., Weis, C.A., Bianconi, F., Melchers, S.M., Schad, L.R., Gaiser, T., Marx, A., Zöllner, F.G., 2016. Multi-class texture analysis in colorectal cancer histology. Sci. Rep. 6 (1), 1–11. http://dx.doi.org/10.1038/srep27988, URL https://www.nature.com/articles/srep27988.

Graham, S., Chen, H., Gamper, J., Dou, Q., Heng, P.-A., Snead, D., Tsang, Y.W., Rajpoot, N., 2019. MILD-Net: Minimal information loss dilated network for gland instance segmentation in colon histology images. Med. Image Anal. 52, 199–211. http://dx.doi.org/10.1016/j.media.2018.12.001, arXiv:1806.01963.

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