PathProfiler: Automated Quality Assessment of Retrospective Histopathology Whole-Slide Image Cohorts by Artificial Intelligence – A Case Study for Prostate Cancer Research

Maryam Haghighat[1,†,⋆], Lisa Browning[2,3,†], Korsuk Sirinukunwattana[1], Stefano Malacrino[4], Nasullah Khalid Alham[1], Richard Colling[2,4], Ying Cui[4], Emad Rakha[5], Freddie C. Hamdy[3,4], Clare Verrill[2,3,4, ‡], and Jens Rittscher[1,3,‡]

[1]Institute of Biomedical Engineering (IBME), Department of Engineering Science, University of Oxford, Oxford, UK
[2]Department of Cellular Pathology, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford, UK
[3]NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, Oxford, Oxfordshire, UK
[4]Nuffield Department of Surgical Sciences, University of Oxford, John Radcliffe Hospital, Oxford, UK
[5]School of Medicine, University of Nottingham, Nottingham, UK
[†, ‡] Denotes equal contribution
[⋆]Corresponding authors: {maryam.haghighat| jens.rittscher}@eng.ox.ac.uk

ABSTRACT

Research using whole slide images (WSIs) of scanned histopathology slides for the development of artificial intelligence (AI) algorithms has increased exponentially over recent years. Glass slides from large retrospective cohorts with patient follow-up data are digitised for the development and validation of AI tools. Such resources, therefore, become very important, with the need to ensure that their quality is of the standard necessary for downstream AI development. However, manual quality control of such large cohorts of WSIs by visual assessment is unfeasible, and whilst quality control AI algorithms exist, these focus on bespoke aspects of image quality, e.g. focus, or use traditional machine-learning methods such as hand-crafted features, which are unable to classify the range of potential image artefacts that should be considered.

In this study, we have trained and validated a multi-task deep neural network to automate the process of quality control of a large retrospective cohort of prostate cases from which glass slides have been scanned several years after production, to determine both the usability of the images for research and the common image artefacts present.

Using a two-layer approach, quality overlays of WSIs were generated from a quality assessment undertaken at patch-level at 5X magnification. From these quality overlays the slide-level quality scores were predicted and then compared to those generated by three specialist urological pathologists, with a Pearson correlation of 0.89 for overall ‘usability’ (at a diagnostic level), and 0.87 and 0.82 for focus and H&E staining quality scores respectively. We subsequently applied our quality assessment pipeline to the TCGA prostate cancer cohort and to a colorectal cancer cohort, for comparison.

Our model, designated as PathProfiler, indicates comparable predicted usability of images from the cohorts assessed (86-90%), and perhaps more significantly is able to predicts WSIs that could benefit from re-scanning or re-staining for quality improvement.

We have shown in this study that AI can be used to automate the process of quality control of large retrospective cohorts to maximise research outputs and conclusions.

1 Introduction

Research using digitised (scanned) histopathology slides (whole slide images; WSIs) for the development of artificial intelligence (AI) algorithms has increased markedly over recent years[1–4], with pathology having been highlighted as being ‘ripe’ for such innovation in the UK Government’s Industrial Life Sciences Strategy[5]. Whilst such algorithms have the potential to support pathologists diagnostically, thus improving safety, efficiency and quality of reporting, perhaps more significant is their promise in the derivation of novel insights into disease biology and behaviour which are unachievable with a human observer. The impact of AI in histopathology is foreseen by many to be revolutionary with anticipated impact across the whole patient journey from the diagnosis of disease, assessment of prognostic and predictive features, and even stratification of patients for optimal clinical management including allocation to treatment within large clinical trials. The door is being opened to the prospect of deeper understanding of morphomolecular correlations, spatially linking morphology with molecular alterations. Such innovations require big datasets and for some diseases, in order to derive novel prognostic insight, there is the requirement for associated clinical outcome data.

Managing and profiling such large image datasets is time consuming and needs to be automated before such studies can be scaled up. Quality-related issues are recognised to be critical to both the development of AI algorithms and thereafter with their clinical utility and transferability between diagnostic laboratories. This is particularly relevant in relation to WSIs of historic glass slides where artefacts are likely to be more commonly encountered, and to cohorts from multiple diagnostic laboratories with variability in pre-analytical tissue processing and slide storage[6].

Large retrospective cohorts are a rare commodity, particularly those with accompanying curated clinical data. These then become hugely valuable in the quest for AI-derived predictive features, whether they be morphological, morphogenetic, or biomarker-related; insights which will be truly transformative. This is particularly highlighted in relation to prostate cancer research where the long natural history of the disease poses challenges with only the intermediate and long-term endpoints of metastasis free survival or overall survival being considered valid outcomes[7]. Prostate cancer research exemplifies therefore the ongoing significance of historic or retrospective cohorts to scientific discovery, despite the ability now for diagnostic laboratories with access to digital pathology (DP) to prospectively collate digitised glass slides for research, including algorithm development.

Historic cohorts are often digitised many years after the glass slides have been produced and archived, as slide scanning technology has evolved and become more readily available, thus introducing the risk of age-related artefacts in relation to the glass slides and subsequently to the quality of the WSI. Indeed it was our own experience of the digitisation of the ProMPT (Prostate Cancer Mechanisms of Progression and Treatment) prostate cancer cohort locally that highlighted to us the impact of age-related factors on WSI quality.

Consideration of quality issues is also relevant in the context of the already publicly available multi-institutional cohorts, one of the most well-known of which is The Cancer Genome Atlas (TCGA); a cohort example with huge potential in terms of facilitating AI development[8]. Such a huge and varied dataset lends itself to the development of AI tools, and examples of successes to date include 'decision support' tools to facilitate tumour classification[9]. However, in spite of these successes, there is a need to remain mindful of the variable nature of the H&E slides present within such cohorts, and therefore also the potential variability in WSI quality. Whilst some datasets may have been curated in terms of a diagnostic 'label' associated with a WSI, curation of WSI for quality assurance purposes is not necessarily guaranteed; often it is unclear what quality assurance process has been conducted, if any.

Quality of WSI can be impacted by artefacts present intrinsically within the tissue, by artefacts introduced during slide preparation (including H&E staining) and digitisation (e.g. scanner/focus issues), and potentially by those introduced as a result of ageing and long-term slide storage, the latter of which are encountered in an unpredictable manner in archived slides. Examples include variability in tissue section thickness, tissue folding, H&E staining, air bubbles and/or dirt under the coverslip. Whilst variation in some features may not be such an issue for WSI of glass slides from a single laboratory, and particularly with quality assurance schemes in place for clinical laboratories, such as through the UK Accreditation Service (UKAS), such variation may become more relevant when collating WSIs from multi-institutional cohorts and particularly with older glass slides.

It is recognised that in order to avoid introducing bias in algorithm development, it is ideal for datasets to be truly representative, encompassing the range of 'data' that would be expected in real-life[10], including both the expected range of (normal and pathological) tissue features and the anticipated variation in tissue and slide preparation between laboratories. If presented with a sufficiently large training dataset, AI algorithms may 'learn' to tolerate up to a level of the artefact that has already been encountered. On the other hand, it is essential to identify sub-optimal or unusable images to ensure the reliability of the results of subsequent analyses performed on the dataset[11, 12]. Deployment of algorithms on images that do not satisfy certain criteria does not guarantee reliable results.

Whilst the quality of the WSI is therefore an essential consideration during the development and deployment of any algorithm[6, 11], currently this is generally reliant upon manual curation of images being considered for analysis, which is time consuming and inefficient, particularly in the context of a lack of pathologist resource. Furthermore, the reliance upon human observers for quality assessment is perhaps questionable given the recognised subjectivity and poor inter-observer concordance in such tasks, even amongst expert pathologists[13, 14]. This issue is further complicated by the interpretation of what is appropriate quality or 'usable' which for the expert pathologist is dependent upon the diagnostic question.

It would seem then that an automated method for the assessment of image quality offers the potential to improve efficiency and consistency in such a task. However, quality assessment in relation to histopathology images is complex. Whilst tools exist for image quality assessment of natural scenes[15], they cannot be directly applied to histological images[16] which pose unique challenges related to the natural complexity of tissue features, the multitude of artefacts that can be encountered, and their distinction from artefacts in natural images. This is complicated further by the need to determine a threshold for acceptability of the image, which is best defined by an expert, and a means to predict this based on the features assessed.

As a result of these challenges, the availability of quality assessment tools is currently limited, with only a small number developed specifically for histopathology slides[17–23]. To date, the available tools employ traditional hand-crafted features rather than learned ones[20,23], or tend to be limited to identification of out-of-focus regions only[17–20] or identification of one artefact per image[21,22]. However, assessment for a combination of artefacts is more meaningful as in real life image artefacts are rarely limited to one feature such as poor staining or tissue folding, particularly with older glass slides. Furthermore, potential quality improvement intervention such as re-scanning or re-staining, are more efficiently recognised if all artefacts present in the image are known. For instance, re-scanning would not resolve the quality issue when there are other artefacts such as dirt or ink over the tissue area. Currently available tools also do not indicate the ‘usability’ of the WSI, which is an important parameter when deciding whether a WSI should be included within a cohort for algorithm development.

To address this gap we have developed PathProfiler; an AI tool to automate the quality assessment of WSIs of a retrospective cohort, using the ProMPT prostate cohort as an example. We demonstrate the reliability of PathProfiler in predicting the presence of multiple artefacts in a WSI together with an indication of their impact on image quality, as assessed at the level of usability for clinical diagnosis and highlight the need for collaborative efforts for further development and utility.

PathProfiler provides an image quality assessment at both patch-level and slide-level. Our pipeline generates whole slide quality overlays and predicts the overall usability of each WSI (a value in the range of 0-1 that can be binarised to 0 or 1), and a score 0-10 for quality of focus and H&E staining from the lowest quality to the highest quality. Through the identification of low-quality images at a patch-level, the uncertainty of developed algorithms can be managed by excluding the low-quality patches from data analysis.

2 Methods

The size and complexity of our prostate cancer cohort which motivated this work is described in Section 2.1. The rationale for developing PathProfiler is discussed in Section 2.2. The application of the tool to our prostate cancer cohort, prostate cancer slides from TCGA, and other histology cohorts, is presented in Section 2.3. We selected HistoQC[23] to benchmark PathProfiler. While other quality tools are available, they typically only assess specific artefacts as for example quality of focus. HistoQC exploits a set of hand-crafted features to train supervised classifiers. The challenges of hand-crafted features for quality assessment of our prostate specimen cohort are discussed in Section 3.3.

2.1 Prostate cancer cohort selection

We utilised two cohorts of H&E slides from formalin-fixed paraffin embedded (FFPE) prostate tissue for algorithm development: a retrospective cohort of historic slides from the ProMPT study, and a ‘contemporary’ cohort from the routine diagnostic workflow. These included biopsies and transurethral resection specimens from the prostate (TURP).

The ProMPT cohort comprises 4732 histology slides (3819 H&Es and the remainder immunohistochemistry) collated between 2001-2018 as part of a UK-based observational study in which participants underwent routine histological assessment as part of diagnostic pathways. All slides digitised for this study were from one site (the lead site - a large academic teaching hospital) which stores slides older than 2 years in an off-site accredited archival storage facility. The combination of the clinical outcome data together with the histological images for this cohort makes it a rare and highly significant resource for image analysis-based research, whilst also providing a relevant example cohort for the purpose of this study.

The ProMPT slides were retrospectively digitised during the period 2017-2021 and are therefore considered representative of a historic glass slide cohort. The collection includes predominantly prostate biopsies, together with transurethral resection specimens from the prostate (TURPs), and radical prostatectomies (RPs). In the initial stages of the digitisation process, it was recognised that there were quality issues with the scanned images of a small proportion of the cases that might potentially have an impact on downstream image analysis; features considered common to other historic cohorts. These included out-of-focus regions, variability in H&E staining quality (fading/loss of contrast), tissue folding, air/glue ‘bubbles’ under the coverslip, dirt, pigments, cover slip edge, and surgical diathermy (such as illustrated in Figure 1A).

For training the model, a random 10% of cases (glass slides from 2008-2016) from the ProMPT cohort were identified from which we selected one H&E WSI (the first available). There were in total 107 WSI from H&E slides, representing 99 biopsies and 8 TURPs.

For quality comparison and enhancing the training dataset with a contemporary cohort, the first available H&E WSI was selected from a set of 91 consecutive prostate biopsy cases from the contemporary diagnostic workflow which had appropriate patient consent. These slides had been scanned immediately after being generated as part of the diagnostic workflow.

The H&E slides from both cohorts were produced and scanned in the Cellular Pathology Department at Oxford University Hospitals NHS Foundation Trust (OUHFT), scanning on a Philips UFS scanner.

The retrospective study was conducted under the ProMPT ethics (reference MREC 01/4/61). The prospective study was conducted under the Pathology image data Lake for Analytics, Knowledge and Education (PathLAKE) research ethics committee approval (reference 19/SC/0363) and Oxford Radcliffe Biobank research ethics committee approval (reference 19/SC/0173). Patients were not identifiable from the material. The research was performed in accordance with the Declaration of Helsinki.

2.2 PathProfiler - A pipeline for comprehensive quality assessment

Working with large cohorts requires an automated method for assessing whether a WSI (or a region of a WSI) is usable or not at the diagnostic level, with the assumption that an image of appropriate quality for diagnosis would be suitable for downstream computational pathology, including algorithm development, with diagnostically unusable images considered inappropriate for such use. If a WSI is considered unusable, then ideally there should be an indication as to whether an intervention, namely re-scanning or re-staining (then re-scanning) could potentially resolve the quality issue that had been detected. This would require also the categorisation of the WSI in terms of common artefacts present, and separately, where appropriate, an indication of the severity of the artefact (e.g. see table in Figure 1A).

We thereby designed PathProfiler (illustrated in Figure 1B) to indicate simultaneously the ‘usability’ of an image and the presence of artefacts. In this way, the algorithm can predict if a WSI is unusable, but also if it is associated only with severe issues with either focus or H&E staining. This facilitates the identification of the image within the pipeline for consideration of re-scanning or re-staining in order to improve the quality and render it ‘usable’. Alternatively, this approach facilitates the user to identify when re-scanning or re-staining would not resolve the quality issue such as when presented with data for a WSI with a significant number of intrinsic artefacts, dirt, ink or bubbles in the tissue area.

Firstly, a tissue segmentation model[24] extracts tissue regions. The current version of the algorithm requires patches of 256×256 at 5X magnification as input. Extracted patches are resized to 224×224 to accommodate for a ResNet18 CNN model. A multi-label pre-trained model predicts a set of quality measures for each patch, which includes the H&E staining and image focus, both indicated by a three-level score; 0 = no quality issue, 0.5 = slight quality issue, 1.0 = severe quality issue. This models the subjective pathologist assessment, whereby a score of 0.5 is assigned for a quality issue felt to be minimal and insufficient to render the WSI unusable, whereas a score of 1 implies severe artefact, potentially (but not necessarily) rendering the WSI unusable. Hence the predicted staining and focus scores can be interpreted as the severity of the artefact. The model also predicts the presence of additional artefacts, categorised as ‘tissue folding’ or ’other’, the latter including dirt, diathermy artefact, etc, although these are not associated with a severity ‘score’.

The predicted patch-level quality measures are collated to generate a WSI quality overlay for each category. Using statistical parameters of the quality overlays, we predict the subjective quality scores (i.e. those expected from an expert pathologist) at slide-level. These slide-level quality scores include overall usability of the WSI (binary 0 or 1), and a score from 0 (lowest quality) to 10 (highest quality) that predicts the overall quality of H&E staining and of image focus. In our study this score was in accordance with a 10-point UKAS-approved qualitative scoring scale utilised within the Cellular Pathology Department for assessment of H&E quality on glass slides. According to this system, a score of <=4 is considered a ‘fail’ for quality, 5-6 = pass, 7-8 = acceptable (good) level of quality, 9-10 = excellent (high) level of quality. With this information it is possible to predict the pathologists’ subjective quality scores indicating whether the WSI should be accepted as being of appropriate diagnostic quality, or rejected, and this is also considered indicative of the potential need for re-scanning or re-staining.

2.2.1 Data annotation

A subset of our prostate cancer cohort was annotated by specialist urological pathologists to facilitate training and testing. As described above, this dataset comprised 107 H&E stained WSIs of prostate tissue (biopsies and TURPs) from the ProMPT cohort and 91 H&E stained WSIs of contemporary prostate biopsy cases from the contemporary diagnostic archive. The selected slides provided a dataset of 1711 annotated image patches which was divided into the training set (80%), test set (10%), and validation set (10%), and a dataset of 198 annotated whole slides which was divided into the training set (60%) and test set (40%).

The breakdown of the annotated dataset of image patches is shown in Figure 1C, and the distributions of the annotated dataset of image slides for usability, focus and H&E staining is shown in Figure 1D.

Image patches were manually selected for annotation to cover various combinations of artefacts such as focus or H&E staining quality issues, tissue folding, dirt, and damage due to diathermy. While the pipeline used image patches of 256×256 at 5X, for the annotation task during development, we opted to use patches of 512×512 at 10X magnification as this was considered by the pathologists to be more representative of the way in which the WSI would be viewed in the diagnostic setting, thus facilitating more realistic images for subjective scoring. The area for pathologist annotation was a designated highlighted area of 712×1024 pixels within a 512×512 box in the centre of the image patch; providing the highlighted image within the larger patch for assessment was considered to provide a context more representative of the diagnostic setting. The annotation

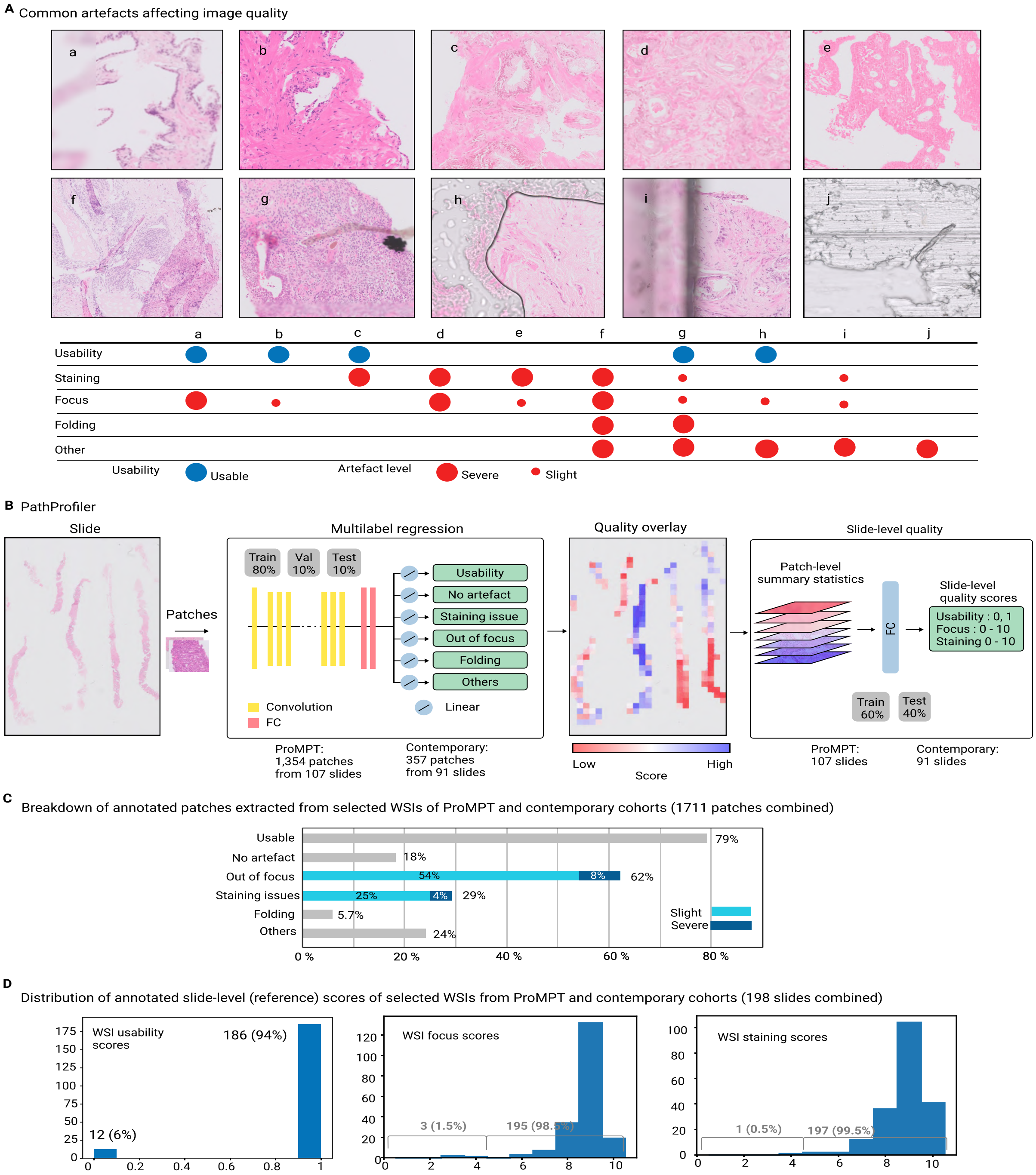


Figure 1. (**A**) H&E images demonstrating commonly encountered artefacts affecting digital WSI quality. Image quality may be affected by multiple artefacts, or just one, and focus quality can be specifically related to scanning/focus issues or it can be a result of other artefacts. (a,b) focus issue, (c) fading/loss of contrast of H&E stain, (d-e) both staining and focus issues (f-g) a combination of focus and staining issues with tissue folding and dirt, (h) bubble under the coverslip and slight focus issues in visible tissue area, (i) edge of the coverslip (affecting focus), (j) coverslip glue. (**B**) PathProfiler quality assessment pipeline. After tissue segmentation, patches of 256×256 are extracted at 5X magnification and resized to 224×224 to accommodate for the ResNet18 CNN model. For each patch, the trained model predicts the presence of an artefact, and for focus and H&E staining artefacts it also predicts a quality score for each patch, 0 = no quality issue, 0.5 = slight quality issue, 1.0 = severe quality issue. A quality overlay is generated for each output category. In the next step, we map the predicted quality overlays to the slide-level standardised scoring system. For this, statistical parameters of quality overlays are used to predict slide-level quality scores; overall usability of the WSI (binary 0 or 1), and a score 0-10 for quality of focus and H&E staining from the lowest quality to highest quality, where the cut-off score for acceptable quality for diagnostic purposes is 4. (**C**) The composition of the pathologist-annotated image patches extracted from selected WSIs of ProMPT and contemporary cohorts (combined) (**D**) The distribution of pathologist-annotated (reference) quality scores of WSIs selected from ProMPT and contemporary cohorts (combined) for WSI usability (binary 0 or 1), focus and H&E staining quality (0-10, as above).

Label	Criteria	Value
y_1	Usability	1 - appropriate for diagnosis 0 - otherwise
y_2	No artefact	1 - no presence of any slight or severe artefacts 0 - otherwise
y_3	Staining artefacts	1 - severe staining or H&E contrast issues 0.5 - slight staining or H&E contrast issues 0 - no staining or contrast issues
y_4	Focus artefacts	1 - severe focus artefacts 0.5 - slight focus artefacts 0 - no focus artefacts
y_5	Tissue folding	1 - the presence of tissue folding 0 - otherwise
y_6	Other artefacts	1 - the presence of other artefacts such as dirt, glue, ink, cover slip edge, diathermy, bubbles, calcification and tissue tearing. 0 - otherwise

Table 1. Patch level labels. The proposed annotation protocol captures that a given image patch can be corrupted by multiple different artefacts. This protocol also captures if an image patch is diagnostically usable despite being affected by some artefacts.

task was then performed by the pathologist using a digital pathology workstation clinically validated for diagnostic work. For each image the pathologist provided quality data within a drop-down menu in terms of the overall usability of the image patch (for diagnostic interpretation), the presence and severity of focus and/or H&E contrast issues, the presence or absence of additional specific artefacts; folding or other artefacts. The corresponding labels are listed in Table 1. Subsequently, a multi-label target $\mathcal{Y} = [y_1, ..., y_6]$ was associated with each image patch. The composition of our annotated (combined ProMPT and contemporary cohort) multi-label patch dataset is illustrated in Figure 1C.

For whole slide quality assessment, three specialist urological pathologists independently assessed the 198 WSIs from the combined ProMPT and contemporary dataset for: (1) overall usability - a binary label if the slide can be used for diagnosis, (2) focus quality, and (3) H&E staining quality. Quality of focus and H&E staining assessment were standardised from 0 (lowest focus quality) to 10 (highest focus quality), utilising the qualitative scale discussed before. A score of $<= 4$ is considered a fail for quality, with 9/10 being equivalent to excellent quality. To aggregate individual assessments, the focus and staining quality scores were averaged between the three assessors. However, we used the “AND” logical operator to aggregate the diagnostic usability of slides such that a slide is rendered usable only if all three assessors considered it as being usable. Whilst this approach may increase the rate of false-positive for identifying unusable slides, it was preferred over missing such cases. The aggregated reference WSI quality scores from the ProMPT cohort included 97 (91%) usable and 10 (9%) unusable WSIs, and from the contemporary cohort 89 (98%) usable and 2 (2%) unusable WSIs. This subjective assessment mirrored the overall perception of the pathologists that the majority of WSIs in both cohorts were of suitable quality for diagnosis. The distribution of annotated quality-related scores of the combined WSIs (198) are shown in Figure 1D.

2.2.2 Multi-label neural network training

In our multi-label dataset the d-dimensional input features in $\mathcal{X} = \mathbb{R}^d$ are associated with the multi-label target space $\mathcal{Y} = [y_1, ..., y_c]$ where $[y_3, y_4] \in \{0, 0.5, 1\}^2$ and $[y_1, y_2, y_5, y_6] \in \{0, 1\}^4$ with a total of $3^2 \times 2^4 = 144$ possible combination of unique labels. So, the multi-label model learns a function $f : \mathcal{X} \to \mathcal{Y}$ from the training data $\mathcal{D} = \{(\mathcal{X}_i, \mathcal{Y}_i) | i = 1, ..., m\}$.

Whilst the problem here could be treated as a combination of single regression or single binary classification tasks, we found it more efficient to use a multivariate regression model. First, a multi-label model can efficiently exploit correlations among labels[25]. Second, we ran multiple tests and observed that label noise in dataset preparation for studying quality was inevitable. This was mainly due to the inherent challenges in subjective quality assessment, including both intraobserver and interobserver variation and in the definition of acceptable quality thresholds (for diagnosis), which again varied. We attempted to reduce the “noise” in this data through provision of guidance for standardisation of quality scoring (e.g. the UKAS approved 0-10 scale), and through regular discussion amongst the pathologists and the wider group to encourage consistency in interpretation of quality, with the provision of example images. The study took place over a prolonged period, making such standardisation important, to reduce the impact of noise. The pathologists used the same workstation throughout for assessment of WSI to minimise bias.

Strong classification loss functions do not always perform well under noisy labels[26]. The Binary Cross Entropy (BCE) loss function for instance highly penalises large loss values, leading to a model learning in favour of outliers rather than inliers. Whilst different approaches are proposed to efficiently learn from noisy data[27], we employed the Huber loss function which is in the category of robust loss functions to label noise. It is shown that Huber loss can provide efficient prediction in deep learning under minimal assumptions on the data[26]. Huber loss function combines the benefits of mean squared loss for small loss values while rejecting the dominance of outliers similar to mean absolute loss for large loss values[26].

For data augmentation, we arbitrary rotated patches (0°, 90°, 180°, 270°) and flipped them along the vertical or horizontal axis. A random affine transform including shear from -10 to 10 degrees and translation by -5 to 5 percent per axis, were also applied.

We trained an 18-layer deep ResNet[28] model (ResNet18) which takes 3-channel input images at a size of 224×224 pixels. The last fully connected layer was modified to output six classes with linear activation functions. Multi-label classification was performed with a Huber loss with $\delta = 1$, as shown in equation (1) and an Adam optimiser.

$$L_\delta(x,y) = \begin{cases} \frac{1}{2}(x-y)^2 & \text{for } |x-y| \leq \delta, \\ \delta\,|x-y| - \frac{1}{2}\delta^2 & \text{otherwise.} \end{cases} \tag{1}$$

During training, we down-sampled annotated image patches of 512×512 at 10X to the network 224×224 pixels input requirement. Therefore, we effectively evaluated slide quality at 5X magnification.

Training with initialised weights, pre-trained on the ImageNet dataset, was performed with hyper-parameters of batch size= 100 and learning rate= 1e-4. The dataset was split into three partitions for training (80%), testing (10%) and validation (10%) with stratification based on their unique multi-label combinations. We ran the model training for 200 epochs and selected the model based on minimum loss for the validation set. To partially handle the label imbalance, a weighted batch sampler was used.

2.3 Quality assessment of cohorts

While the models were trained on annotated data from prostate slides we also tested the model on other cohorts to assess its generalisability. Firstly though we applied PathProfiler on the entirety of the available H&E stained WSIs (3819 WSIs) from the ProMPT cohort to provide a prediction of usability of the whole cohort, and to assess the burden of artefacts present within the cohort, seeking to potentially identify WSIs with quality issues that might be resolved.

To analyse the performance of our tool on other previously unseen WSI of prostate tissue, we selected slides from the TCGA multi-institutional collection. All 449 examples of H&E WSIs of prostate tissue in TCGA were analysed. These included mostly RP cases, and a range of preparations, including frozen tissue as well as the FFPE tissue, the latter more routinely used in clinical practice[29]. Noteworthy is that the focus of the tissue samples collected within the TCGA cohort is different to that of our cohorts, with samples collected primarily for molecular analysis, hence the predominance of H&Es from frozen sections in TCGA rather than FFPE tissue such as that present in our training/test cohort. As such we anticipated that the cohort would vary from that on which PathProfiler had been trained. The quality data sought from the analysis included the parameters assessed on our original cohorts, i.e. overall usability, focus and H&E staining quality, and presence of other artefacts, however the model provided estimates of these parameters rather than predictions as we did not have a reference standard pathologist quality assessment for the TCGA WSIs.

Finally, WSIs from another cancer cohort were assessed to determine the functionality of our model on other tissue types. For this purpose we opted to utilise PathProfiler on the FOCUS cohort (Fluorouracil, Oxaliplatin, CPT11 [irinotecan]: Use and Sequencing), which is a collection of 788 WSIs from a dataset of colorectal cancer specimens taken from the MRC FOCUS clinical trial[30]. These were WSIs of resected tumour from 375 advanced colorectal cancer patients. Glass H&E slides from this cohort were reviewed by a specialist gastrointestinal pathologist; tumour and the associated intratumoural stroma were annotated and used to guide RNA and DNA extractions for the purpose of other studies. All H&E slides were later scanned (in 2016) at high resolution on an Aperio scanner at a total magnification of 20X. WSIs were then re-reviewed by a second gastrointestinal pathologist and tumour annotations were traced to generate region annotations for machine learning classification[31].

The quality data sought from the analysis of the FOCUS cohort included the parameters assessed on our original prostate cohorts and the TCGA cohort. However, for FOCUS, slide-level scores were calculated based on tumour area only. It was considered more appropriate to limit the assessment to the tumour because there were large areas of tissue, such as fat, in many of the WSI that were not relevant diagnostically and we wanted to avoid this tissue skewing the data. However, at patch-level the average estimated quality scores were calculated for patches extracted from all tissue regions.

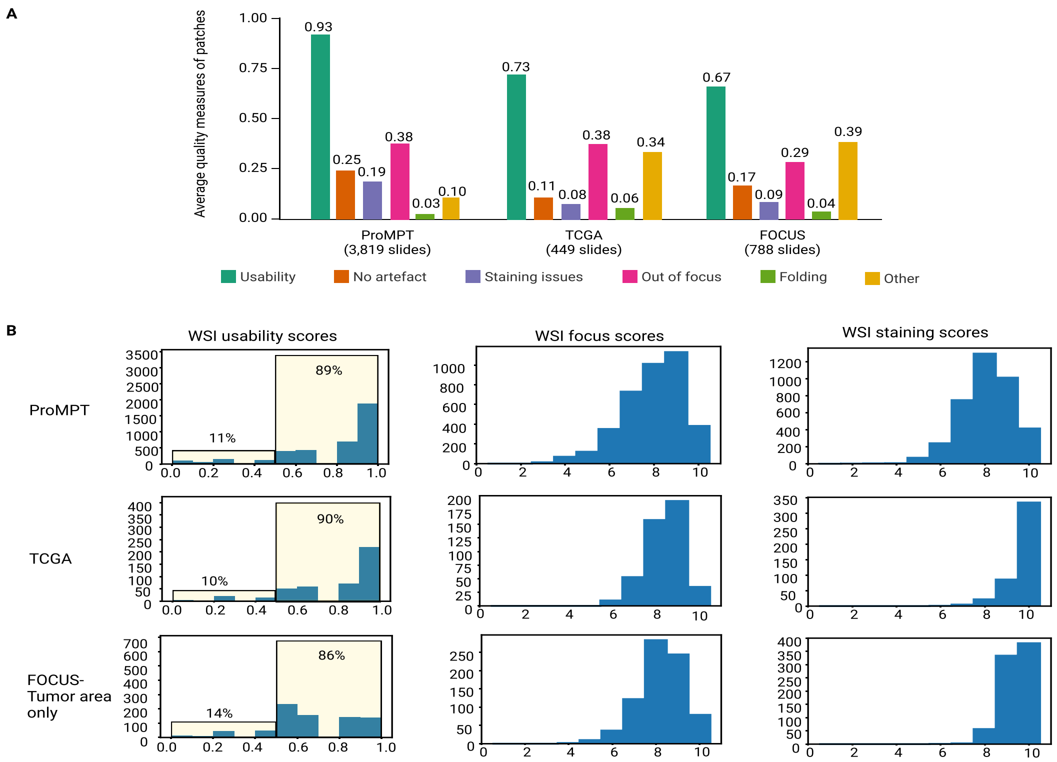


Figure 2. (**A**) Average estimated quality measures of patches extracted from the entire ProMPT cohort (3819 WSIs), and the TCGA-prostate (449 WSIs), and FOCUS (788 WSIs) cohorts. (**B**) Distribution of estimated usability scores of WSIs in each cohort (entire ProMPT cohort, TCGA, and FOCUS) and distribution of estimated scores for WSI focus and H&E staining quality (0-10, lowest quality to highest quality). WSI scores for the FOCUS cohort are calculated for tumour regions only.

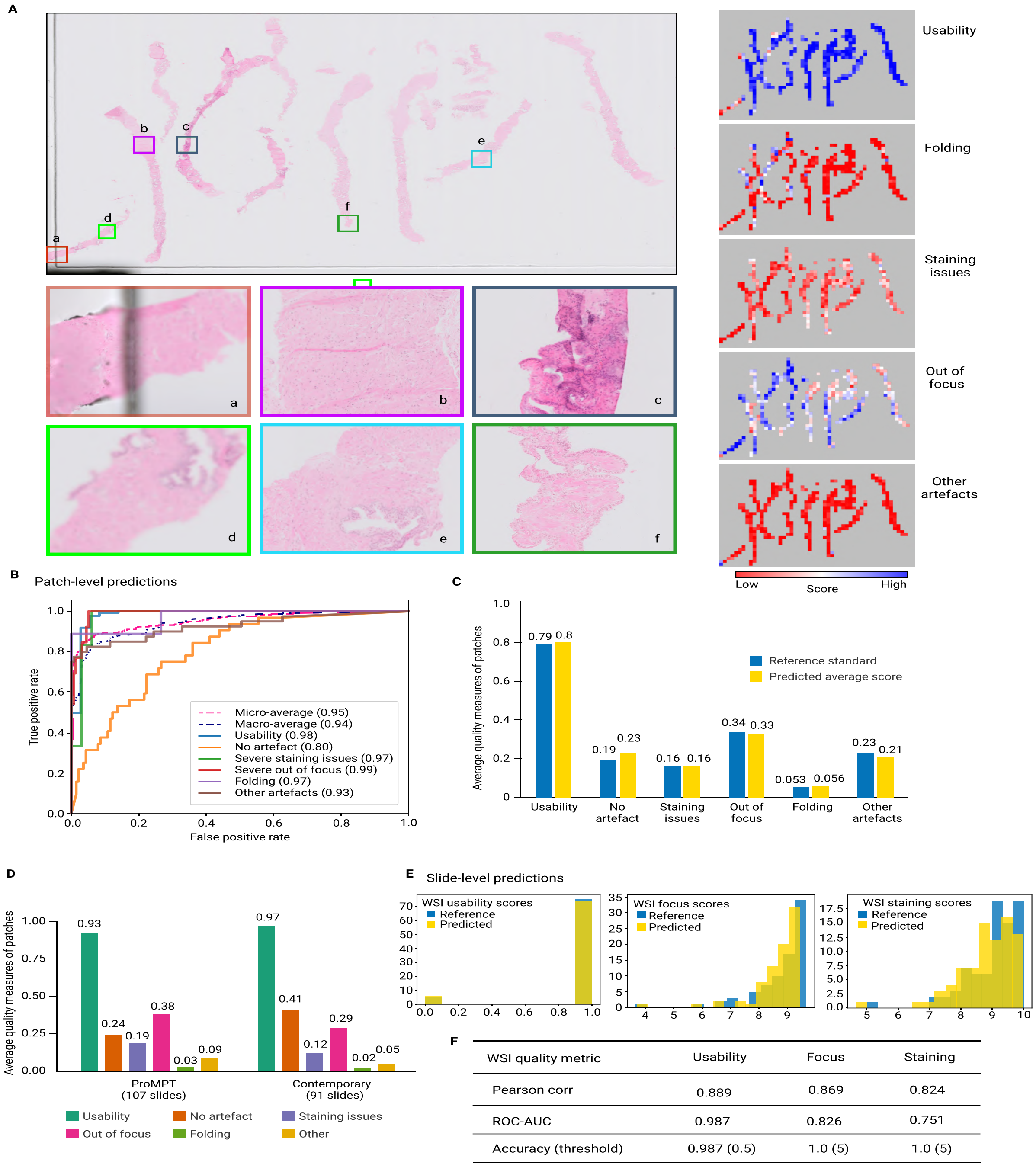
3 Results

3.1 Quality assessment of prostate cancer cohorts

We utilised PathProfiler to analyse the entire ProMPT cohort and the 449 WSIs of prostate tissue from TCGA. The breakdown of predicted patch-level artefacts is shown in Figure 2A. Our model predicts that on average 93% of image patches in the ProMPT and 73% of image patches in the TCGA prostate cohort are 'usable' according to our criteria (and threshold). The main quality impacting artefacts in the ProMPT cohort are predicted to be related to focus and staining issues, with an average probability of 0.38 for focus quality issues and an average probability of 0.19 for staining quality issues.

Image patches extracted from TCGA slides are highly associated with 'other' artefacts, affecting the usability of the slides in comparison with the ProMPT cohort (average patch-level probability of 0.34 vs 0.1). On direct visualisation of the TCGA WSIs it is seen that many of the 'unusable' regions are related to the presence of 'ink' applied to the glass slide (prior to scanning) by pathologists to highlight areas of interest on the glass slide and which are retained on the WSI. Ink artefact by comparison is rarely seen in the ProMPT and contemporary prostate biopsy cohorts as the glass slides were cleaned of ink to enable the scanner to focus on the tissue rather than the ink.

Predicted slide-level quality scores are shown in Figure 2B. Our model predicts that 11% of WSIs in the ProMPT cohort are unusable. Whilst we have defined the threshold for usability at the subjective level of that regarded as minimum standard for diagnostic interpretation, the threshold for the cut-off means that a slight shift of predicted score impacts significantly on the binary 'usable' vs 'unusable' and perhaps it is better to regard the term 'unusable' as the cut off at which an image is flagged for manual quality review. According to our criteria, 10% of WSIs in TCGA are estimated to be unusable. In addition, at least 2% of slides in ProMPT are predicted to benefit from re-scanning (84 slides) while at least 0.45% are predicted to benefit from re-staining (17 slides). These values are reported based on the cut-off threshold (<=4) for focus and staining quality scores. However, our investigations of the pathologists' annotations show that many slides annotated as 'unusable' with quality scores of 5 or 6 are still suggested for a re-scan/re-stain to solve the quality problem. Therefore, from the distribution of WSI focus and staining scores shown in Figure 2B, we can conclude that most of 'unusable' slides in the ProMPT cohort would benefit



WSI quality metric	Usability	Focus	Staining
Pearson corr	0.889	0.869	0.824
ROC-AUC	0.987	0.826	0.751
Accuracy (threshold)	0.987 (0.5)	1.0 (5)	1.0 (5)

Figure 3. (**A**) Quality overlays for slide 49_1 from the ProMPT cohort. H&E WSI (top left) and predicted quality overlays (heatmaps, right). The predicted usability overlay suggests that the probability of most image patches being usable is high. The individual artefact heatmaps indicate patches predicted to show the various artefacts, e.g., a = 'other' artefact, in this case coverslip edge and unusable, b & c = folded tissue (although the probability of b being usable is higher than c), d & e = severe focus issue predicted, affecting usability of d but without predicted impact on usability of e, f = slight H&E staining issue predicted (again with minimal impact on usability). Predictions like these could be used in an image analysis pipeline or be made available to a pathologist for regional investigation of artefacts. (**B** & **C**) Model performance for test dataset of image patches (combined ProMPT/contemporary): (**B**) ROC-AUC curves for each category of artefacts and overall usability, and (**C**) average predicted quality measures versus the reference standard as assessed by the pathologist. (**D**) Average predicted quality measures of all image patches in the 'selected' ProMPT slides (107 slides) and contemporary archive (91 slides). Such summarised patch-level quality assessment of images provides general cohort-level quality metrics for computational pathology. (**E**) Distribution of 'reference standard' vs 'predicted' slide-level usability and quality scores for focus and H&E staining (0-10, low to high quality) for the test dataset of image slides (combined ProMPT/contemporary). (**F**) ROC-AUC and Pearson correlation coefficient of the slide-level predictions for usability and quality scores (0-10, low to high quality) and the reference standard values for the test dataset of image slides (combined ProMPT/contemporary).

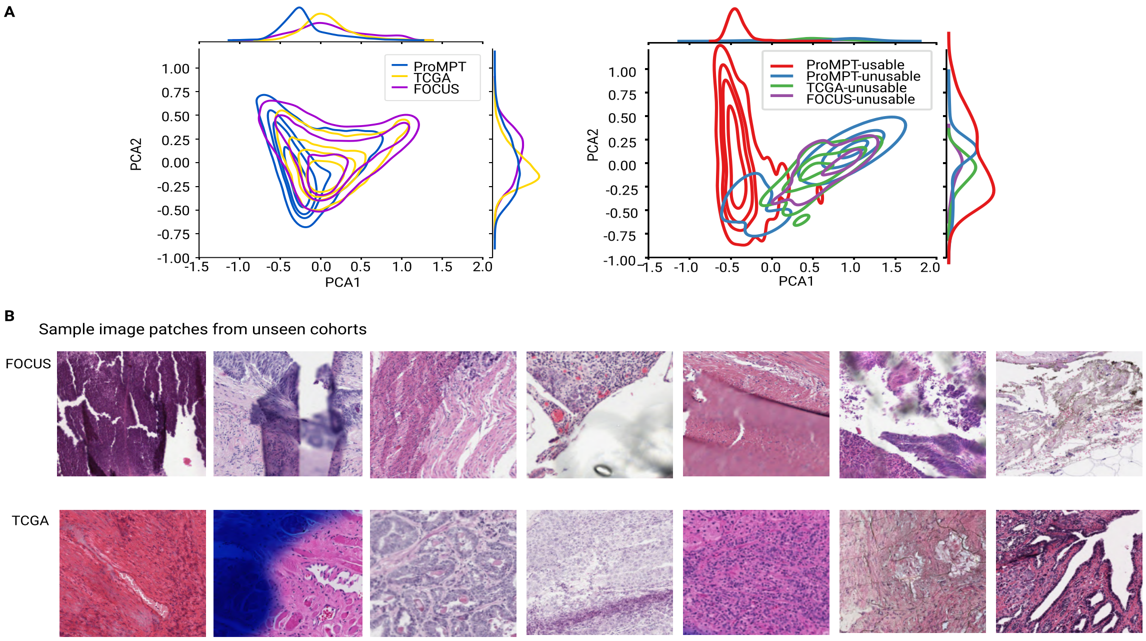


Figure 4. (**A**) Left: Kernel Density Estimation (KDE) plot of 2D PCA feature space for 16000 random patches extracted from ProMPT, TCGA and FOCUS cohorts. Whilst there is still room for improvement, the overlap between feature spaces of different cohorts suggests that we have a reasonable domain-invariant set of features. Right: Feature space of annotated patches in our patch test set from ProMPT cohort vs a set of annotated patches from TCGA and FOCUS cohorts. (**B**) Sample image patches predicted by PathProfiler as unusable from non-prostate tissue (FOCUS colorectal tissue) and TCGA (prostate tissue) cohorts.

from either re-scanning or re-staining.

3.1.1 Patch-level quality assessment for selected ProMPT and contemporary WSI cohorts

A sample WSI from the ProMPT cohort with the detailed quality overlays is presented in Figure 3A. A heatmap is generated for the overall usability of each WSI at a patch-level, providing a simple visual cue as to the quality of regions of the image, clearly indicating poor quality areas for further assessment. The quality heatmaps are colour-coded to indicate the probability of a feature being present. As seen in Figure 3A the heatmap indicates that the probability of most of the patches being ‘usable’ is high, with lower probability of staining issues, folding and other artefacts. The heatmap for ‘focus’ however indicates more patches predicted to have high probability of a focus-related quality issue (note that here the white colour indicates slight while blue indicates severe focus issues), although correlating this with the ‘usability’ heatmap it can be seen that these issues do not invariably impact overall diagnostic ‘usability’. Thus, by comparing the overall usability heatmap with the individual heatmaps (indicate the various specific quality measures analysed) it is then possible to identify precisely which artefact(s) impact on the overall quality, and then assess whether an intervention, such as re-scanning, would improve the image. Heatmaps such as these could be integrated into a research pipeline and used to highlight artefact regions within a WSI for further assessment.

The performance of the proposed multivariate model on the test dataset of image patches is reported in Figures 3B and 3C. The ROC-AUC for out of focus and staining artefacts is 0.85 and 0.84 respectively. However, as reported in Figure 3B, the ROC-AUC scores for detection of focus and stain artefacts when these are severe, and therefore likely significant in relation to usability, indicate a higher model accuracy for such cases (0.99 and 0.97 respectively). This is likely a reflection of greater inter- and intraobserver agreement for severe artefact, in comparison with what was interpreted as a ‘slight’ artefact, with many of the latter cases showing minimal focus/staining artefact which in fact really fall on the borderline with "no artefact".

A summary of predictions for the test dataset of image patches (combined ProMPT and contemporary) versus the reference standard (pathologist assessment), is shown in Figure 3C. This shows that the average model predictions of the presence or absence of artefact, and for the ‘usability’ of an image, closely matches the subjective assessment by the pathologist.

We then utilised the validated model to predict the average quality measures for all patches in the selected ProMPT cohort and all patches in the contemporary prostate cohort separately (Figure 3D). The results indicate that the average usability of image patches is predicted to be 93% for the selected ProMPT cohort and 97% for the contemporary cohort.

3.1.2 Slide-level quality assessment

A predicted quality assessment at slide-level can be achieved by summarising the assessment made at patch-level. Whilst such a summarised quality assessment is useful to identify dominant artefacts in a WSI, and thus those likely to impact on usability, it does not express a standardised scoring system. We therefore sought to fit a simple model to map the predicted patch-level measures to the subjective quality scores (0-10, from low to high quality) in accordance with the reference scores provided by an expert pathologist.

The slide-level subjective quality scores include a binary usability score of 0 or 1, and a standardised H&E staining score and a focus score, both from 0 to 10 (low quality to high quality, as previously). The dataset including whole slide subjective quality scores, was split into 60% train and 40% test with stratification based on the quality scores. Three separate linear regression models were then fitted to predict WSI usability, focus and staining scores from the mean and variance of selected quality overlays.

Reference standard subjective scores implied that 9% of WSIs in the selected ProMPT cohort (107 slides) and 2% of WSIs in the contemporary cohort (91 slides) were unusable, whereas our model predicted this to be 12% and 2% respectively. Such slide-level analysis therefore implies that the quality/usability of WSIs in the ProMPT cohort is overall less than that of the contemporary cohort of cases, which is in keeping with the perception of the pathologists regarding the cohorts overall.

The predicted slide-level subjective scores can then be used to provide summary metrics for the whole cohort, including both the percentage of usable/unusable slides, and a histogram of WSI quality scores. As shown in Figure 3E, our model-predicted assessment of scores for the selected slides of ProMPT and contemporary cohorts are highly comparable with the subjective pathologist assessment of the WSIs.

The Pearson correlation coefficient and the ROC-AUC of the predicted usability, focus and staining scores (versus reference standard) for the combined test set of slides is reported in Figure 3F. This shows that the quality measures as predicted by PathProfiler closely align with the reference standard. The calculated accuracy of binary focus and staining scores with the cut-off $<= 4$, is 1, and the accuracy of the predicted usability score is 0.987.

3.2 Non-prostate tissue cohort

Patch-level predictions of quality measures for all patches extracted from the 788 WSIs from the FOCUS cohort are presented in Figure 2A. As for TCGA, many patches from this cohort are affected by 'ink', categorised as 'other' artefacts in our pipeline. Slide-level usability and standardised focus and staining quality scores for this cohort have been calculated for tumour regions only (Figure 2B). Our model predicts that 86% of WSIs and 67% of image patches in the assessed FOCUS cohort are 'usable' according to our criteria (Figures 2A and 2B)

Whilst we have trialed our quality assessment pipeline to provide a quality estimate for an external prostate tissue cohort (TCGA) and non-prostate tissue cohort (FOCUS), using the same measures as we have assessed for the local ProMPT and contemporary prostate WSI cohorts, we must caveat that the model has not been trained on external prostate tissue WSIs nor on non-prostate tissues, both of which may harbour novel artefacts that the model has not been exposed to. For this reason, we do not claim that the results presented are fully reliable for these other cohorts. However, we have taken measures to assess for the presence of bias which may impact on the results, and we have shown that there is a good overlap between the feature space of the different cohorts, as shown in Figure 4A.

Figure 4A (left image) illustrates a 2D PCA feature space of 16000 patches randomly extracted from each of the ProMPT, TCGA, and FOCUS cohorts. We also selected a small set of patches (66 from TCGA and 37 from FOCUS) predicted as unusable, some of which are included as examples in Figure 4B. We then plotted the feature space of the unusable patches versus a set of usable and unusable patches from the ProMPT cohort. As seen in Figure 4A (right image), the feature space of unusable image patches from all cohorts has a good overlap, suggesting that we have a set of features that are less biased to our cohort and maybe applicable to other cohorts to some extent. In our opinion therefore we have observed that PathProfiler-predicted quality measures for usability, focus, and other artefacts seems to provide a meaningful overall quality assessment of other cohorts.

3.3 Comparison with other quality assessment tools

We used our annotated dataset of image patches at 10X and a set of hand-crafted features exploited in HistoQC to generate separate Random Forest classifiers to identify usability and each of the artefacts. Features used from HistoQC include TenenGrad contrast, RMS contrast, Michelson contrast, grayscale image mean, median and variance, mean of filtered image with Gaussian, Laplace, Frangi and Gabor filters, Local binary patterns (LBP), per channel mean of RGB and HSV and deconvolved H&E stain channels (available at `https://github.com/choosehappy/HistoQC`).

A

ROC-AUC

	Usability	No artefact	Staining issues	Severe staining issues	Out of focus	Severe out of focus	Folding	Other
HistoQC	0.86	0.67	0.56	0.49	0.78	0.59	0.49	0.81
Ours- hand crafted	0.92	0.64	0.70	0.66	0.72	0.77	0.55	0.86
PathProfiler (deep features)	0.98	0.80	0.84	0.97	0.85	0.99	0.97	0.93

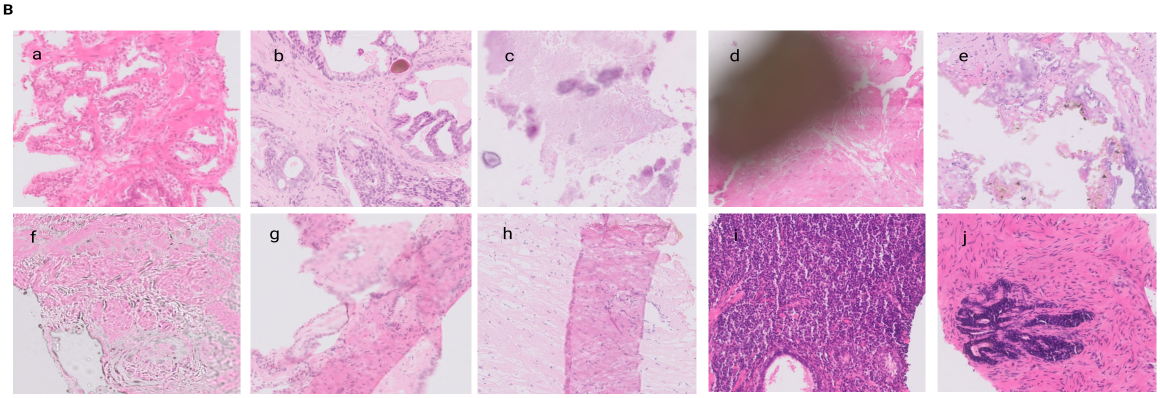


Figure 5. (**A**) Quality assessment of our annotated patch dataset (from ProMPT and contemporary cohorts) using other quality tools. (**A**) Comparison of performances of HistoQC, our proposed set of hand-crafted features, and PathProfiler for usability and artefact classification. (**B**) Examples of sample artefacts that are challenging to identify using hand-crafted features; (a-b) staining issues cannot be generally related to the brightness of images, (c-d) 'other' artefacts such as calcification (c) and dirt (d) that cause regions of low intensity variation are confused with out-of-focus regions. This may falsely reinforce recommendation by the algorithm for a slide re-scan, although this will not resolve those artefacts, (e-f) simple hand-crafted features such as Laplacian filtering misclassify unusable regions that contain rapid intensity changes as usable, (g-h) hand crafted features mostly associate folded tissue with darker colours in the data and therefore cannot detect folded areas within the range of average data colours, and (i-j) thicker tissue in a section may be misclassified as folded tissue.

We also tried a different set of hand-crafted features that performed better overall on our dataset in comparison with HistoQC feature set, as shown in Figure 5A (denoted 'Ours-hand crafted'). The features include Modified Laplacian (LAP2)[32] focus measure, Variance of Laplacian, TenenGrad contrast, average difference of image from Gaussian filtered image, average variance along R, G and B channels (along third axis), per channel mean of RGB, HSV and deconvolved H&E stain channels and Otsu threshold. The ROC-AUC of classifiers generated using HistoQC features and our set of hand-crafted features versus PathProfiler are reported in Figure 5A. As seen, the performance of our multi-task deep learning model in PathProfiler was superior to classifiers that used HistoQC or our 'hand-crafted' features.

We found multiple challenges in using hand-crafted features for our prostate specimen cohorts, including those previously proposed by Janowczyk *et al*[23]. For example, H&E staining issues detected in the ProMPT cohort cannot be directly correlated with image brightness, and we found that overall the available hand-crafted features were not successful in the detection of various staining issues, such as those illustrated in Figure 1A (c-f) and Figure 5B (a-b).

Furthermore, many available quality assessment tools measure the intensity variation in an image (e.g. using edge-detection filters) to identify blurred or unusable regions[13,20,23]. We found that this resulted in large gland-free stromal regions in our data being designated as 'unusable', an issue also previously recognised by Janowczyk *et al*[23].

We faced further challenges; artefacts that result in reduced variation in colour or intensity, such as H&E poor staining Figure 1A (c-e) and Figure 5B (a), cover slip edges Figure 1A (i), calcification and ink/dirt Figure 5B (c,d) are confused with out-of-focus regions. This is undesirable as this will incorrectly identify the slide as needing a rescan, which will not resolve the quality issue. On the other hand, unusable regions that contain rapid intensity changes such as Figure 5B (e,f) were detected as usable with hand-crafted features.

Finally, we found that the hand-crafted features dramatically failed in the detection of folded tissue in our dataset. This would appear to be the result of the association of tissue folding with darker (H&E) colours[23] while it rather seems to be recognisable by more complicated features (Figure 5A). For example, hand-crafted features failed in the detection of folded regions in Figure 5B (g,h) and Figure 1A (f) as the colour intensities in these images are generally in the range of average colour values in the dataset. On the other hand, areas where the tissue section was thick and the H&E consequently appeared dark, such as in Figure 5B (i,j), were misclassified as folded tissue.

4 Discussion

Given the importance of retrospective cohorts in the current research landscape, perhaps surprising is that to-date the factors impacting on the quality of glass slides and subsequently on WSIs, have not been fully assessed, and yet as DP becomes more readily available, more retrospective cohorts are progressively being digitised in academic centres for purposes such as algorithm development.

Digital image quality issues recognised within a diagnostic setting are either related to intrinsic tissue factors or to slide processing factors such as H&E staining, dirt, tissue folding, and to digital scanning per se, such as focus issues. Whilst these issues are also applicable to images evaluated in a retrospective research setting, the window between glass slide creation and digitisation also introduces additional quality challenges related to slide ageing , such as fading and drying back of the coverslip glue.

In this study we have investigated the potential for development through deep-learning of an AI model to automate the assessment of WSIs for quality related measures, using a subjective cut-off threshold for diagnostic usability of a WSI, as assessed by an expert pathologist, as being the minimum acceptable quality standard for a WSI for computational pathology. We set out to achieve this utilising the ProMPT cohort as a representative and academically important example of a historic glass slide cohort. In so doing, we have developed a pipeline for quality assessment which facilitates the identification of quality-related issues for a WSI in the form of a heatmap, providing focus for further direct visualisation of problematic areas of an image in an attempt to determine whether a specific intervention such as re-scanning or re-staining (H&E) of a glass slide will be of benefit in improving the image quality. Whilst we have developed our model, designated as ‘PathProfiler’, on local prostate tissue cohorts including ProMPT, we have also attempted to demonstrate the functionality of PathProfiler on other tissue cohorts. Finally, we have assessed the functionality of PathProfiler in comparison with other currently available quality assessment methods.

Intuitively, quality-related issues might be expected to be more frequently encountered in cohorts of older slides due to degradation, however to-date there is little evidence to support such an assumption. We therefore sought such evidence at the outset of our study by investigating the frequency of significant quality issues, as judged by a specialist pathologist on the basis of whether an image was usable for diagnostic purposes. We compared the findings within the WSIs of a cohort of historic glass slides with those from a newer ‘contemporary’ cohort. As an example of an historic cohort, we had access to the collection of almost 4000 glass H&E slides of prostate tissue (mainly biopsies) from the ProMPT observational study, which were collated between 2001-2018, and retrospectively digitised. For the purpose of algorithm development, a representative set of WSIs from 10% of cases (107 WSIs) from this cohort were subjectively assessed for quality related features and overall diagnostic usability, and 9% of these were considered to be ‘unusable’ in their current state. This compared with 2% of cases from a comparably sized contemporary cohort of WSIs of prostate biopsies from 2019, scanned at the time of slide preparation. These figures may be an underestimate of quality given that they are based upon the requirement for subjective agreement on usability from all three participating specialist pathologists, however for algorithm development purposes it was felt that capture of all potential quality issues would be more beneficial at the outset. From the annotated WSIs it appears that the impact on ‘usability’ is likely to be attributable to more significant focus quality issues seen in the ProMPT cohort compared with the newer cohort, although this subjective assessment indicated that the quality of H&E staining was also an issue more frequently seen in the ProMPT WSIs, as might be expected given the recognition that H&E staining fades over time. These glass slides from both cohorts had been prepared at the same institution and whilst there may have been minor variation in H&E preparation given the period over which they were produced, this would be expected to be minimal, and therefore any differences can reasonably be considered to be related to slide age. What became apparent also was that the un-usability of images was not necessarily related to a scanning issue per se; other intrinsic tissue-related features such as folding and other artefacts such as cover slip edge, dirt and calcifications also impacted on usability (sample images and their quality assessment are shown in Figure 1A). This was important to recognise at the outset as we sought to develop a pipeline that could predict not only the presence of a significant quality related issue, but one that could potentially be remedied by rescanning or restaining the tissue section. The ability of the model to both detect and classify quality issues that could not be remedied was therefore important in ensuring the accuracy of the tool in this task.

At both patch-level and slide-level for these prostate cohorts, the PathProfiler model we have developed is able to accurately predict the usability of an image, with an ROC-AUC at slide-level of 0.987 (compared with the reference standard pathologist assessment, Figure 3F). At slide-level the ROC-AUC for focus and H&E staining overall is 0.826 and 0.751 respectively. At patch-level, if evaluation is limited to severe artefact in these categories, which is of significance in terms of usability, the

ROC-AUC is improved from 0.85 to 0.99 and from 0.84 to 0.97 respectively (Last table row in Figure 5A). As such, PathProfiler predicts that 12% of the selected ProMPT cohort (107 WSI) and 2% of the contemporary cohort (91 WSI) are 'unusable' at a diagnostic threshold, which is very similar to the reference standard (9% and 2% respectively). It should be noted that the cut-off threshold from the predicted scores was at a level at which interobserver agreement may not be high, i.e. at the mid-range of the scale (Figure 3E). In reality, predicted values are from a continuous range (0 to 1 for usability, and 0 to 10 for focus and staining scores) and therefore the cut-off thresholds could feasibly be shifted in accordance with local preference.

We went on to use PathProfiler to assess the same quality measures across the entirety of the ProMPT cohort (3819 WSI), and as such the predicted overall usability (of the WSIs) of the whole cohort is 89% at slide level. This figure would seem to be aligned with that expected on the basis of the original subjectively assessed WSI cohort, and with the perception of the pathologists working with the cohort, and perhaps it is a useful benchmark for expectations for similar single-institutional historic cohorts, although this is unknown. Potentially more usefully, PathProfiler predicts that at least 2% of the ProMPT cohort (84 slides) could see quality improvement from re-scanning, and at least 0.45% (17 slides) from re-staining. Such relatively minor interventions would be worthwhile in such a precious cohort, and anecdotally we have certainly seen improvements in image quality from slides in this cohort which have been re-stained. Furthermore, specifically in relation to prostate biopsy cases, even if the originally scanned H&E WSI is eventually deemed unusable, the routine availability of multiple H&E levels on a block, often put onto different H&E glass slides, offers the potential opportunity to scan alternative slides from the same case.

For our model to provide potentially meaningful impact in the setting of curation of WSI cohorts for research, we sought to embed it within a quality assessment 'pipeline', which would provide user friendly output data to enable the operator to identify WSI for further attention. This might in the end require their removal from a cohort, but it could be that the identification of a slide for re-scanning or re-staining may result in resolution of the quality issue. To this end our output data was assimilated from patch-level to slide-level predictions in the form of a heatmap providing a visual cue as to the 'unusable' areas of a WSI, and then with overlays corresponding to the classification of the artefact, Figure 3A. This quickly enables the user to focus attention on a problem area to assess whether this can be remedied. In addition, PathProfiler provides image quality scores at both patch-level and slide-level, the former being particularly relevant during algorithm development with the possibility of excluding poor-quality patches from assessment, rather than removing the whole image being rendered unsuitable.

The ProMPT prostate cohort was considered at the outset to be representative of other retrospective historic cohorts in terms of quality in that it was a non-curated resource at the time of scanning. However, we wished to investigate the functionality of PathProfiler on other prostate tissue cohorts, and for this purpose we identified the TCGA prostate cohort, being an example of a publicly available multi-institutional dataset that is already in use for computational pathology. We also applied PathProfiler to a non-prostate tissue cohort, the MRC FOCUS colorectal cancer dataset. These additional assessments were undertaken with caution given that features associated with image quality assessment can vary widely among different cohorts due to inherent tissue characteristics, differences in tissue handling (e.g. section thickness) and preparation (FFPE versus frozen section), and to differences in digitisation of the WSI not only associated with digital system used. This was potentially of particular significance in relation to the TCGA dataset, wherein most of the tissue sections are from frozen rather than FFPE tissue. While there are techniques to reduce domain-related biases in algorithm development, such as data augmentation, they cannot be freely used for learning features that describe image quality, due mainly to the fact that many of these techniques change the quality of an image to augment the data, which would be undesirable for a study of inherent WSI quality. For example, transfer functions that change colour saturation, contrast and brightness, add structures or noise, or use filters such as Gaussian blur would not be recommended in this context.

Whilst we therefore have trialed PathProfiler on the cohorts that are different from ProMPT, we do not claim that the results are fully reliable on those cohorts, although the demonstration of overlap of feature space for the different cohorts (ProMPT, TCGA, FOCUS, Figure 4A) provides reassurance of a domain invariant set of features for our model. Within this context, PathProfiler does appear to provide meaningful predicted measures for focus, usability and other artefacts for an overall quality assessment of other cohorts, and predicts 90% of WSIs and 73% of images patches within the TCGA cohort, and 86% of WSIs and 67% of image patches in the FOCUS dataset to be usable according to our criteria (Figures 2A and 2B). Perhaps significantly it has also been revealed through PathProfiler that the 'other' artefact category appears to be responsible for many of the 'unusable' areas on the TCGA cohort (Figure 2A), which relates to the presence of ink. By comparison, the ink was removed prior to the digitisation of the ProMPT cohort which may go some way to account for the differences between cohorts, although this is supposition. It is a consideration for those embarking on retrospective digitisation of glass slide cohorts, although the actual impact downstream on algorithm development that the presence of ink artefact has is unknown. This ink[33] is unlikely to be an issue in prospectively digitised images if these are produced within a routine digital pathology laboratory diagnostic workflow.

Overall, we feel that our data appears to provide a useful indication of quality measures within other cohorts, indicating a broadly similar level of quality between tissue cohorts which may be of relevance to those considering such cohorts in future for computational pathology. PathProfiler provides an indication of overall predicted quality within a cohort of WSI and in a timely and user-friendly manner and is able to classify quality impacting features within the cohort for further investigation.

Comparing this model with other available tools seemed the intuitive next step, and for the purpose of comparing a tool with the ability to assess a range of artefacts we selected HistoQC[23] which uses a set of hand-crafted features for supervised classification. We quickly recognised challenges with this approach, in particular in relation to classification of issues associated with H&E staining quality and tissue folding. We found that these methods were not satisfactory in the classification of the range of artefacts that impacted on image usability, being unable to separate factors that could be remedied by re-scanning or re-staining from those that could not. As a result, our deep learning model overall appears superior in performance, although we recognise that our datasets for testing are relatively small.

Further development of PathProfiler would therefore seem to be potentially worthwhile, and tools such as this will likely become increasingly important. As AI algorithms become available for use within the diagnostic setting, the quality assurance related to their development will likely be scrutinised[10, 34], much in the same way as for other medical devices. It is foreseeable that in the future there will be a need to provide evidence of the quality of cohorts used for the development of AI tools being considered for clinical use, or of a means to evidence that quality-related issues had been considered and accounted for during the study design. Consideration of quality is therefore paramount during algorithm development. Tools that can detect quality issues with WSIs are therefore important/relevant in the collation of image repositories for research purposes and are further justifiable given the significant resources needed to undertake the task of digitising the slides. Ideally, a quality indicator could be provided alongside a WSI library to inform potential researchers of the basic quality of the cohort and the nature of artefacts present. However, in our opinion the development of such a tool capable of analysis across tissue types now requires a concerted effort to collate larger and more varied sample cohorts to enable exposure to the full range of potential variables, ensuring reliability and lack of bias. As such, we have made the code, trained models and generated quality overlays available at https://github.com/MaryamHaghighat/PathProfiler for further studies.

PathProfiler may equally have the future potential for utility within a workflow for diagnostic pathology in the clinical setting and given that the algorithm runs on images at relatively low magnification the impact on throughput time in a clinical pipeline would not be significant. Whilst the benefit for a clinical workflow in highlighting sub-optimal WSIs is evident, there is also a downstream benefit for research, and increasingly so. Retrospective cohorts of cases have been the mainstay of data resources to date for computational pathology, and for research requiring long-term patient outcome data they will remain so, however, the roll-out of digital pathology within the diagnostic setting has opened the doors for prospective collation of whole slide images (WSI) for research purposes. As a result, resources are being ploughed into forward-looking schemes which involve the prospective collation of WSI directly from the diagnostic workflow. The PathLAKE project (https:www.pathlake.org) is an exemplar of such a scheme drawing on WSIs from multiple academic clinical settings to populate a repository that can be utilised for algorithm training, in a pioneering partnership between industry and the NHS. Whilst it is anticipated that these contemporary diagnostic slides would be afflicted with fewer quality issues and we have seen this in our data 12% vs 2% old vs new, artefacts are still recognised and quality assurance is still very relevant.

Whilst large prospective cohorts may be suitable for the development of AI tools, for example to aide cancer detection or assess biomarker status, retrospective cohorts with linked outcome data remain necessary in the development of algorithms to detect novel prognostic tissue features, for example. For this reason, algorithms developed for quality assessment will remain relevant. It is for this reason that we consider the output from our study as particularly significant.

Finally, in addition to the impact on cohort quality, a tool such as PathProfiler that has been developed and validated with specialist pathologist annotated datasets can be time-saving for both pathologists and data scientists. An algorithm to assess diagnostic usability will avoid the need for manual curation of datasets and, in contrast to prior work, our proposed model is reliable in distinguishing artefacts such as out-of-focus regions from glue or ink, and in identifying potentially remediable quality issues. Furthermore, through the identification of ‘unusable’ images at a patch-level, data scientists can decrease their model uncertainty by restricting the input only of ‘usable’ image patches to train their AI algorithms.

4.1 Future work

As part of future work, we will optimise the model to be deployed as part of the data acquisition effort. Our annotated data size is relatively small and does not cover various artefacts seen in different cohorts. We are aiming to improve the tool performance on other cohorts by aggregating more data. This can be done through semi-supervised learning methods and information extracted from diagnostic reports to limit input required from experts. Furthermore, with the help of the community to collect various artefacts in different tissue types and settings, we believe PathProfiler can be extended to a comprehensive and clinically relevant quality assessment tool.

The employed CNN architecture in this work was chosen based on available resources. Further investigation is required to find an optimal network architecture and hyper-parameters. In addition, we trained one model to simultaneously predict the usability of an image, measure the quality of focus and H&E staining and detect folding and other artefacts. At the expense of higher computation time, training separate models for each task may improve predictions.

Conclusions

We have presented the development of an algorithm 'PathProfiler' which is able to reliably predict a clinically relevant quality assessment of WSIs within a DP quality assessment pipeline. The output provision of user-friendly quality evaluation data facilitates the further assessment of areas of sub-optimal image quality in a WSI, with the individual classification of artefacts allowing the user to identify quality issues that are potentially remediable from those that are not. In the context of DP, as advances in AI algorithm development for improvements in patient care become reality, we foresee that quality control of input data will become an essential and scrutinised component of study design. As such, we predict a significant role for a quality tool such as ours in the collation of WSIs from retrospective cohorts of glass slides, and potentially also for prospective collections. PathProfiler has been developed on a cohort of WSIs of historic glass slides from the ProMPT prostate cohort, however, we feel that our work could potentially be enhanced by further training on additional datasets through a community effort, with the anticipation that this could greatly expand the utility of the tool for our academic community.

Compliance with Ethical Standards

Tissue materials were collected in accordance with the ethical standards of the institutional and/or national research committee.

Acknowledgements

This paper is supported by the PathLAKE Centre of Excellence for digital pathology and AI which is funded from the Data to Early Diagnosis and Precision Medicine strand of the government's Industrial Strategy Challenge Fund, managed and delivered by Innovate UK on behalf of UK Research and Innovation (UKRI). PathLAKE funding reference: 104689 / Application number: 18181. The national ProMPT (Prostate cancer Mechanisms of Progression and Treatment) collaborative (grant G0500966/75466) supported sample collections. The University of Oxford sponsors ProMPT.

We acknowledge the contribution to this study made by the Oxford Centre for Histopathology Research and the Oxford Radcliffe Biobank (ORB) which is supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre. Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre.

JR is adjunct professor of the Ludwig Oxford Branch and the Wellcome Centre for Human Genetics. CV, LB and FCH are part funded by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). Funding is via the Molecular Diagnostics Theme. The views expressed are those of the author(s) and not necessarily those of the PathLAKE Consortium members, the NHS, Innovate UK, UKRI, the NIHR or the Department of Health. The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Author contributions

Development of the idea and computational approach (MH, JR), Experiments and data analysis (MH), Code development (MH, KS, SM), Identification of the clinical application (LB, CV, RC), Image scanning (YC), Image exporting (NA, SM), Data annotation (LB, CV, RC), Manuscript drafting (MH, LB), Manuscript revision (all authors).

Competing interests

JR and KS are co-founders of Ground Truth Labs. PathLAKE has received in kind industry investment from Philips. University of Oxford, Oxford University Hospitals NHS Foundation Trust and University of Nottingham are part of the PathLAKE consortium. CV is the principal investigator of a study evaluating Paige Prostate.

Data availability

The datasets generated during and/or analysed during the current study are not publicly available due to the terms of the PathLAKE Consortium Agreement and other agreements in place but a subset of the data could be made available via the corresponding author on reasonable request. The software is open source.

References

1. Ahmad, Z., Rahim, S., Zubair, M. & Abdul-Ghafar, J. Artificial intelligence (AI) in medicine, current applications and future role with special emphasis on its potential and promise in pathology: present and future impact, obstacles including costs and acceptance among pathologists, practical and philosophical considerations. a comprehensive review. *Diagn. Pathol.* **16**, 1–16 (2021).
2. Niazi, M. K. K., Parwani, A. V. & Gurcan, M. N. Digital pathology and artificial intelligence. *The lancet oncology* **20**, e253–e261 (2019).
3. Serag, A. *et al.* Translational AI and deep learning in diagnostic pathology. *Front. medicine* **6**, 185 (2019).
4. van der Laak, J., Litjens, G. & Ciompi, F. Deep learning in histopathology: the path to the clinic. *Nat. medicine* **27**, 775–784 (2021).
5. Bell, J. Life sciences industrial strategy—a report to the government from the life sciences sector. 2017 (2017).
6. Jahn, S. W., Plass, M. & Moinfar, F. Digital pathology: Advantages, limitations and emerging perspectives. *J. Clin. Medicine* **9**, 3697 (2020).
7. Gharzai, L. A. *et al.* Intermediate clinical endpoints for surrogacy in localised prostate cancer: an aggregate meta-analysis. *The Lancet Oncol.* **22**, 402–410 (2021).
8. Cooper, L. A. *et al.* Pancancer insights from the cancer genome atlas: the pathologist's perspective. *The J. pathology* **244**, 512–524 (2018).
9. Kalra, S. *et al.* Pan-cancer diagnostic consensus through searching archival histopathology images using artificial intelligence. *NPJ digital medicine* **3**, 1–15 (2020).
10. Abels, E. *et al.* Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the digital pathology association. *The J. pathology* **249**, 286–294 (2019).
11. Tarroni, G. *et al.* Large-scale quality control of cardiac imaging in population studies: Application to uk biobank. *Sci. reports* **10**, 1–11 (2020).
12. Schömig-Markiefka, B. *et al.* Quality control stress test for deep learning-based diagnostic model in digital pathology. *Mod. Pathol.* 1–11 (2021).
13. Totu, T. *et al.* An objective scoring framework for histology slide image mosaics applicable for the reliable benchmarking of image quality assessment algorithms. *IEEE Access* **6**, 53080–53091, DOI: 10.1109/ACCESS.2018.2868127 (2018).
14. Chen, Y. *et al.* Assessment of a computerized quantitative quality control tool for whole slide images of kidney biopsies. *The J. Pathol.* **253**, 268–278, DOI: https://doi.org/10.1002/path.5590 (2021). https://onlinelibrary.wiley.com/doi/pdf/10.1002/path.5590.
15. Talebi, H. & Milanfar, P. Nima: Neural image assessment. *IEEE Transactions on Image Process.* **27**, 3998–4011 (2018).
16. Stanciu, S. G., Ávila, F. J., Hristu, R. & Bueno, J. M. A study on image quality in polarization-resolved second harmonic generation microscopy. *Sci. reports* **7**, 1–12 (2017).
17. Campanella, G. *et al.* Towards machine learned quality control: A benchmark for sharpness quantification in digital pathology. *Comput. Med. Imaging Graph.* **65**, 142 – 151, DOI: https://doi.org/10.1016/j.compmedimag.2017.09.001 (2018). Advances in Biomedical Image Processing.
18. Wang, Z., Hosseini, M. S., Miles, A., Plataniotis, K. N. & Wang, Z. Focuslitenn: High efficiency focus quality assessment for digital pathology (2020). 2007.06565.
19. Senaras, C., Niazi, M. K. K., Lozanski, G. & Gurcan, M. N. Deepfocus: detection of out-of-focus regions in whole slide digital images using deep learning. *PloS one* **13**, e0205387 (2018).
20. Wu, H. *et al.* Detection of blur artifacts in histopathological whole-slide images of endomyocardial biopsies. In *2015 37th annual international Conference of the IEEE Engineering in Medicine and biology society (EMBC)*, 727–730 (IEEE, 2015).
21. Zhang, T. *et al.* Slidenet: Fast and accurate slide quality assessment based on deep neural networks. In *2018 24th International Conference on Pattern Recognition (ICPR)*, 2314–2319 (IEEE, 2018).
22. Babaie, M. & Tizhoosh, H. R. Deep features for tissue-fold detection in histopathology images. In *Digital Pathology*, 125–132 (Springer International Publishing, Cham, 2019).

23. Janowczyk, A., Zuo, R., Gilmore, H., Feldman, M. & Madabhushi, A. Histoqc: An open-source quality control tool for digital pathology slides. *JCO clinical cancer informatics* **3**, 1–7, DOI: 10.1200/CCI.18.00157 (2019).

24. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241 (Springer, 2015).

25. Zhang, M.-L. & Zhou, Z.-H. A review on multi-label learning algorithms. *IEEE transactions on knowledge data engineering* **26**, 1819–1837 (2013).

26. Lederer, J. Risk bounds for robust deep learning. *arXiv preprint arXiv:2009.06202* (2020).

27. Karimi, D., Dou, H., Warfield, S. K. & Gholipour, A. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Med. Image Analysis* **65**, 101759, DOI: https://doi.org/10.1016/j.media.2020.101759 (2020).

28. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778, DOI: 10.1109/CVPR.2016.90 (2016).

29. Abeshouse, A. *et al.* The molecular taxonomy of primary prostate cancer. *Cell* **163**, 1011–1025 (2015).

30. Seymour, M. T. *et al.* Different strategies of sequential and combination chemotherapy for patients with poor prognosis advanced colorectal cancer (mrc focus): a randomised controlled trial. *The Lancet* **370**, 143–152 (2007).

31. Sirinukunwattana, K. *et al.* Image-based consensus molecular subtype (imcms) classification of colorectal cancer using deep learning. *Gut* **70**, 544–554 (2021).

32. Nayar, S. K. & Nakagawa, Y. Shape from focus. *IEEE Transactions on Pattern analysis machine intelligence* **16**, 824–831 (1994).

33. Ali, S., Alham, N. K., Verrill, C. & Rittscher, J. Ink removal from histopathology whole slide images by combining classification, detection and image generation models. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 928–932 (IEEE, 2019).

34. Colling, R. *et al.* Artificial intelligence in digital pathology: a roadmap to routine use in clinical practice. *The J. pathology* **249**, 143–150 (2019).