Check for updates

Automated curation of large-scale cancer histopathology image datasets using deep learning

Lars Hilgers,[1,2] Narmin Ghaffari Laleh,[1,2] Nicholas P West,[3] Alice Westwood,[3] Katherine J Hewitt,[1,2] Philip Quirke,[3] Heike I Grabsch,[3,4] Zunamys I Carrero,[2] Emylou Matthaei,[2] Chiara M L Loeffler,[2] Titus J Brinker,[5] Tanwei Yuan,[6] Hermann Brenner,[6,7,8] Alexander Brobeil,[9,10] Michael Hoffmeister[6] & Jakob Nikolas Kather[2,3,11]

[1]*Department of Medicine III, University Hospital RWTH Aachen, Aachen,* [2]*Else Kroener Fresenius Center for Digital Health, Medical Faculty Carl Gustav Carus, Technical University Dresden, Dresden, Germany,* [3]*Pathology & Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK,* [4]*Department of Pathology, GROW - Research Institute for Oncology and Reproduction, Maastricht University Medical Center+, Maastricht, The Netherlands,* [5]*Digital Biomarkers for Oncology Group, German Cancer Research Center (DKFZ),* [6]*Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ),* [7]*Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT),* [8]*German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ),* [9]*Institute of Pathology, University Hospital Heidelberg,* [10]*Tissue Bank, National Center for Tumor Diseases (NCT), University Hospital Heidelberg and* [11]*Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany*

Automated curation of large-scale cancer histopathology image datasets using deep learning

Background: Artificial intelligence (AI) has numerous applications in pathology, supporting diagnosis and prognostication in cancer. However, most AI models are trained on highly selected data, typically one tissue slide per patient. In reality, especially for large surgical resection specimens, dozens of slides can be available for each patient. Manually sorting and labelling whole-slide images (WSIs) is a very time-consuming process, hindering the direct application of AI on the collected tissue samples from large cohorts. In this study we addressed this issue by developing a deep-learning (DL)-based method for automatic curation of large pathology datasets with several slides per patient.

Methods: We collected multiple large multicentric datasets of colorectal cancer histopathological slides from the United Kingdom (FOXTROT, $N = 21,384$ slides; CR07, $N = 7985$ slides) and Germany (DACHS, $N = 3606$ slides). These datasets contained multiple types of tissue slides, including bowel resection

Address for correspondence: Jakob Nikolas Kather, Else Kroener Fresenius Center for Digital Health, Technical University Dresden, Fetscherstrasse 74, Dresden 01307, Germany. e-mail: jakob-nikolas.kather@alumni.dkfz.de
Lars Hilgers and Narmin Ghaffari Laleh contributed equally.

Abbreviations: AI, artificial intelligence; AUROC, area under the receiver operating curve; BE, biopsy and endoscopic resection; CI, confidence interval; CNN, convolutional neural network; CRC, colorectal cancer; DL, deep learning; GPU, graphics processing unit; GradCAM, gradient-weighted class activation mapping; H&E, hematoxylin and eosin; IHC, immunohistochemistry; LN, lymph node; MIL, multiple instance learning; NCT, National Center for Tumor Diseases; NT, non-tumor tissue; SSL, self-supervised learning GradCAM - gradient-weighted class activation mapping; T, tumor tissue; TMA, tissue microarrays; WSI, whole slide image.

specimens, endoscopic biopsies, lymph node resections, immunohistochemistry-stained slides, and tissue microarrays. We developed, trained, and tested a deep convolutional neural network model to predict the type of slide from the slide overview (thumbnail) image. The primary statistical endpoint was the macro-averaged area under the receiver operating curve (AUROCs) for detection of the type of slide.

Results: In the primary dataset (FOXTROT), with an AUROC of 0.995 [95% confidence interval [CI]: 0.994–0.996] the algorithm achieved a high classification performance and was able to accurately predict the type of slide from the thumbnail image alone. In the two external test cohorts (CR07, DACHS)

AUROCs of 0.982 [95% CI: 0.979–0.985] and 0.875 [95% CI: 0.864–0.887] were observed, which indicates the generalizability of the trained model on unseen datasets. With a confidence threshold of 0.95, the model reached an accuracy of 94.6% (7331 classified cases) in CR07 and 85.1% (2752 classified cases) for the DACHS cohort.

Conclusion: Our findings show that using the low-resolution thumbnail image is sufficient to accurately classify the type of slide in digital pathology. This can support researchers to make the vast resource of existing pathology archives accessible to modern AI models with only minimal manual annotations.

Introduction

Over the course of time, we have observed a continuous increase in the amount of digital histopathology image data that is readily available. Furthermore, there has also been an exponential growth in the number of new artificial intelligence (AI) approaches using deep learning (DL) in digital histopathology of cancer.[1–4] AI has been applied to numerous tasks based on information that can be extracted from histology slides, including cancer detection,[5,6] predicting the origin in cancer of unknown primary,[7] survival prediction,[8–10] genetic subtyping,[4,11,12] and prediction of treatment response.[13] These methods are valuable research tools, which are also being incorporated into clinical routines as diagnostic algorithms approved by regulatory entities. While a substantial number of published studies have only relied on 100s or 1000s of digitized whole-slide images (WSIs), there are currently large academic and commercial consortia that aim to expedite the digitalization and accessibility of hundreds of thousands of pathological slides.[3,14]

The majority of published studies were carried out on highly selective image collections, where only one WSI is assumed to be representative of the entire patient case. In reality, in many cases the histopathological analysis is not limited to a single slide for a given patient.[15,16] For example, colorectal cancer (CRC) resection specimen cases routinely comprise over 25 slides, and this number can increase when the tumour is large, numerous lymph nodes are identified, or immunohistochemistry (IHC) is required.[17] Although crucial, these slides are usually not labelled

and it is not routinely reported which slides contain which tissue types. As a result, dozens of unlabelled slides are usually available for a single patient. WSIs have been used for a multitude of research applications such as molecular subtyping,[12,18–20] survival prediction,[21,22] response prediction,[23] or to identify risk factors for lymph node metastasis,[24] but a manual selection step by an expert pathologist is usually required to select WSIs that contain the desired tissue type (tumour tissue, normal tissue, lymph node tissue, IHC, etc.) and are of good quality. Previous work has shown that "search and retrieve" approaches can be implemented by extracting visual features from high-resolution tiles generated from WSI.[25] However, this is computationally expensive. Pathologists can often identify the tissue slides without the aid of a microscope, by simply observing a glass slide with the naked eye. For instance, in CRC pathology human experts can easily distinguish tumour slides from lymph nodes or normal mucosa just by looking at the glass slide without any magnification. While it has been shown that DL models can efficiently identify tissue characteristics, such as lymph nodes using low-resolution images,[26] there is still a clear need for automated curation of large histopathological datasets via an algorithm that can efficiently recognize and classify different tissue types to presort large collections of WSIs for subsequent DL applications. It is only via the availability of such systems that advanced AI algorithms may be deployed in a fully automatic way in routine diagnostic workflows.

Therefore, we hypothesized that DL can assist the curation of large collections of WSI at a low

resolution, using only the "thumbnail" images. We developed and validated DL-based models to classify large and unsorted collections of WSIs—in our case, CRC cases—into tissue slide categories. We externally validated the performance of the model in two additional CRC datasets to provide definitive evidence for the generalizability of the model beyond the dataset it was initially trained on. To further investigate the performance of the model, we used relevant explainability methods like gradient-weighted class activation mapping (Grad-CAM) to gain insights into the features and regions of the input images that the model is relying on for its predictions. Additionally, we analysed the misclassified cases from a pathological point of view to define limitations of the model and potential areas for improvement.

Materials and Methods

ETHICS STATEMENT

This study was carried out in accordance with the Declaration of Helsinki. The collection of the tissue samples for the cohorts FOXTROT-CRC and CR07-CRC was granted by the Northern and Yorkshire Research Ethics Committee (Jarrow, UK; Unique Reference Number: 07/MRE03/24).[27,28] The analysis of the second testing cohort DACHS-CRC (an epidemiological study which is led by the German Cancer Research Center, DKFZ, Heidelberg, Germany) was approved by the Ethics Committee of the Medical Faculty, University of Heidelberg under 310/2001.[29–31] The overall analysis was approved by the Ethics Committee of the Medical Faculty of Technical University of Dresden (BO-EK-444102022).

PATIENT COHORTS

In this study we analyzed digital WSIs from three large multicentric patient cohorts (Figure 1A-F). All WSIs were stored in the SVS format. Details and clinicopathological characteristics of all the samples are shown in Table 1. We used the "Fluoropyrimidine, Oxaliplatin, and Targeted Receptor pre-Operative Therapy for colon cancer cohort" (FOXTROT, $N = 1006$ patients, $N = 21,384$ WSIs, Figure S1)[32] cohort as a training set and then used "Medical Research Council CR07" (CR07, $N = 608$ patients, $N = 7985$ WSIs, Figure S2)[28] and "Darmkrebs: Chancen der Verhütung durch Screening" study (DACHS, $N = 2448$ patients, $N = 3606$ WSIs, Figure S3)[29] cohorts as external test sets. All three cohorts represent multicentre, large-scale clinical trials / cohort

studies. Histopathology slides for each cohort were mostly created using routine diagnostic pipelines at each specific trial centre. For the FOXTROT cohort, 90% of slides were created locally and 10% were created centrally at St. James University Hospital in Leeds, UK. For the CR07 cohort, 74% of slides were created locally and 26% were created centrally at St. James University Hospital in Leeds, UK. For the DACHS cohort, all tissue was processed at the individual trial centres. For all cohorts, slides were scanned using Leica Aperio slide scanners. Slides for FOXTROT and CR07 were scanned at St. James University Hospital in Leeds, UK. For DACHS, slides were scanned at the Tissue Bank at the National Center for Tumour Diseases (NCT) in Heidelberg, Germany.

From each WSI, we generated a low-resolution thumbnail image at a fixed resolution of 32 micrometres per pixel using an automated script. Thumbnails were saved in the JPEG format. Low-resolution thumbnail images were then resized to 224×224 pixels with zero padding to preserve proportions. No other preprocessing steps were applied to the images.

CLASSIFICATION OBJECTIVE AND GROUND TRUTH LABELS

We aimed to train a DL model that is capable of classifying colorectal WSIs into seven sample categories including (1) tumour tissue (T), (2) nontumour tissue (NT), (3) lymph node (LN), (4) biopsy and endoscopic resection (BE), (5) fat, (6) IHC, and (7) tissue microarray (TMA) (full descriptions for each category can be found in Table S1). All classes except IHC represented slides stained with haematoxylin and eosin (H&E). During annotation, we removed cases that were too heavily artefacted or didn't fit into any of the defined classes. During model deployment, the class "Undecided" was assigned to a WSI if the prediction score of the classification model was below a defined confidence threshold. Due to the large size of the training cohort (FOXTROT), ground-truth labels were generated in a semisupervised way (Figure 1D). Two observers, a trainee pathologist (K.J.H.) and J.N.K. manually labelled a random subset (41%) of all FOXTROT WSIs. This subset had 8814 WSIs ($N = 5961$ T, $N = 1105$ NT, $N = 296$ LN, $N = 207$ BE, $N = 105$ fat, $N = 1008$ IHC, $N = 132$ TMA). We used this accurate and error-free annotated subset of data to train a very simple convolutional neural network (CNN). The characteristics of this model can be found in the "Implementation and parameters" section. We used this simple classifier to generate noisy labels for the rest of the training cohort (59% of
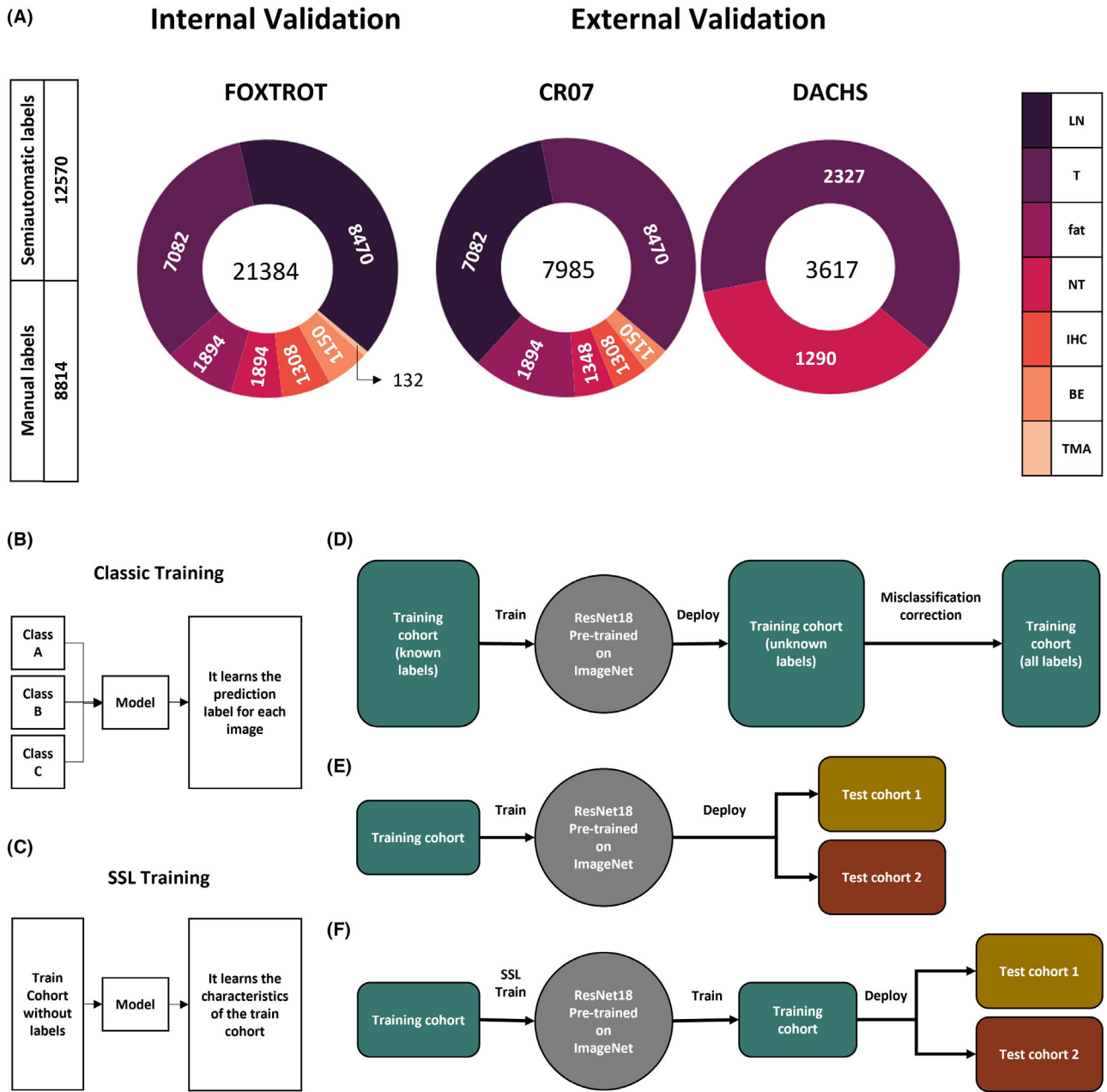
Figure 1. Cohort description and experimental design. (A) Internal and external validation cohorts characteristics. (B) Classic training overview. (C) SSL training overview. (D) Workflow for semiautomatic label generation. (E) Classic training and validation experiment workflow. (F) SSL training and validation experiment workflow. BE, biopsy & endoscopic resection; IHC, immunohistochemistry; LN, lymph node; NT, nontumour tissue; SSL, self-supervised learning; T, tumour tissue; TMA, tissue microarray.

FOXTROT). In this stage, all the noisy labels assigned by DL were manually checked and corrected by the two observers. As a result of this procedure, we were able to more quickly and efficiently annotate 21,384 WSIs in the training cohort (i.e. FOXTROT). Labels for CR07 were manually generated by a trainee pathologist (K.J.H.) for every single image. Available categories in the CR07 cohort are T ($N = 3130$

WSIs), NT ($N = 397$ WSIs), LN ($N = 2800$ WSIs), BE ($N = 268$ WSIs), fat ($N = 1036$ WSIs), and IHC ($N = 354$ WSIs). Labels for DACHS were available in the original study database, where they had been added by pathologists of the National Center for Tumour Diseases (NCT) biobank at the Institute of Pathology of the University of Heidelberg, Germany. For DACHS, only the categories T ($N = 2319$) and

Table 1. Clinicopathological features of all cohorts

	FOXTROT	CR07	DACHS
Origin	United Kingdom	United Kingdom	Southwest Germany
Dataset type	Clinical trial	Clinical trial	Cohort study
Trial centres	85	80	22
Number of patients with tissue available (%)	1006 (95.5)	608 (45.1)	2448 (99.6)
Number of WSIs	21,384	7985	3606
WSI format	SVS	SVS	SVS
Ground truth labels generated by	Partly manually, partly semi-automatic	Full manual annotation	Full manual annotation
Mean age (years) [±SD]	63.0 [±9.7]*	65.1 [±9.2]	68.5 [±10.8]
Female, n (%)	673 (63.9)*	179 (29.4)	1012 (41.3)
Male, n (%)	380 (36.1)*	429 (70.6)	1436 (58.7)

Clinicopathological data were provided by the respective study principal investigators.

n, number of cases; SD, standard deviation; WSI, whole slide image.

*Age and gender data refers to the whole FOXTROT cohort. Subset data were not available for this study.

NT ($N = 1287$) were present in the dataset. A very important aspect of the DACHS cohort is that ~82% of the WSIs contain very clear ink marks. Our purpose in selecting this cohort was to test the sensitivity of the model to possible artefacts on the WSIs.

EXPERIMENTAL DESIGN AND DL TECHNIQUES

We employed two experimental strategies: Strategy #1, supervised learning in which the full training cohort was used (Figure 1E). Strategy #2, supervised learning in which only a very small subset of labelled data from the training cohort were used during training. For this strategy, the number of instances in the training set was limited to 2, 4, 8, 16, 32, and 64 samples per class. For both strategies, we ran two experiments: Experiment #1, an "internal classification experiment" on the FOXTROT cohort only, in which we used threefold crossvalidation to assess within the cohort classification performance. Experiment #2, for which we retrained a classifier on the FOXTROT cohort and then externally tested the

performances in CR07 and DACHS. Finally, we employed two different techniques for each experiment: Technique #1, classical transfer learning, in which a CNN model that was pretrained on the ImageNet database was retrained on the task at hand (Figure 1B). Technique #2, self-supervised pretraining, in which a pretrained CNN was first trained on FOXTROT in a self-supervised way (without labels) and later on retrained in a supervised way (with labels) (Figure 1C,F). Altogether, two strategies for two experiments and two techniques yielded eight separate experimental runs.

IMPLEMENTATION AND PARAMETERS

For the noisy label generating model we trained an ImageNet-pretrained Resnet18 network for a maximum number of 100 epochs with a batch size of 128, a learning rate of 10^{-4}, and a weight decay of 10^{-4} as per the default settings obtained in previous research projects.[33] Early stopping was used with a minimum number of epochs of 30 and a patience value of 10.

Afterwards, for supervised training, we also trained an ImageNet-pretrained Resnet18 network, now using the fully curated labels, with the same parameters mentioned for the noisy label generating model. For self-supervised training we used SimCLR,[34] a method for contrastive self-supervised learning (SSL). We trained this network for 500 epochs with a batch size of 256, a learning rate of 10^{-4}, and a weight decay of 10^{-5}. Augmentations for contrastive SSL were applied to the images, as described previously[35] (Table S2). SSL was conducted using Python's Lightly package to set up the SSL method, including image augmentations, and Python's PyTorch Lightning package for model training. No further hyperparameter tuning was performed.

STATISTICS AND EXPLAINABILITY

The primary endpoint was the area under the receiver operating curve (AUROC). We assumed that an AUROC of above 0.90 would represent a very good classifier. Specifically, we used the macro-averaged AUROCs, where the AUROC for each class is calculated separately and then the average is taken across all classes. The 95% confidence intervals (CIs) of the AUROC values were calculated using the quantiles obtained through 1000-fold bootstrapping with resampling. We also calculated the overall accuracy of the network with fixed classification thresholds of 0.5 and 0.95 applied to the output neurons. These

thresholds resulted in a new category for the classification named 'Undecided'. To provide explainability for the model's decisions, we applied Gradient-weighted Class Activation Mapping (GradCAM)[36] to the images to visualize pixels that were important for classification.

HARDWARE

All experiments were run on a desktop workstation running Windows Server 2019 with 64 GB of RAM and a Nvidia RTX A6000 graphics processing unit (GPU).

CODE AVAILABILITY

All source codes and trained models for DL are open source and available at https://github.com/KatherLab/thumbnail-classification. A script for automated thumbnail generation is available at https://github.com/KatherLab/preprocessing-ng.

Results

DL REACHES HIGH TISSUE CLASSIFICATION PERFORMANCE BASED ON THUMBNAIL IMAGES

First, we assessed the predictability of tissue classes from thumbnail images in a multiclass classification approach. Our baseline approach, internal crossvalidation on the FOXTROT cohort (N = 21,384 slides), with a standard transfer learning approach, yielded a near-perfect AUC of 0.995 [95% CI: 0.994–0.996] (Figure 2A and Figure 2F) and accuracy of 99.8% and 99.9% for predefined classification thresholds of 0.5 and 0.95, respectively. The number of classified cases—meaning cases that were classified by the model with a prediction probability above the chosen threshold—was 21,384 (100% of all cases) for the 0.5 confidence threshold and 21,336 (99.78% of all cases) for the 0.95 confidence threshold. Thus, most cases were still confidently classified by the classifier, even when considerably raising the confidence threshold for classification (Table 2, Table S3).

DL CLASSIFICATIONS GENERALIZE WELL TO EXTERNAL COHORTS

Next, to assess how well the model generalizes, external validation was conducted on independent CRC cohorts, namely, CR07 (N = 7985 slides) and DACHS (N = 3606 slides). The model performance reached accuracies of 91.7% (Figure 2B) and 94.6% (Figure 2C) on the CR07 cohort, and 78.0% (Figure 2D) and 85.1% (Figure 2E) on the DACHS cohort for the predefined classification thresholds of 0.5 and 0.95, respectively. Macro-averaged AUROC for the CR07 cohort was 0.982 [95% CI: 0.979–0.985] (Figure 2F) and 0.875 [95% CI: 0.864–0.887] (Figure 2G) for the DACHS cohort. Similar to the results obtained in the internal crossvalidation experiment, a decrease in the number of classified cases for CR07 was also observed when applying the 0.95 confidence threshold with 7331 (91.81%) classified cases compared to 7948 (99.54%) for the 0.5 classification threshold. For DACHS, the number of classified cases was 2752 (85.1%) for the 0.95 classification threshold and 3560 (98.72%) for the 0.5 classification threshold. Based on these results, the model performs well even when deployed on datasets from different institutions. Remarkably, model performance remained good even on the DACHS cohort, which, as we mentioned earlier, contains heavy ink marks on many of its slides, further showing that our models are quite robust to slide artefacts and other confounders.

DL SHOWS ROBUST RESULTS WHEN TRAINED ON SMALL TRAINING SUBSETS

Furthermore, we investigated the performance of our models on the given task when trained on a comparatively smaller dataset. For this, we trained a supervised model with small subsets of the original dataset for training (2, 4, 8, 16, 32, and 64 samples per class). Overall accuracy remained above 0.5, with accuracies of 58.4% (0.5 threshold; 30.22% cases classified) and 99.4% (0.95 threshold; 0.80% cases classified) for internal validation on the FOXTROT cohort, 58.2% (0.5 threshold; 28.98% cases classified) for external validation on CR07, and 58.8% (0.5

Figure 2. Performance on all datasets. Confusion matrices and receiver operating characteristic (ROC) curves. (A) Interval validation on FOXTROT, threshold 0.5. (B) External validation on CR07, threshold 0.5. (C) External validation on CR07, threshold 0.95. (D) External validation on DACHS, threshold 0.5. (E) External validation on DACHS, threshold 0.95. (F) ROC curves for internal validation on FOXTROT. (G) ROC curves for external validation on CR07. (H) ROC curves for external validation on DACHS. BE, biopsy & endoscopic resection; IHC, immunohistochemistry; LN, lymph node; NT, nontumour tissue; T, tumour tissue; TMA, tissue microarray.
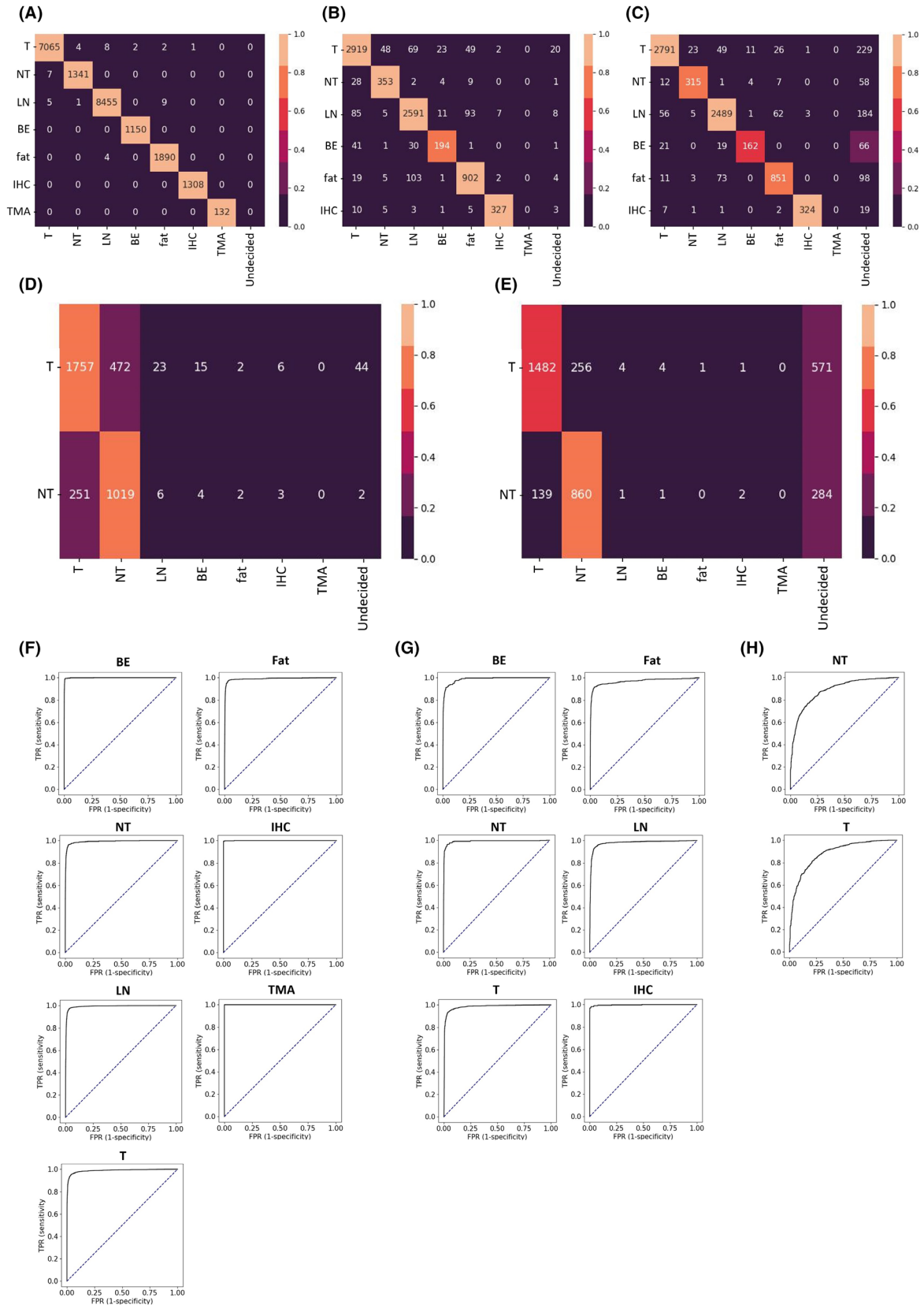
threshold; 10.37% cases classified) for external validation on DACHS, even when the network was given only two cases of each class for training (Table 3). For few-shot experiments with these very small training datasets, the model accuracy for higher thresholds, such as 0.95, fluctuated significantly, since the number of confidently classified cases—with prediction values higher than the set threshold—was very small (one for CR07 and zero for DACHS for the $n = 2$ case). However, model performance continually improved with more training data. When trained on the maximum number of 64 cases per class, performance rose significantly, showing high classification accuracies of 89.5% (96.72% classified cases) and 97.1% (57.75% classified cases) for FOXTROT 87.0% (94.60% classified cases) and 97.2% (44.18% classified cases) for CR07. Accuracy for DACHS also increased to 68.7% (91.18% classified cases) and 85.9% (21.27% classified cases), respectively. Together, these data highlights how models can be trained on relatively small amounts of data while still retaining good classification performances, even when working with low-resolution thumbnails.

SELF-SUPERVISED LEARNING

Additionally, we explored training the model in a self-supervised way with the aim of improving model performance even further. With the inclusion of an SSL step into our workflow, the model showed similar performance to the classic approach. Overall accuracy for testing on the complete datasets was very high, with accuracies of 99.8% (99.99% classified cases) and 99.9% (99.78% classified cases) for FOXTROT, 91.4% (99.70% classified cases) and 94.6% (91.98% classified cases) for CR07, and 79.5% (98.89% classified cases) and 85.5% (78.87% classified cases) for DACHS for both confidence thresholds, respectively (Table 2, Table S4). Similarly, when tested on the few-shot learning task, performance was comparable to the classic approach, with an overall accuracy that remained above 0.5. Furthermore, when the network was given only two cases of each class for training on the given classification task, we were able to observe improved classification results when compared to the classic approach for CR07 and DACHS, with accuracies of 64.5% (34.06% classified cases) and 100% (0.65% classified cases) for FOXTROT, 57.6% (30.77% classified cases) and 100% (only one classified case) for CR07, and 67.7% (24.24% classified cases) and 100% (only one classified case) for DACHS. Once again, when trained on 64 cases per class for the classification task, performance rose to 86.0% (97.54% classified cases) and 94.0% (66.89% classified cases) for FOXTROT, 83.4% (96.46% classified cases) and 92.8% (58.66% classified cases) for CR07, and 75.7% (93.70% classified cases) and 87.8% (42.98% classified cases) for DACHS. In general, the SSL approach showed at least parity to the classic approach for all conducted experiments. All data can be found in Table 4.

IDENTIFICATION OF POSSIBLE REASONS FOR MISCLASSIFIED SAMPLES

In order to gain insight into the misclassified cases, we analysed the cases that were misclassified by our

Table 2. Strong supervision experiments using all the samples

Model pretraining	Test strategy	AUROC [95% CI]	Accuracy %, threshold 0.5	N above Threshold cases for 0.5	Accuracy %, threshold 0.95	N above Threshold cases for 0.95	# total slides
ImageNet	Cross-validation on FOXTROT	0.995 [0.994–0.996]	99.8	21,384	99.9	21,336	21,384
	External test on CR07	0.982 [0.979–0.985]	91.7	7948	94.6	7331	7985
	External test on DACHS	0.875 [0.864–0.887]	78.0	3560	85.1	2752	3617
Pathology SSL	Crossvalidation on FOXTROT	0.999 [0.999–1.000]	99.8	21382	99.9	21,337	21,384
	External test on CR07	0.982 [0.978–0.985]	91.4	7961	94.6	7345	7985
	External test on DACHS	0.871 [0.860–0.883]	79.5	3566	85.5	2844	3617

For each approach—one using a standard pretrained Resnet-18 model and one using a Resnet-18 model that was pretrained on the FOXTROT dataset using a self-supervised learning (SSL) approach—three experiments were conducted. First internal validation on FOXTROT, then external validation on CR07 and DACHS. AUROC and accuracy values are given for each experiment for two different confidence thresholds, alongside the number of cases that were confidently classified by the model for each experiment.

Table 3. Few-shot learning

Test strategy	Number of items per each class	AUROC [95% CI]	Accuracy %, threshold 0.5	N above-threshold cases for 0.5	Accuracy %, threshold 0.95	N above-threshold cases for 0.95
Cross-validation on FOXTROT (number of total slides = 21,384)	2	0.907 [0.904–0.909]	58.4	6462	99.4	172
	4	0.934 [0.932–0.936]	69.8	12,388	92.9	519
	8	0.978 [0.977–0.979]	84.1	16,632	99.0	1367
	16	0.980 [0.979–0.981]	83.2	19,038	96.6	5612
	32	0.988 [0.987–0.989]	85.9	20,343	96.4	9747
	64	0.991 [0.990–0.992]	89.5	20,682	97.1	12350
External test on CR07 (number of total slides = 7985)	2	0.840 [0.832–0.847]	58.2	2314	0	1
	4	0.858 [0.852–0.865]	65.4	4501	75.0	48
	8	0.947 [0.942–0.952]	80.8	5769	98.8	163
	16	0.956 [0.951–0.960]	81.2	6828	96.4	1408
	32	0.966 [0.962–0.969]	82.3	7353	96.5	2492
	64	0.976 [0.972–0.979]	87.0	7554	97.2	3528
External test on DACHS (number of total slides = 3617)	2	0.655 [0.641–0.669]	58.8	374	—	0
	4	0.647 [0.618–0.654]	50.3	1848	100	1
	8	0.689 [0.673–0.705]	65.6	2078	100	17
	16	0.715 [0.699–0.732]	63.3	3070	86.0	444
	32	0.803 [0.789, 0.818]	72.4	3189	89.5	550
	64	0.774 [0.759, 0.790]	68.7	3288	85.9	767

Model pretrained on ImageNet. Few-shot learning experiments were conducted using two cases as a minimum and 64 cases as a maximum number of cases for model training. First internal validation on FOXTROT, then external validation on CR07 and DACHS. AUROC and accuracy values are given for each experiment for two different confidence thresholds, alongside the number of cases that were confidently classified by the model for each experiment.

classic approach model when applied to the CR07 dataset (0.5 confidence threshold; 662 cases; 8.3% misclassification rate). The most prevalent reason for misclassification (31.47% of misclassified cases) was the occurrence of key features of multiple classes within the image, which makes classification into one distinct class difficult. Most cases in this category were slides that contained mainly adipose tissue, but also included small lymph nodes that reasonably could be classified as "fat" as well as "lymph node". Other examples included slides that contained invasive tumours but also prominent lymph nodes. These cases were then often classified as "lymph node" even though a human pathologist would in all cases classify these slides as "tumour tissue", since that is the more clinically relevant class. The second most

prevalent reason was misclassification (27.08%), which describes cases where, after a second review, the ground truth generated by the human pathologist was not correct. In most of these cases the model actually classified the cases correctly, but usually these cases also contained key features of multiple classes, which made finding a clear ground truth difficult. Misleading features accounted for 18.31% of misclassified cases. This category contains any slides with tissue features that apparently were misinterpreted by the model. For example, some cases in this category contained mucosa-associated lymphoid tissue alongside normal intestinal mucosa, which the model apparently interpreted as an invasive tumour, therefore labelling the case "tumour tissue". In other cases, tissue marking dye used to highlight resection

Table 4. Few-shot learning

Test strategy	Number of items per each class	AUROC [95% CI]	Accuracy %, threshold 0.5	N above-threshold cases for 0.5	Accuracy (%), threshold 0.95	N above-threshold cases for 0.95
Crossvalidation On FOXTROT (number of total slides = 21,384)	2	0.890 [0.887, 0.892]	64.5	7284	100	138
	4	0.948 [0.946, 0.949]	63.3	9698	88.2	331
	8	0.971 [0.970, 0.972]	77.7	16,117	95.9	1526
	16	0.982 [0.980, 0.982]	83.1	19,052	95.3	5086
	32	0.969 [0.966, 0.973]	80.6	20,239	94.4	10422
	64	0.991 [0.990, 0.991]	86.0	20,858	94.0	14303
External test on CR07 (number of total slides = 7985)	2	0.799 [0.792, 0.806]	57.6	2457	100	1
	4	0.889 [0.884, 0.894]	59.5	3408	74.3	35
	8	0.943 [0.938, 0.947]	77.0	5648	95.5	290
	16	0.955 [0.951, 0.959]	78.5	6835	93.5	1246
	32	0.969 [0.966, 0.973]	80.6	7441	93.7	3237
	64	0.975 [0.972, 0.978]	83.4	7702	92.8	4684
External test on DACHS (number of total slides = 3617)	2	0.659 [0.645, 0.674]	67.7	874	100	1
	4	0.718 [0.703, 0.735]	63.5	1467	100	3
	8	0.728 [0.712, 0.745]	68.9	2401	94.9	39
	16	0.792 [0.776, 0.807]	72.2	3066	90.7	343
	32	0.865 [0.853, 0.877]	79.6	3490	94.5	1268
	64	0.823 [0.810, 0.836]	75.7	3379	87.8	1550

Model pretrained using pathology SSL. Few-shot learning experiments were conducted using two cases as a minimum and 64 cases as a maximum number of cases for model training. First internal validation on FOXTROT, then external validation on CR07 and DACHS. AUROC and accuracy values are given for each experiment for two different confidence thresholds, alongside the number of cases that were confidently classified by the model for each experiment.

margins was confused for invasive tumour growth. Slide artefacts made up 14.98% of all misclassified cases. These include ink markings, air bubbles, dust, and other contaminants. Staining issues are also included in this category. A small number of misclassified cases (8.17%) was due to tissue structures the model was not familiar with for a certain class. Among others this includes cases where normal intestinal mucosa was cut at an angle, leading to tissue architectures that differ from the norm or tumour cases where the tumour seemed to be surrounded by connective tissue from all sides (Figure 3). In summary, we were able to describe five different types of possible reasons that provide a likely explanation to why each case was misclassified.

To further provide explanations for the model's decisions and also to corroborate which factors contributed to misclassification of images, we decided to explore visual explanations by applying Gradient-weighted Class Activation Mapping (GradCAM) to the images. GradCAM overlays a heatmap onto the original image, highlighting individual pixels and regions that were important for the model's classification decision. For example, size and shape appeared to be important in both the accurate classification and misclassification of biopsy and lymph node images. GradCAM revealed the models' ability to detect cancer cell invasion into subepithelial tissue as well as identifying tissue regions where normal epithelium transitions to invasive cancer. Additionally, GradCAM also highlighted some of the models' weaknesses. For example, subtle colour deviations within the tissue could sometimes lead to misclassification. GradCAM was also able to visualize when the model
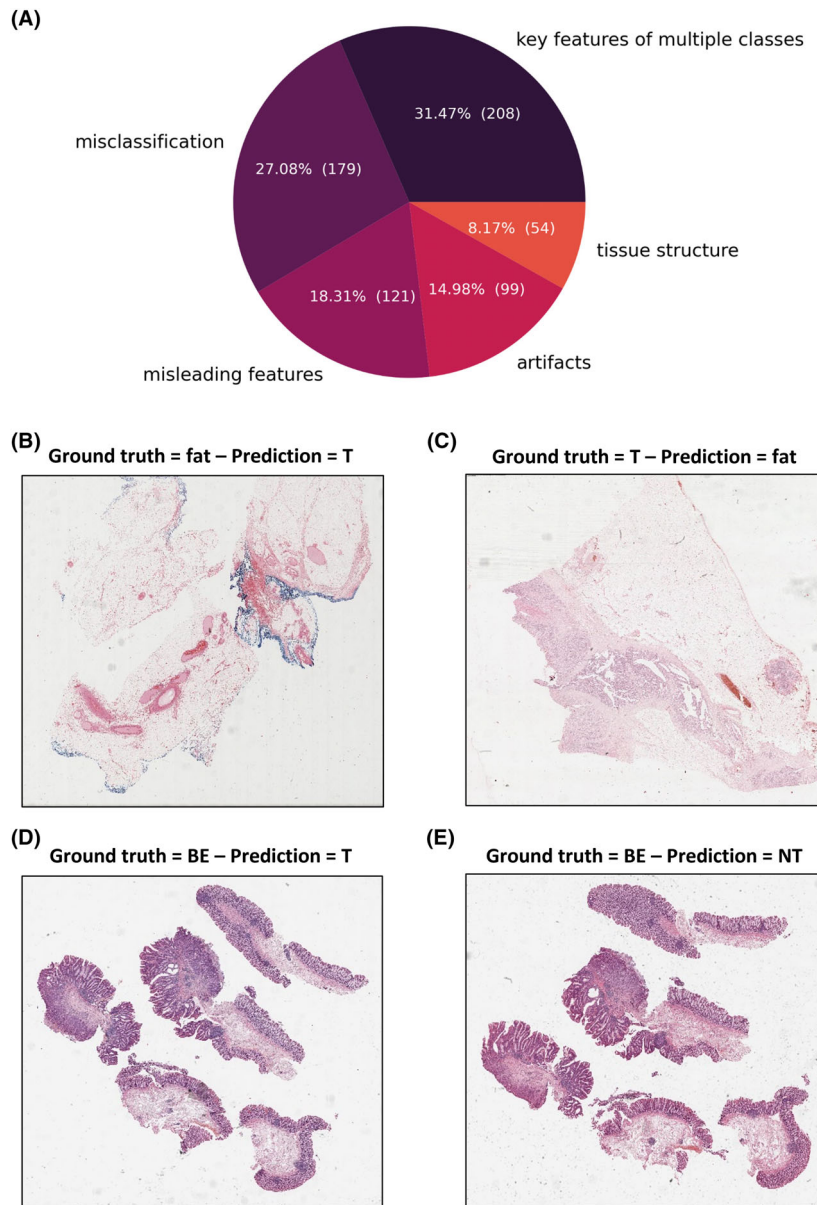
Figure 3. Misclassified cases. (A) Pie chart showing the distribution of reasons for misclassification (standard approach model deployed on CR07. Classification threshold 0.5. Number of misclassified cases $N = 662$). (B) Marking dye confused for invasive tumour cells. (C) Invasive tumour was not detected because of staining issues. (D) Large endoscopic resection tissues and prevalent invasive tumours lead to misclassification. (E) Tissue from the same patient as in (D) but classified differently.

misinterpreted certain tissue features for features from a different class (Figure 4).

In conclusion, this shows that while our models performed very well, there are still limitations to tissue classification. This is particularly true for borderline cases, where even finding a clear ground truth can be difficult, given that tissues are usually very complex in their structure and often may contain characteristics of multiple classes.

Discussion

A ROLE FOR LOW-RESOLUTION "THUMBNAIL" IMAGE ANALYSIS IN COMPUTATIONAL PATHOLOGY

Computational pathology studies almost exclusively address the classification of WSIs at high magnification. Due to the gigapixel size of images at this
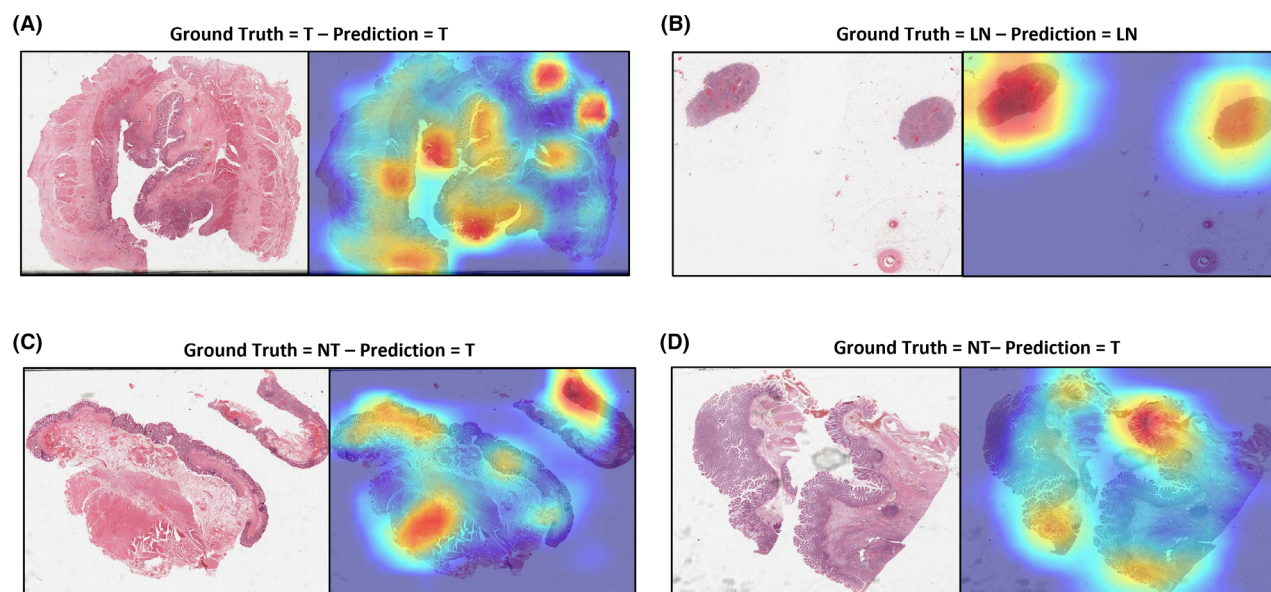
Figure 4. GradCam examples. (A,B) Correctly classified cases in external validation on CR07. (C,D) Misclassified cases in external validation on CR07. LN, lymph node; NT, nontumour tissue; T, tumour tissue.

magnification, complex pipelines are usually used to tessellate a gigapixel image, run predictions on individual tiles, and aggregate predictions, often with multiple instance learning (MIL).[3] Such approaches have also been used for "search and retrieval" problems in computational pathology.[25] However, these intricate workflows are highly computationally expensive, and therefore costly. Furthermore, software pipelines should follow the principle of Occam's razor: they should not be unnecessarily complex if a simple workflow is sufficient. Here, we pursued a much simpler approach: we used low-resolution thumbnail images of whole slides to perform the essential classification task of tissue classification. Surprisingly, even very simple approaches such as the thumbnail-based classification with few-shot learning with simple transfer learning yielded a near-perfect performance. Testing in external cohorts yielded a slightly lower performance, for which a partial remedy was the use of SSL to pretrain the models. Additionally, training and testing our models on datasets from multicentre, large-scale clinical trials, showing good performance across all cohorts, highlights the robustness and flexibility of our models, even when faced with the variability of real-world data.

Thus, we show that clinically relevant image classification tasks can be efficiently solved at very low resolution. Our approach also has significant potential for implementation into clinical workflows. In the busy histopathology department, our pipeline could be used to presort slides by clinical relevance. For example, in resection cases with a large number of slides, this would allow the pathologist to direct their immediate attention to images with higher diagnostic importance, such as tumour.

LOW-RESOLUTION PRESORTING OF SLIDES IN LARGE COHORTS

The relevance of this text revolves around the significant role that our proposed approach could play in enhancing complex computational pathology biomarker studies. Currently, tasks such as molecular subtyping[4] pose a substantial challenge, and researchers often need to manually preprocess and select a single tumour-bearing tissue slide for further investigation. This process is both time-consuming and reduces the quantity of data that can be utilized in these studies. Our approach seeks to automate the preselection process of tumour-bearing tissue slides, thereby significantly expanding the pool of available data for these studies. It serves as a supportive tool to improve the efficiency of complex biomarker research by automating this preliminary task. This improvement becomes critically important when dealing with large clinical trials or clinical routine cohorts, which may contain tens of thousands, if not more, slides that require presorting. Automating this process paves the way for more efficient extraction of computational pathology biomarkers from large datasets,

hence strengthening the potential of biomarker-based clinical research.

LIMITATIONS

Our approach has a number of limitations. For instance, the way we created our ground truth labels was very simplistic. We assumed that every image belongs to exactly one class, but some slides contained two classes, such as a lymph node next to the primary tumour tissue. In such cases, tumour was given priority when assigning a class, as this was deemed the more diagnostically relevant. Additionally, some of this uncertainty can indeed be attributed to the complex architecture of these tissue samples, with a single slide of colon resection containing a plethora of different types of tissues. Together with the variety of tissue compositions found throughout the slides, this presents a challenging factor that ultimately makes it infeasible for some cases to be assigned a single class label. This limitation is aggravated by the fact that our models are incapable of multilabel classification, and will always output a single label for each slide. Furthermore, artefacts, pen markings, and other slide alterations are technical issues that are frequently present in these data formats, potentially limiting our approach. Establishing new practices that can improve these technical and human inaccuracies would inherently lead to even more robust performances of these kinds of tissue classification models. Nonetheless, we were able to show that, even across different cohorts, the results remained consistent, with only a moderate reduction in classification accuracy. To address some of these limitations we introduced a pretraining step using SSL. In all instances, this approach demonstrated parity to the classic approach and in certain cases even slightly improved on the performance of the classic model. Therefore, we would generally recommend the use of an SSL pretrained model for these kinds of classification tasks.

FUTURE WORK

In the present study we have demonstrated that the curation of large datasets can be accomplished through the utilization of thumbnail representations of WSIs and a CNN classifier. However, it should be noted that this approach has only been trained and tested within the context of CRC datasets. Therefore, future research must focus on assessing the model's generalizability to other types of cancers. A key direction for subsequent studies would be to validate this model across diverse cohorts of different cancer and tissue types. By doing so, we can confirm the applicability and robustness of this model beyond CRC, enhancing its ability to be used in various cancer research. Additionally, iterations of this model might be useful for identifying the most representative cases within a cohort in order to make clinical trials more efficient or for multilabel classification of tissue slides. Ultimately, the objective is to evolve this model into a universal tool that can expedite the curation of large datasets across all cancer and tissue types. By achieving this, we can significantly accelerate the processing time, a major bottleneck in medical AI research, and make data more readily available for the broader research community.

Author contributions

LH and NGL performed the experiments. TY, HB, AB, and MH contributed and reviewed data for DACHS. PQ, NW, and AW created the FOxTROT and CR07 scanned slide collections and have reviewed the slides of both trials. All authors contributed to analysing and interpreting the results. LH, NGL, and JNK wrote the first version of the article. All authors edited the article and made the decision to submit this article for publication.

Conflict of interest

JNK declares consulting services for Owkin, France; DoMore Diagnostics, Norway, and Panakeia, UK; furthermore, he holds shares in StratifAI GmbH and has received honoraria for lectures by AstraZeneca, Bayer, Eisai, MSD, BMS, Roche, Pfizer, and Fresenius. TJB is the owner of Smart Health Heidelberg GmbH (Handschuhsheimer Landstr. 9/1, 69120 Heidelberg, Germany, https://smarthealth.de), which develops mobile apps, outside of the submitted work. PQ declares consulting services and research funding from Roche and Amgen. NW declares consulting services for Bristol Myers Squibb, Astellas, Amgen, GSK, and Pfizer as well as research funding from Pierre Fabre. AW has no conflict of interest. No other potential conflicts of interest are declared by any of the authors.

Acknowledgements

Funding information

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

1. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nat. Rev. Clin. Oncol.* 2019; **16**; 703–715.
2. Echle A, Rindtorff NT, Brinker TJ, Luedde T, Pearson AT, Kather JN. Deep learning in cancer pathology: a new generation of clinical biomarkers. *Br. J. Cancer* 2020; **124**; 686–696.
3. Shmatko A, Ghaffari Laleh N, Gerstung M, Kather JN. Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. *Nat. Can.* 2022; **3**; 1026–1038.
4. Cifci D, Foersch S, Kather JN. Artificial intelligence to identify genetic alterations in conventional histopathology. *J. Pathol.* 2022; **257**; 430–444.
5. Kleppe A, Skrede O-J, De Raedt S, Liestøl K, Kerr DJ, Danielsen HE. Designing deep learning studies in cancer diagnostics. *Nat. Rev. Cancer* 2021; **21**; 199–211.
6. Campanella G, Hanna MG, Geneslaw L *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* 2019; **25**; 1301–1309.
7. Lu MY, Chen TY, Williamson DFK *et al.* AI-based pathology predicts origins for cancers of unknown primary. *Nature* 2021; **594**; 106–110.
8. Kather JN, Krisam J, Charoentong P *et al.* Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. *PLoS Med.* 2019; **16**; e1002730.
9. Kleppe A, Skrede O-J, De Raedt S *et al.* A clinical decision support system optimising adjuvant chemotherapy for colorectal cancers by integrating deep learning and pathological staging markers: a development and validation study. *Lancet Oncol.* 2022; **23**; 1221–1232.
10. Saillard C, Schmauch B, Laifa O *et al.* Predicting survival after hepatocellular carcinoma resection using deep-learning on histological slides. *Hepatology* 2020; **72**; 2000–2013.
11. Coudray N, Ocampo PS, Sakellaropoulos T *et al.* Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat. Med.* 2018; **24**; 1559–1567.
12. Kather JN, Pearson AT, Halama N *et al.* Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* 2019; **25**; 1054–1056.
13. Heinz CN, Echle A, Foersch S, Bychkov A, Kather JN. The future of artificial intelligence in digital pathology—results of a survey across stakeholder groups. *Histopathology* 2022; **80**; 1121–1127.
14. Moulin P, Grünberg K, Barale-Thomas E, van der Laak J. IMI-Bigpicture: a central repository for digital pathology. *Toxicol. Pathol.* 2021; **49**; 711–713.

15. Hwang C, Lee SJ, Lee JH *et al.* Stromal tumor-infiltrating lymphocytes evaluated on H&E-stained slides are an independent prognostic factor in epithelial ovarian cancer and ovarian serous carcinoma. *Oncol. Lett.* 2019; **17**; 4557–4565.

16. van Diest PJ, Huisman A, van Ekris J *et al.* Pathology image exchange: the Dutch digital pathology platform for exchange of whole-slide images for efficient teleconsultation, Televerision, and virtual expert panels. *JCO Clin. Cancer Inform.* 2019; **3**; 1–7.

17. G049-Dataset-for-histopathological-reporting-of-colorectal-cancer.pdf.

18. Echle A, Grabsch HI, Quirke P *et al.* Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning. *Gastroenterology* 2020; **159**; 1406–1416.e11.

19. Hildebrand LA, Pierce CJ, Dennis M, Paracha M, Maoz A. Artificial intelligence for histology-based detection of microsatellite instability and prediction of response to immunotherapy in colorectal cancer. *Cancers* 2021; **13**; 391.

20. Yamashita R, Long J, Longacre T *et al.* Deep learning model for the prediction of microsatellite instability in colorectal cancer: A diagnostic study. *Lancet Oncol.* 2021; **22**; 132–141.

21. Bychkov D, Linder N, Turkki R *et al.* Deep learning based tissue analysis predicts outcome in colorectal cancer. *Sci. Rep.* 2018; **8**; 3395.

22. Courtiol P, Maussion C, Moarii M *et al.* Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat. Med.* 2019; **25**; 1519–1525.

23. Harder N, Schönmeyer R, Nekolla K *et al.* Automatic discovery of image-based signatures for ipilimumab response prediction in malignant melanoma. *Sci. Rep.* 2019; **9**; 7449.

24. Brockmoeller S, Echle A, Ghaffari Laleh N *et al.* Deep learning identifies inflamed fat as a risk factor for lymph node metastasis in early colorectal cancer. *J. Pathol.* 2022; **256**; 269–281.

25. Chen C, Lu MY, Williamson DFK, Chen TY, Schaumberg AJ, Mahmood F. Fast and scalable search of whole-slide images via self-supervised deep learning. *Nat. Biomed. Eng.* 2022; **6**; 1420–1434.

26. Beuque M, Magee DR, Chatterjee A *et al.* Automated detection and delineation of lymph nodes in haematoxylin & eosin stained digitized slides. *J. Pathol. Inform.* 2022; **14**: 100192. https://doi.org/10.2139/ssrn.4207480.

27. West NP, Morris EJA, Rotimi O, Cairns A, Finan PJ, Quirke P. Pathology grading of colon cancer surgical resection and its association with survival: a retrospective observational study. *Lancet Oncol.* 2008; **9**; 857–865.

28. Sebag-Montefiore D, Stephens RJ, Steele R *et al.* Preoperative radiotherapy versus selective postoperative chemoradiotherapy in patients with rectal cancer (MRC CR07 and NCIC-CTG C016): a multicentre, randomised trial. *Lancet* 2009; **373**; 811–820.

29. Carr PR, Weigl K, Edelmann D *et al.* Estimation of absolute risk of colorectal cancer based on healthy lifestyle, genetic risk, and colonoscopy status in a population-based study. *Gastroenterology* 2020; **159**; 129–138.e9.

30. Hoffmeister M, Bläker H, Jansen L *et al.* Colonoscopy and reduction of colorectal cancer risk by molecular tumor subtypes: a population-based case–control study. *Am. J. Gastroenterol.* 2020; **115**; 2007–2016.

31. Brenner H, Chang-Claude J, Seiler CM, Stürmer T, Hoffmeister M. Does a negative screening colonoscopy ever need to be repeated? *Gut* 2006; **55**; 1145–1150.

32. Morton D, Seymour M, Magill L *et al.* Preoperative chemotherapy for operable colon cancer: mature results of an international randomized controlled trial. *J. Clin. Oncol.* 2023; **41** (8):1541. JCO2200046.

33. Ghaffari Laleh N, Muti HS, Loeffler CML *et al.* Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. *Med. Image Anal.* 2022; **79**; 102474.

34. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In Iii HD, Singh A eds. *Proceedings of the 37th international conference on machine learning*, 2020, PMLR, 13–18 July, pp. 1597–1607.

35. Stacke K, Unger J, Lundström C, Eilertsen G. Learning representations with contrastive self-supervised learning for histopathology applications. *Journal of Machine Learning for Biomedical Imaging* 2022:023. pp 1–33. http://arxiv.org/abs/2112.05760.

36. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. IEEE International Conference on Computer Vision (ICCV), IEEE2017 https://doi.org/10.1109/iccv.2017.74.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. CONSORT chart for FOXTROT.

Figure S2. CONSORT chart for CR07.

Figure S3. CONSORT chart for DACHS.

Table S1. Class descriptions.

Table S2. Image augmentations used for contrastive learning and their probabilities of application.

Table S3. Precision and recall for the strong supervision experiments using the whole cohorts and classic training.

Table S4. Precision and recall for the strong supervision experiments using the whole cohorts and self-supervised learning.