Medical Image Analysis 92 (2024) 103047

Contents lists available at ScienceDirect

Medical Image Analysis

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

CoNIC Challenge: Pushing the frontiers of nuclear detection, segmentation, classification and counting

Simon Graham [1,2,*,a,b], Quoc Dang Vu [1,2,a,b], Mostafa Jahanifar [1,a], Martin Weigert [3], Uwe Schmidt [4], Wenhua Zhang [5], Jun Zhang [6], Sen Yang [7], Jinxi Xiang [8], Xiyue Wang [9], Josef Lorenz Rumberger [10,11,12], Elias Baumann [13], Peter Hirsch [10,11], Lihao Liu [14], Chenyang Hong [15], Angelica I. Aviles-Rivero [14], Ayushi Jain [16,17], Heeyoung Ahn [18], Yiyu Hong [18], Hussam Azzuni [19], Min Xu [19], Mohammad Yaqub [19], Marie-Claire Blache [20], Benoît Piégu [20], Bertrand Vernay [21,22,23,24], Tim Scherr [25], Moritz Böhland [25], Katharina Löffler [25], Jiachen Li [26], Weiqin Ying [26], Chixin Wang [26], David Snead [2,27,28,a], Shan E. Ahmed Raza [1,a], Fayyaz Minhas [1,a], Nasir M. Rajpoot [1,2,27,a], The CoNIC Challenge Consortium [c]

[1] *Tissue Image Analytics Centre, University of Warwick, Coventry, United Kingdom*
[2] *Histofy Ltd, Birmingham, United Kingdom*
[3] *Institute of Bioengineering, School of Life Sciences, EPFL, Lausanne, Switzerland*
[4] *Independent Researcher, Dresden, Germany*
[5] *The Department of Computer Science, The University of Hong Kong, Hong Kong*
[6] *Tencent AI Lab, Shenzhen, China*
[7] *College of Biomedical Engineering, Sichuan University, Chengdu, China*
[8] *Department of Precision Instruments, Tsinghua University, Beijing, China*
[9] *College of Computer Science, Sichuan University, Chengdu, China*
[10] *Max-Delbrueck-Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany*
[11] *Humboldt University of Berlin, Faculty of Mathematics and Natural Sciences, Berlin, Germany*
[12] *Charité University Medicine, Berlin, Germany*
[13] *University of Bern, Bern, Switzerland*
[14] *Department of Applied Mathematics and Theoretical Physics, University of Cambridge, United Kingdom*
[15] *Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong*
[16] *Softsensor.ai, Bridgewater, NJ, United States of America*
[17] *PRR.ai, TX, United States of America*
[18] *Department of R&D Center, Arontier Co. Ltd, Seoul, Republic of Korea*
[19] *Computer Vision Department, Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates*
[20] *CNRS, IFCE, INRAE, Université de Tours, PRC, 3780, Nouzilly, France*
[21] *Institut de Génétique et de Biologie Moléculaire et Cellulaire, Illkirch, France*
[22] *Centre National de la Recherche Scientifique, UMR7104, Illkirch, France*
[23] *Institut National de la Santé et de la Recherche Médicale, INSERM, U1258, Illkirch, France*
[24] *Université de Strasbourg, Strasbourg, France*
[25] *Institute for Automation and Applied Informatics Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany*
[26] *School of software engineering, South China University of Technology, Guangzhou, China*
[27] *Department of Pathology, University Hospitals Coventry and Warwickshire NHS Trust, Coventry, United Kingdom*
[28] *Division of Biomedical Sciences, Warwick Medical School, University of Warwick, Coventry, United Kingdom*

ARTICLE INFO

Dataset link: https://conic-challenge.grand-challenge.org

Keywords:

ABSTRACT

Nuclear detection, segmentation and morphometric profiling are essential in helping us further understand the relationship between histology and patient outcome. To drive innovation in this area, we setup a community-wide challenge using the largest available dataset of its kind to assess nuclear segmentation and cellular

* Correspondence to: Department of Computer Science, University of Warwick, United Kingdom.
E-mail address: simon.graham@warwick.ac.uk (S. Graham).
[a] Challenge organisers.
[b] First authors contributed equally.
[c] See Appendix for more details.

https://doi.org/10.1016/j.media.2023.103047
Received 15 May 2023; Received in revised form 19 September 2023; Accepted 29 November 2023
Available online 13 December 2023

Computational pathology
Nuclear recognition
Deep learning

composition. Our challenge, named CoNIC, stimulated the development of reproducible algorithms for cellular recognition with real-time result inspection on public leaderboards. We conducted an extensive post-challenge analysis based on the top-performing models using 1,658 whole-slide images of colon tissue. With around 700 million detected nuclei per model, associated features were used for dysplasia grading and survival analysis, where we demonstrated that the challenge's improvement over the previous state-of-the-art led to significant boosts in downstream performance. Our findings also suggest that eosinophils and neutrophils play an important role in the tumour microevironment. We release challenge models and WSI-level results to foster the development of further methods for biomarker discovery.

1. Introduction

Analysis of nuclei in a histopathology tissue slide can provide key information for identifying the presence or state of a disease. For example, their shape and appearance can be used to determine cancer grade, whereas the co-occurrence and distribution of different nuclei can be indicative of diagnosis and patient outcome. In particular, epithelial nuclear pleomorphism is a major component of the Nottingham Grading System for breast cancer (Rakha et al., 2008), while increased amounts of immune cells may be a sign of certain conditions, such as inflammatory bowel disease (Lennard-Jones, 1989; Magro et al., 2013) Two particularly well-studied prognostic tissue-based biomarkers are Tumour-Infiltrating Lymphocytes (TILs) (Ropponen et al., 1997; Salgado et al., 2015) and Cancer-Associated Fibroblasts (CAFs) (Sahai et al., 2020; Tommelein et al., 2015) Here, TILs have been linked to positive patient outcome and immunotherapy response, while CAFs are generally associated with poor outcome due to their role in promoting tumour development.

To assist with nuclear analysis, Deep Learning (DL) methods like HoVer-Net (Graham et al., 2019) and StarDist (Schmidt et al., 2018) have been used to automate nucleus recognition. However, these models require a large amount of labelled data to perform accurately. Obtaining pixel-level annotations is a time-consuming task that requires pathologist input, often leading to small datasets. To overcome this, recent semi-automatic methods involving pathologists have led to the collection of large datasets for nuclear segmentation. For example, PanNuke (Gamper et al., 2019, 2020) collected point annotations of over 200,000 different nuclei with collaborating pathologists, and utilised a semi-automatic method for generating the nuclear boundaries (Koohbanani et al., 2020). The Lizard dataset (Graham et al., 2021) employed an iterative approach to label nearly half a million nuclei, using a combination of semi-automatic and manual pathologist-involved refinement steps to ensure accurate annotation. Datasets at such scales pave the way for the development and reliable evaluation of advanced DL models for nuclear recognition.

AI competitions have been pivotal in helping to drive forward the development of innovative DL models in Computational Pathology (CPath) (Bejnordi et al., 2017; Bulten et al., 2022; Sirinukunwattana et al., 2017; Da et al., 2022), where carefully curated datasets are made available to participants around the world. However, even though there have been several previous competitions for automatic identification of nuclei in H&E images (Kumar et al., 2019; Vu et al., 2019), all tend to suffer from a similar set of limitations. For example, the previous largest competition for nuclear segmentation and classification (Verma et al., 2020) used a dataset consisting of around 47 thousand nuclei, where only 15 thousand of these were used for evaluation. Furthermore, the evaluation images were available to participants, meaning that models could be tuned until a satisfactory visual performance was observed. Of course, this is not reflective of clinical practice and may ultimately lead to overfitting. Instead, it is desirable for images to be hidden from participants during evaluation to ensure reliable assessment of model performance and to minimise the risk of test data hacking. Despite competitions being a great way of accelerating research for AI-based nuclear recognition, the ultimate aim is to enable the extraction of interpretable biomarkers and use them in downstream clinical tasks, such as cancer grading (Shaban et al., 2020), finding origins for cancers of unknown primary (CUP) (Lu et al., 2021) or improved patient stratification (Wulczyn et al., 2020; Zhu et al., 2020; Kather et al., 2019). However, no previous AI competition for nuclear identification has performed an analysis on how the performance of submitted algorithms impacts downstream applications. We consider this to be particularly important because up until now, there has been limited understanding into what level of performance is required for automatic nuclear identification.

To counter the above limitations, we organised the Colon Nuclei Identification and Counting (CoNIC) Challenge that invited participants from around the world to develop solutions aimed at solving the following two tasks: (1) nuclear segmentation and classification and (2) prediction of cellular composition. The CoNIC Challenge uses an extension of the current largest available dataset for nuclear instance segmentation and classification, consisting of over 535,000 unique nuclei from 16 centres, which is over 11 times the number of nuclei used in the previous largest challenge (Verma et al., 2020). In addition to using a large dataset to ensure reliable evaluation, we also required participants to submit their algorithms, rather than the results, enabling test images to remain hidden and guaranteeing unbiased evaluation. Results were announced at the International Symposium for Biomedical Imaging (ISBI) 2022 in Kolkata.

Furthermore, we performed an extensive assessment of the top algorithms on two clinical tasks: dysplasia grading and survival analysis. For this, we processed 1,658 WSIs from two independent colorectal cohorts with the best performing models and assess the impact of nuclear recognition performance on each downstream application. We identified that the best methods from the challenge are capable of achieving superior performance as compared to previous methods. Nevertheless, our findings also suggest that there exist differences in the most important features identified using each of the state-of-the-art model predictions. As a result, subsequent interpretation of these features should be done with caution. To encourage the development of further downstream methods using nuclear features, we also make these WSI-level results, along with the top-performing algorithms, publicly available.

Taking all of this into account, we believe that the CoNIC Challenge will be pivotal in stimulating the development of interpretable cell-based AI models for CPath. The challenge website can be accessed at https://conic-challenge.grand-challenge.org/.

2. Methods

2.1. CoNIC challenge

To stimulate the development of automatic models for nuclear recognition, we organised an AI competition that invited researchers to develop solutions for two tasks: (1) nuclear segmentation and classification and (2) cellular composition. For this purpose, we extended our recent Lizard dataset (Graham et al., 2021) so that it now contains 535,063 labelled nuclei, making this ten times the size of the previous largest AI competition for automatic nuclear recognition in CPath (Verma et al., 2020). Specifically, participants were required to either segment or predict the counts of the following types of nuclei: epithelial, plasma, lymphocyte, neutrophil, eosinophil and connective tissue. Here, we use connective tissue as a broader category consisting of fibroblasts, muscle and endothelial cells.

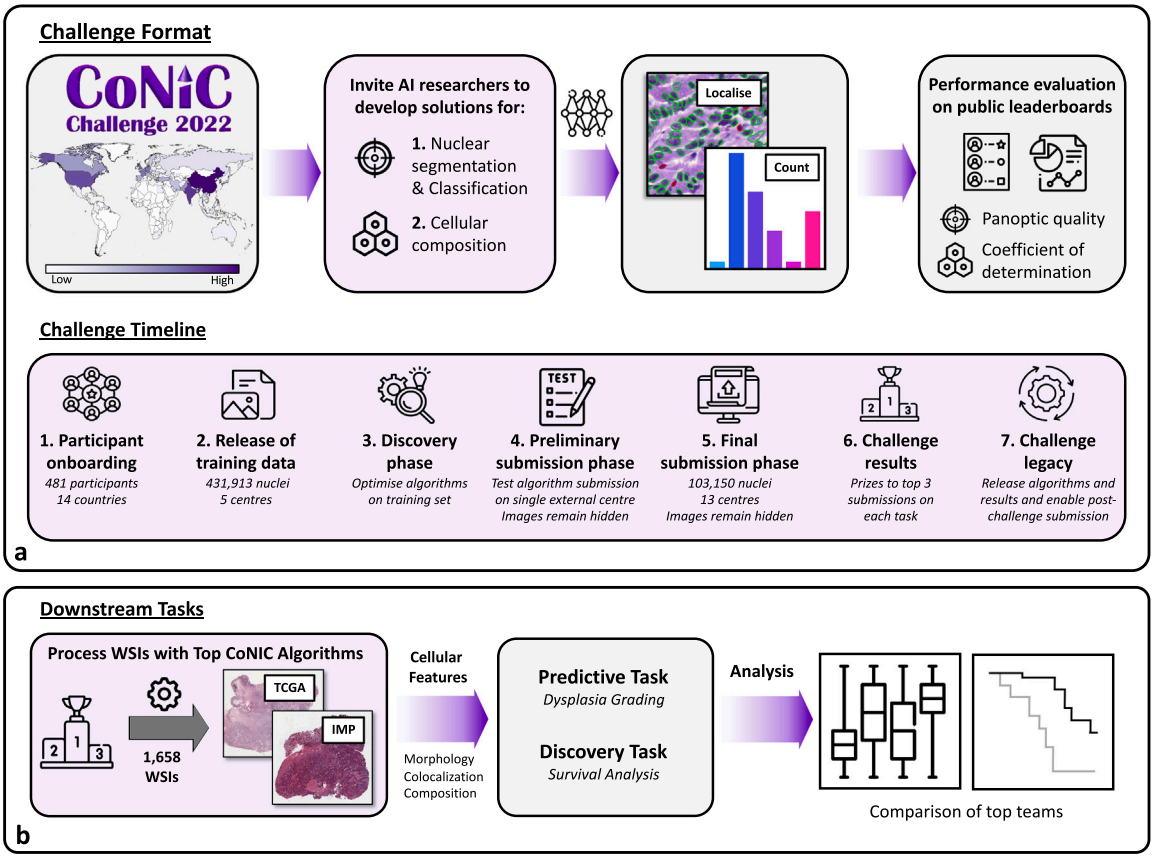

Fig. 1. Overview of the CoNIC Challenge. **a**, Challenge format and timeline. The format describes the main aims of the challenge, which involves developing AI models for (1) automatic nuclear segmentation and classification and (2) cellular composition. The challenge timeline describes the major events during the competition. **b**, Application of the best performing models from the challenge on downstream tasks. We take the best models from the challenge and assess their performance on the tasks of dysplasia grading and survival analysis.

In Fig. 1a we give an overview of the CoNIC Challenge, including the timeline with the following major events: (1) participant onboarding, (2) release of training data, (3) discovery phase, (4) preliminary submission phase, (5) final submission phase, (6) challenge results, and (7) challenge legacy. The competition was hosted on Grand Challenge (https://grand-challenge.org), which enabled seamless participant registration and provided a platform for algorithm submission. Training data was released at the beginning of the competition, where we extracted small image regions (patches) of size 256 × 256 from the original Lizard training set and made them available to download.

We provided a HoVer-Net model (Graham et al., 2019), which was optimised on the training dataset, as the baseline for the competition. Aside from this, we did not permit the organisers to make any submissions to the challenge. There was a two-week preliminary submission phase to allow participants to familiarise themselves with the submission system and improve their algorithm generalisation. To assist with the latter, we utilised a small sample of the full evaluation dataset, which came from a single TCGA centre with images that looked noticeably different from the training data. The final submission phase lasted one week, where participants could only submit once per task. Here, algorithms had to process all 103,150 nuclei in the evaluation dataset within 60 min, where images came from the Lizard test dataset and an additional colon biopsy dataset made purposely for the competition. The results were kept secret until they were revealed at the challenge workshop.

In total, we received 373 submissions during the challenge, where 208 were for the segmentation and classification task and 165 for the cellular composition task. 26 unique teams appeared on the final segmentation and classification leaderboard, whereas 24 teams appeared on the cellular composition leaderboard. Upon conclusion of the challenge, we have kept the portal open for submissions and release the top algorithms to facilitate future developments in the field.

2.1.1. Dataset

To ensure reliable evaluation and to foster the development of generalisable models, it is essential for AI competitions to utilise large datasets. Therefore, in this competition, we used data from our recently curated Lizard dataset (Graham et al., 2021), consisting of 495,179 nuclei in H&E-stained microscopic image regions from 16 different centres and three countries. To gather such a large dataset, we employed an iterative approach to annotate the data, which used a combination of semi-automatic and manual refinement steps with significant pathologist involvement. Utilising this strategy ensured the development of an accurate dataset at scale, resulting in the largest available dataset for nuclear segmentation and classification in CPath. Nuclei were labelled in accordance with their associated cell type and categorised as either: epithelial cell, lymphocyte, plasma cell, neutrophil, eosinophil or connective tissue cell. Here the connective tissue cell category groups endothelial cells, fibroblasts and muscle cells into a single class. Therefore, models trained on this data may be used to help effectively profile the colonic tumour micro-environment. We choose to focus on nuclei from colon tissue to ensure that our dataset contains images from a wide variety of different normal, inflammatory, dysplastic and cancerous conditions in the colon - therefore increasing the likelihood of generalisation to unseen examples.

In addition to the Lizard dataset (Graham et al., 2021), we labelled 39,884 nuclei from an internal colon biopsy dataset. This was done by multiple pathologists in the form of point annotations with consensus review (Wahab et al., 2022). Then, segmentation masks were produced using a semi-automatic method (Koohbanani et al., 2020) and results

were manually refined. This led to a total challenge dataset of 535,063 accurately labelled nuclei. Since the data was not completely manually annotated, some noise in the dataset is inevitable. To investigate this further, we assessed the generated annotation accuracy by comparing it with annotations performed by multiple pathologists (Graham et al., 2021). We found that the level of error was acceptable. For the purpose of the competition, we then extracted patches of size 256 × 256 pixels at 20× objective magnification (approximately 0.5 microns/pixel) and provided the counts within the central 224 × 224 pixel region. This ensured that nuclei were only considered if the majority of pixels were visible. We provide a detailed summary of the breakdown of the dataset in Fig. S1. Only the challenge data provided was allowed for model training. We did not permit the use of any external data.

2.1.2. Evaluation metrics

For each task, we utilised a single metric that was used to rank team submissions. For the segmentation and classification task, the multi-class panoptic quality was used, which has recently been justified as a strong metric by Graham et al. (2019). Here, for each type t, the PQ is defined as:

$$\mathcal{PQ}_t = \underbrace{\frac{|TP_t|}{|TP_t| + \frac{1}{2}|FP_t| + \frac{1}{2}|FN_t|}}_{\text{Detection Quality (DQ)}} \times \underbrace{\frac{\sum_{(x_t,y_t)\in TP} IoU(x_t, y_t)}{|TP_t|}}_{\text{Segmentation Quality (SQ)}} \tag{1}$$

where x denotes a ground truth (GT) instance, y denotes a predicted instance, and IoU denotes intersection over union. Setting $IoU > 0.5$ will uniquely match x and y. This unique matching therefore splits all available instances of type t within the dataset into matched pairs (TP), unmatched GT instances (FN) and unmatched predicted instances (FP). Henceforth, we define the multi-class PQ (mPQ) as the task ranking metric, which takes averages the PQ over all classes. Note, for mPQ we calculate the statistics over all images to ensure there are no issues when a particular class is not present in a patch. This is different to mPQ calculation used in previous publications, such as PanNuke (Gamper et al., 2020), MoNuSAC (Verma et al., 2020) and in the original Lizard paper (Graham et al., 2021), where the PQ is calculated for each image and for each class before the average is taken. Hence, for the purpose of this challenge, we refer to the metric as mPQ^+. As an added benefit, PQ can be easily decomposed into Detection Quality (DQ) and Segmentation Quality (SQ), enabling a more detailed analysis of participants' results. Despite these results not being utilised in the main challenge, we display a summary of the obtained mDQ^+ and mSQ^+ scores obtained for each team in Fig. S4.

For the cellular composition task, we used the multi-class coefficient of determination to determine the correlation between the predicted and true counts. Similar to the previously described metric, the statistic is calculated for each class independently and then the results are averaged. In particular, for each nuclear category t, the correlation of determination is defined as follows:

$$R_t^2 = 1 - \frac{RSS_t}{TSS_t} \tag{2}$$

where RSS stands for the sum of squares of residuals and TSS stands for the total sum of squares after a regression line is fitted to the predicted and actual counts. For additional analysis, we also display additional regression results, using Mean Absolute Error (MAE) and Mean Arctangent Absolute Percentage Error ($MAAPE$), in Fig. S4. As before, both metrics compute the statistics for each class independently and the results are then averaged to give the final score. Unlike other metrics described throughout this paper, low scores for MAE and $MAAPE$ indicate a strong performance.

2.1.3. Submission pipeline

The challenge began on the 20th November 2021, when we released the training data, so that participants could start developing solutions for the two tasks. We also released evaluation code and a HoVer-Net (Graham et al., 2019) baseline model on our challenge GitHub page to help accelerate model development and prevent participants needing to build models from scratch (https://github.com/TissueImageAnalytics/CoNIC). During this time, participants were able to ask questions on the web page forum to further understand the intricacies of the tasks and the baseline model.

In the meantime, we implemented an evaluation framework that enabled participants to submit their algorithms to the competition, allowing us to keep test images hidden and ensuring unbiased evaluation. For this, we utilised the Grand Challenge platform (https://grand-challenge.org) developed by the Diagnostic Image Analysis Group at Radboud University Medical Center. Participants were required to submit their algorithms to the portal as Docker containers, which were then used to process the test images in the cloud using Amazon Web Services (AWS). To help with this, we created detailed videos on the challenge web page (https://conic-challenge.grand-challenge.org/), providing step-by-step instructions on how to create, test and submit the containers to the portal. To facilitate this, we provided a Docker template, along with a specific example using our baseline, that participants could easily adapt for their own solutions. Adhering to this template ensured that containers were able to appropriately read image data in the backend, process the data using their developed algorithms and return outputs easily recognisable in the next step of the evaluation protocol.

For evaluation, we developed a container that took the algorithm outputs and computed the metrics of each submission. These metrics then interacted with the Grand Challenge platform, where the overall results were then displayed on public leaderboards. This overall submission procedure could be tested during the preliminary submission period, that took place between 13th–27th February 2022. At this stage, the organisers were in regular contact with competing teams, advising them on potential reasons for failed submissions. Therefore, not only did this encourage the improvement in the performance of developed models due to its competitive nature, but it also prepared participants for making their final submissions. Teams were allowed one submission per day for each task, leading to many results being displayed on the leaderboards. For each task, a separate submission was required, even if the segmentation output was used to predict cellular composition. However, it was not mandatory to complete both tasks and participants could focus on just one, such as predicting cellular composition. This has recently been done by Dawood et al. (2021) where the counts of different cell types were predicted without explicitly localising each nucleus. The final stage was between 27th February–6th March 2022, where only one successful submission was permitted for each team per task. We allowed a maximum of 60 min to process the full test set, to prevent excessive ensembling. Prizes were awarded to the top three positions of each task and the winners also received an NVIDIA RTX 3070 GPU.

Upon conclusion of the challenge, we invited the top ten teams to send us two Docker containers: (1) original submission and (2) models trained on a specified split of the data, without ensembling. Original algorithms were requested so that we could make them available to the public (for those that granted us permission), whereas retrained models on a single split enabled us to perform a fairer head-to-head comparison between methods. Gathering models that did not use ensembling also unlocked the potential to use them for WSI processing and downstream analysis, due to reasonable inference times. Docker containers are available for download by visiting https://warwick.ac.uk/conic-challenge. Despite the conclusion of the challenge, participants can still submit algorithms and visualise their results on post-challenge leaderboards using the same Docker-based submission protocol as outlined above.

2.2. Post challenge analysis

2.2.1. Clinical datasets

To assess how differences in the results of nuclear detection across various teams impacts the performance of downstream tasks, we collected data containing Haematoxylin and Eosin (H&E) stained WSIs of colorectal tissue from The Cancer Genome Atlas (TCGA) and IMP Diagnostics Laboratory. Here, the TCGA dataset was used to perform survival analysis, whereas the IMP Diagnostics dataset was used for dysplasia grading. TCGA slides were digitised at various institutions and therefore the scan resolution of the original WSIs varies. Slides from IMP Diagnostics were digitised with a Leica GT450 scanner at a pixel resolution of 0.263 microns/pixel. In total, we obtained 526 WSIs of surgical resections from TCGA and 1132 WSIs of endoscopic biopsies from IMP Diagnostics.

To enable survival analysis on TCGA, we extracted the disease specific and overall survival times as well as the respective survival statuses of each patient. Here, overall survival is the time from the initial diagnosis of the disease (in this case, colorectal cancer) until the death of the patient from any cause. On the other hand, disease specific survival is the time to death specifically as a result of colorectal cancer. Within the IMP Diagnostics dataset, each slide is categorised into one of the following groups: non-neoplastic, low-grade lesion or high-grade lesion. Here, non-neoplastic slides contain both normal and inflammatory conditions, low-grade lesions contain conventional adenomas with low-grade dysplasia, and high-grade lesions contain conventional adenomas with high-grade dysplasia, intra-mucosal carcinomas and invasive adenocarcinomas. To reliably evaluate the performance of each downstream task, on both datasets we performed five-fold cross-validation. Each fold was separated into a training (60%), validation (20%) and test set (20%). We repeated this procedure five times with different random seeds, resulting in a total of 25 different splits of the data.

2.2.2. Evaluation metrics

For dysplasia grading, we measured the F_1 and Average Precision (AP) score per category and then calculated the average result, denoted by mF_1 and mAP. In addition, we computed the Quadratic Weighted Kappa (QWK), which measures the agreement between the predictions and true diagnostic categories. For survival analysis, we measured and reported the concordance index (C-Index) between the predicted risk scores and actual events.

2.2.3. Digital features

For both cohorts, we processed slides using the top-performing models from the segmentation and classification task (EPFL | StarDist, MDC Berlin | IFP Bern and Pathology AI) that were trained on a single split of the challenge dataset. For each WSI, we then extracted 222 patient-level features that could be grouped into the following categories: morphological, colocalisation, and density features. Here, morphological features included the best alignment metric (BAM) (Awan et al., 2017), size, eccentricity, major axis, minor axis and perimeter of each nucleus based on its predicted contour. Colocalisation features describe the spatial relationship between different types within several predefined neighbourhoods (200 μm and 400 μm radius (Berry et al., 2021; Vu et al., 2023) For morphological and colocalisation features, we first calculated these statistics per nucleus and then computed the mean and standard deviation across all WSIs belonging to each patient to give the corresponding patient-level features. Density features describe the global ratio of different nucleus types across all tissue samples of a patient. Depending on the settings of subsequent experiments, a patient-level digital descriptor for a WSI can either contain only morphological, colocalisation, density features, or contain a combination of them all. For clarity, we respectively denoted these digital feature sets as D_m, D_c, D_d and D. A comprehensive description of these features is provided in the supplementary material.

2.2.4. Predictors for downstream tasks

In order to identify how the nuclear segmentation results for each team affect downstream tasks, it is important for us to not only use a model with a strong predictive power, but also use one that allows interpretation of which input features are important. With this in mind, we utilised a tree-based method, named gradient boosted trees (implemented with XGBoost), for both grading and survival analysis tasks. Compared to other tree-based implementations, XGBoost is well-known for being computationally efficient and scalable, while still ensuring a strong performance (Chen and Guestrin, 2016). Throughout the paper, we used Random Search on the XGBoost parameter space, with 2048 sampled points, to obtain the best parameters for each set of input features that were obtained from each team and for each of the downstream tasks. The XGBoost parameter sets that have the best validation results across all folds and repetitions are selected for subsequent feature interpretation and analyses on their corresponding testing set.

2.2.5. Feature selection

As commonly described in other works (Kira and Rendell, 1992; Kursa and Rudnicki, 2010) not all input features will necessarily contribute to the considered downstream tasks. Thus, we perform feature selection on D to find the most predictive feature set $\bar{D}$ for the final models. To identify these features, we first performed the previously described random parameter search, but rather than selecting the model that performs best across all folds and repetitions, we selected the best model on each fold, resulting in 25 different models. For each of these 25 models, we examined the impact of the input feature set D on the results (QWK for grading and C-index for survival analysis) using the Permutation Test (Altmann et al., 2010), which gives an importance value for each feature. Then, we averaged the feature importance values across all folds to obtain the overall importance of the features in D. Finally, we selected features whose importance scores were greater than the median value across the 222 features. A rough summary of which features were selected from each team and for each task is reported in Fig. S7.

2.2.6. Dysplasia grading

For this task, we extracted D_m, D_c, D_d, D and $\bar{D}$ from the IMP Diagnostics dataset, denoting the various feature sets, as outlined above. Then, using the previously described procedure, we fit XGBoost models on each feature set extracted from the nuclear segmentation results of each team and performed a comprehensive comparative analysis.

2.2.7. Survival analysis

For this task, as well as comparing the performance of each digital feature set, we also assess how the final selected set of digital features $\bar{D}$ compares with existing clinical features for predicting disease specific and overall survival. In this work, we utilised sex, age and cancer stage, denoted by C, as the set of clinical features. Similar to the grading pipeline, we extracted D_m, D_c, D_d, D and $\bar{D}$ feature sets from the TCGA dataset for utilisation in the downstream survival analysis pipeline. To perform survival analysis, we evaluated the predicted risk scores obtained from fitting XGBoost models on the different feature sets from each team and reported their C-Index on the validation and testing sets.

3. Results

3.1. Preliminary test phase results

As well as preparing participants for their final submissions, the preliminary submission phase provided an opportunity to assess and improve the performance of algorithms developed during the discovery phase of the competition. In Fig. 2 we show the results over the course of the preliminary submission phase for both tasks. In Fig. 2a and

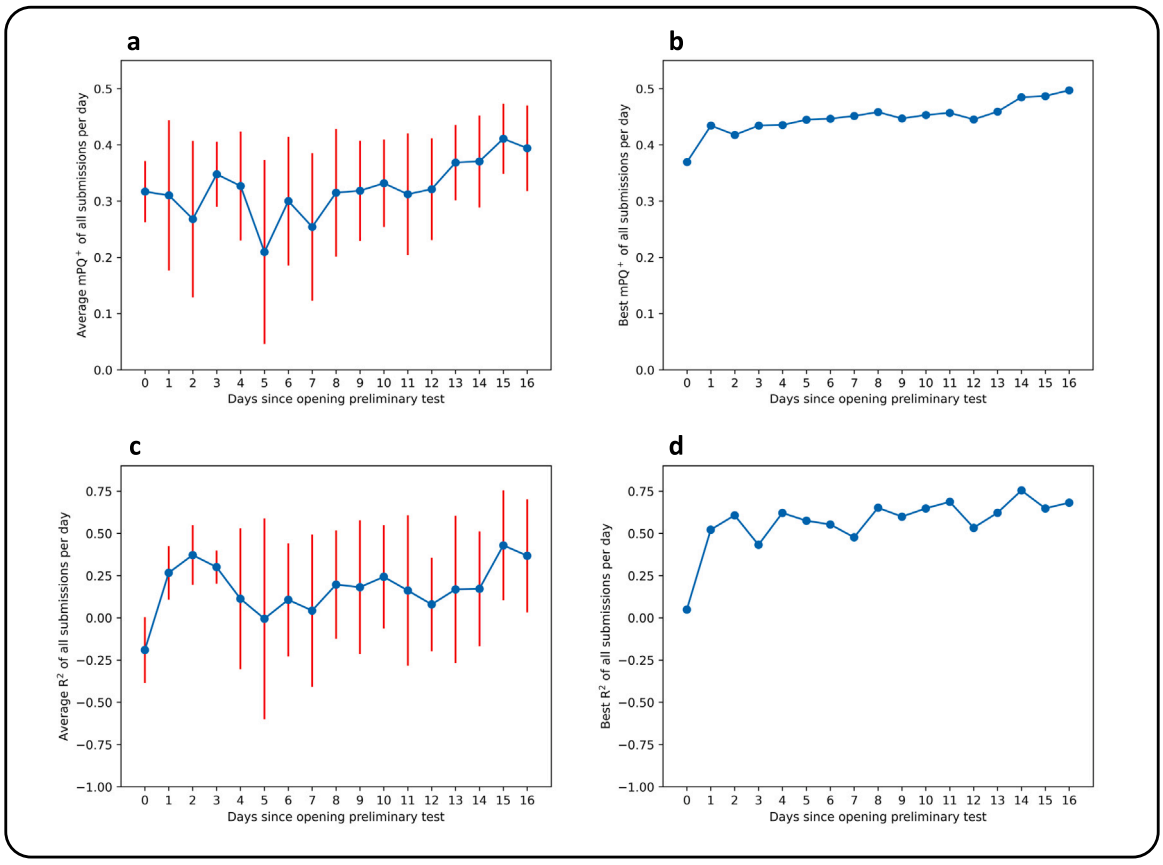

Fig. 2. Results during the preliminary submission phase of the competition for both tasks. **a**, Average score per day for the segmentation task; **b**, best score per day for the segmentation task; **c**, average score per day for the cellular composition task; **d**, best score per day for the cellular composition task.

Fig. 2c, we show the average results of all submissions per day over time, along with the corresponding error bars. In Fig. 2b and Fig. 2d, we show the best performance per day over time. Generally, we can see that the preliminary submission phase was successful in helping to improve the performance of submitted models, due to its competitive nature.

3.2. Segmentation and classification results

In Fig. 3 we display the final competition standings for the segmentation and classification task. These are shown in the form of a heat map, where results are sorted by their final mPQ^+ score. We observe that the epithelial cell, lymphocyte and connective tissue cell classes were the easiest to segment, with average PQ^+ scores across all participants of 0.513, 0.496 and 0.443, respectively. On the other hand, neutrophil and eosinophil classes were the most difficult nuclei, with average PQ^+ scores of 0.213 and 0.305. We believe that this was due to the large class imbalance in the dataset, with significantly fewer neutrophils and eosinophils. EPFL | StarDist, MDC Berlin | IFP Bern and Pathology AI were the top three submissions on this task, with mPQ^+ scores of 0.501, 0.476 and 0.463, respectively. These submissions all used an encoder–decoder based convolutional neural network, a strong instance segmentation target and a strategy to deal with the class imbalance. Dealing with the class imbalance was particularly important to rank highly. We display visual segmentation and classification results for the top participants in Fig. 5, where we observe that the models could successfully delineate the boundaries of different nuclei. It was especially impressive to see that submissions such as EPFL | StarDist were able to detect neutrophils within the lumen in the 3rd row of the figure. It is evident that some models struggled on the external TCGA dataset. For example, in the 5th row of Fig. 5, Pathology AI misclassified plasma cells as epithelial cells, whereas in the bottom row some participants failed to detect various epithelial nuclei. As an alternative form of visualisation, we also show final results as point plots in Fig. S2.

3.3. Cellular composition results

In Fig. 4, we show the final results of the cellular composition task. Results are sorted in order of their final position on the leaderboard, which was determined by the mR^2 score. The standard deviation of the top 20 submissions for mR^2 was 0.095, as opposed to 0.042 for mPQ^+, indicating that there was greater variability in the results for the cellular composition task. It is evident that participants who were able to sustain a good performance across all classes secured strong positions on the final leaderboard. Again, the epithelial, lymphocyte and connective tissue cell classes were the easiest to predict, with average R^2 scores over all submissions of 0.713, 0.722 and 0.673, respectively. The top three submissions for the cellular composition task were Pathology AI, AI_Medical and EPFL | StarDist with final scores of 0.7641, 0.7625 0.7550. Each of these submissions obtained good correlation scores for the minority classes, owing to their strong final positions. Despite us allowing participants to directly predict the counts from the image as a regression task, top results used an initial detection step before counting. However, we received significantly more two-stage submissions and so cannot make any conclusive remarks. Like the segmentation task, it was crucial for participants to employ a technique to deal with the class imbalance to perform well. We show final results as point plots in Fig. S3 and also compare the performance with additional metrics for both tasks in Fig. S4.

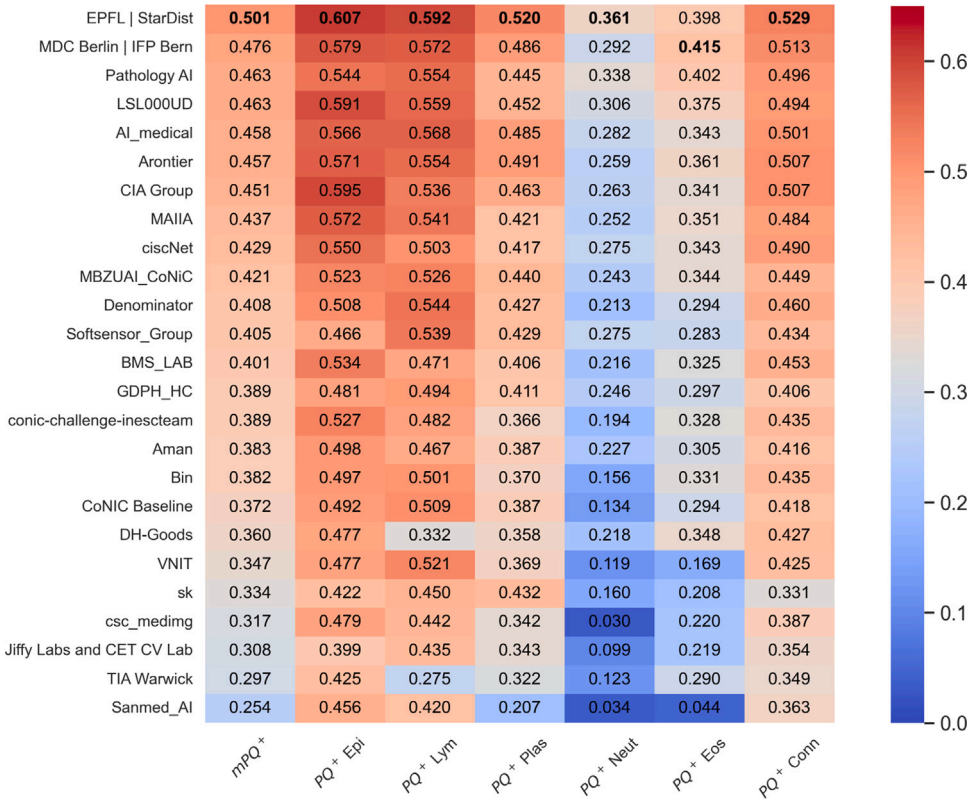

Fig. 3. Segmentation and classification challenge results on the final test set as a heat map. PQ^+ and mPQ^+ refer to the Panoptic Quality per class and averaged over all classes, respectively.

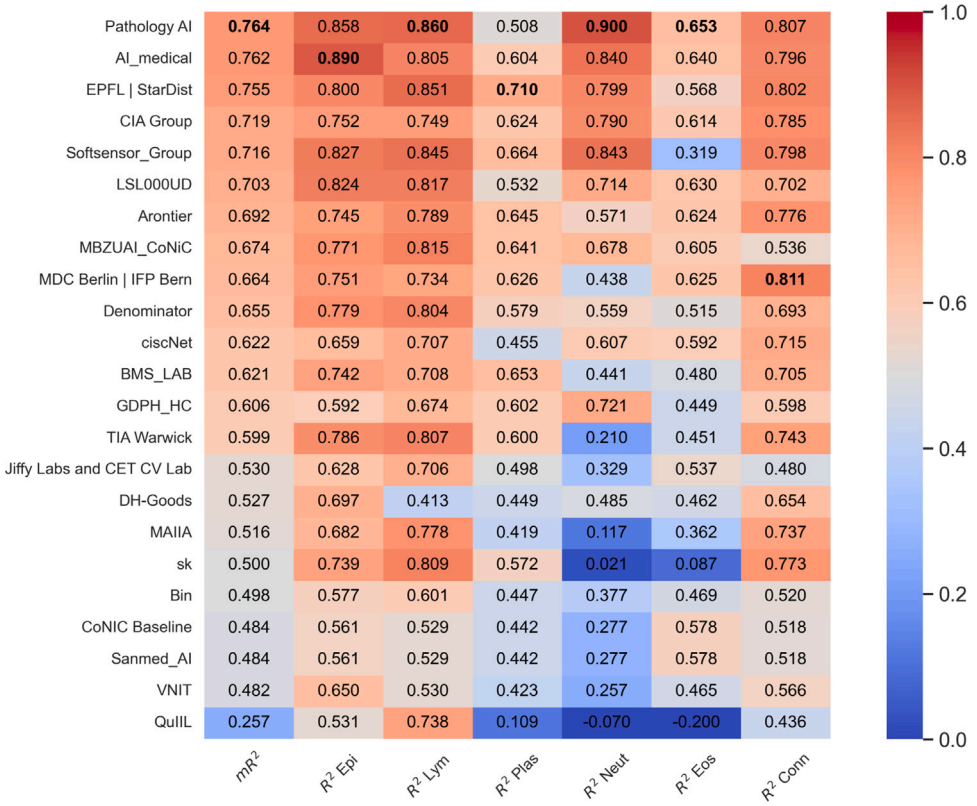

Fig. 4. Cellular composition challenge results on the final test set as a heat map. R^2 and mR^2 are the coefficient of determination per class and averaged over all classes, respectively.

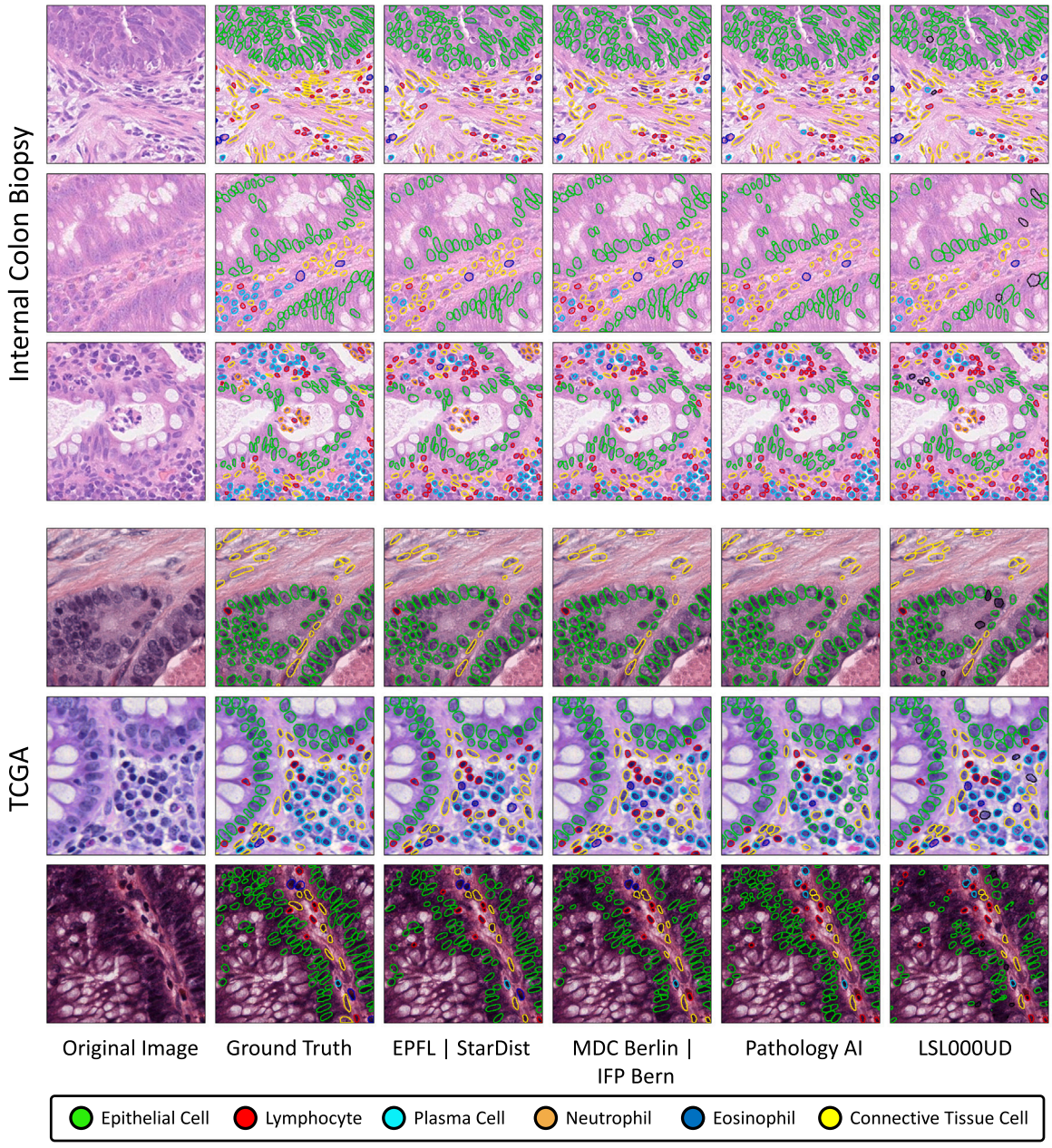

Fig. 5. Visual results from the top participants of the segmentation and classification task. The first 3 rows show results on the internal colon biopsy dataset from UHCW and the bottom 3 rows show results on the TCGA dataset from different submissions as well as the ground truth.

3.4. Impact of model ensembling and bootstrap analysis

To better understand how models compared, we asked the top teams to submit their algorithms that were trained on a specific data split and without ensembling. Here, ensembling involves combining results from multiple runs of the network with different architectures, checkpoints, or transformed images. We present the results in Fig. 6, with Fig. 6a and Fig. 6b showing the results for segmentation and classification, and Fig. 6c and Fig. 6d for cellular composition. Without ensembling, the mean score among these participants decreased from 0.4501 to 0.4330 for segmentation and classification and from 0.6823 to 0.6402 for cellular composition, showing that ensembling has a big impact on the final standing. In Fig. 6a and 6c, we show the heat map of the results, which are ordered by their original ranking in the competition. In parts Fig. 6b and Fig. 6d of the figure, we perform bootstrapping (n=100) of the submissions for each task to get the confidence bounds. The top three submissions for each task still perform well, even under the conditions we set for this experiment. We show the difference in performance between the original and single model submissions in Fig. S5.

3.5. Application of top models to downstream clinical tasks

Accurate recognition of nuclei in histopathology images enables the extraction of interpretable features for downstream clinical pipelines. Therefore, as a next step we assessed the impact of features derived from the top nuclear recognition algorithms on the tasks of dysplasia grading and survival analysis. To enable this, we processed a total of 1,658 WSIs with the top three ranked algorithms from the segmentation and classification task (EPFL | StarDist, MDC Berlin | IFP Bern and Pathology AI) before extracting a series of global cell-level features. Example visual results on two randomly selected WSIs are shown in Fig. 7. We only considered models trained on the single split of data because using models with excessive ensembling, like was done in

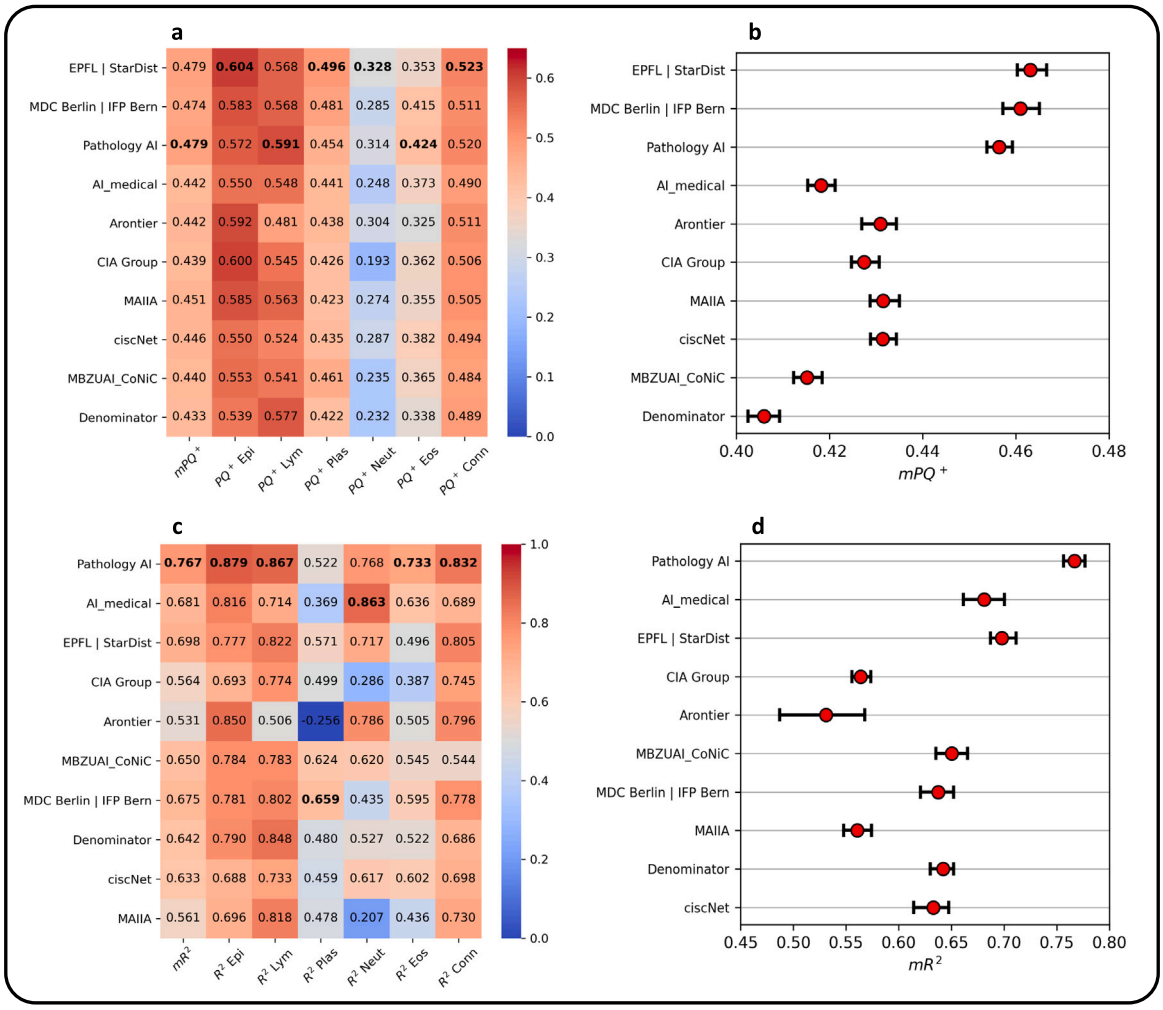

Fig. 6. Challenge results of the top participants trained on a single defined split of the data and without model ensembling. **a**, **c**, Model results on a single split as a heat map. **b**, **d**, Results using bootstrapping of the test set (n=100) to visualise the confidence bounds. **a**, **b**, Results for segmentation tasks and **c**, **d**, results for cellular composition tasks. PQ^{+} and mPQ^{+} refer to the Panoptic Quality per class and averaged over all classes, respectively. Similarly, R^2 and mR^2 are the coefficient of determination per class and averaged over all classes.

many of the original submissions, is not feasible when processing a large amount of WSIs. For our downstream experiments we used a total of 222 features that can be broadly categorised into the following groups: morphological, density and colocalisation features. These features were directly used as input to a machine learning model for automated diagnosis and patient stratification.

3.6. Conic models for cell-based dysplasia grading

To assess the impact of nuclear recognition performance on automated diagnosis, we utilised a dataset of 1132 colon WSIs from IMP Diagnostics Laboratory in Portugal (Oliveira et al., 2021) All data was H&E-stained and labelled as either non-neoplastic, low-grade dysplasia or high-grade dysplasia. We then used the global nuclear features obtained from the results of each of the top teams as input to a gradient boosted random forest, using five-fold cross validation to ensure reliable results. In Fig. 8a, we show the F_1 score and Quadratic Weighted Kappa (QWK) over all experimental runs for each team, where we observe that MDC Berlin | IFP Bern obtains the best performance with average mF_1 and QWK scores of 0.8739 and 0.8463 on the testing set, respectively.

3.7. Conic models for cell-based survival analysis

To assess impact of the nuclear features on being able to successfully stratify patients, we predicted overall and disease specific survival within 526 H&E-stained colorectal WSIs from The Cancer Genome Atlas (TCGA). For this experiment, we utilised the features as input to gradient boosted trees (XGBoost) and performed five-fold cross validation to predict a risk score for each patient. In Fig. 8b we show the concordance indices (C-indices) obtained when using the features from each of the top teams for survival tasks. Here, we see that EPFL | StarDist obtains the best performance with an average C-index of 0.6554 and 0.6456 respectively for predicting disease specific survival and overall survival. Detailed results for both dysplasia grading and survival analysis can be found in Section S3 of the supplementary material.

4. Discussion

The CoNIC Challenge, characterised by its competitive nature, has spurred rapid advancements in deep learning-based nuclear identification methods. This progress has resulted in numerous final submissions outperforming the prior state-of-the-art techniques (Graham et al., 2019). A predominant strategy among participants involved the

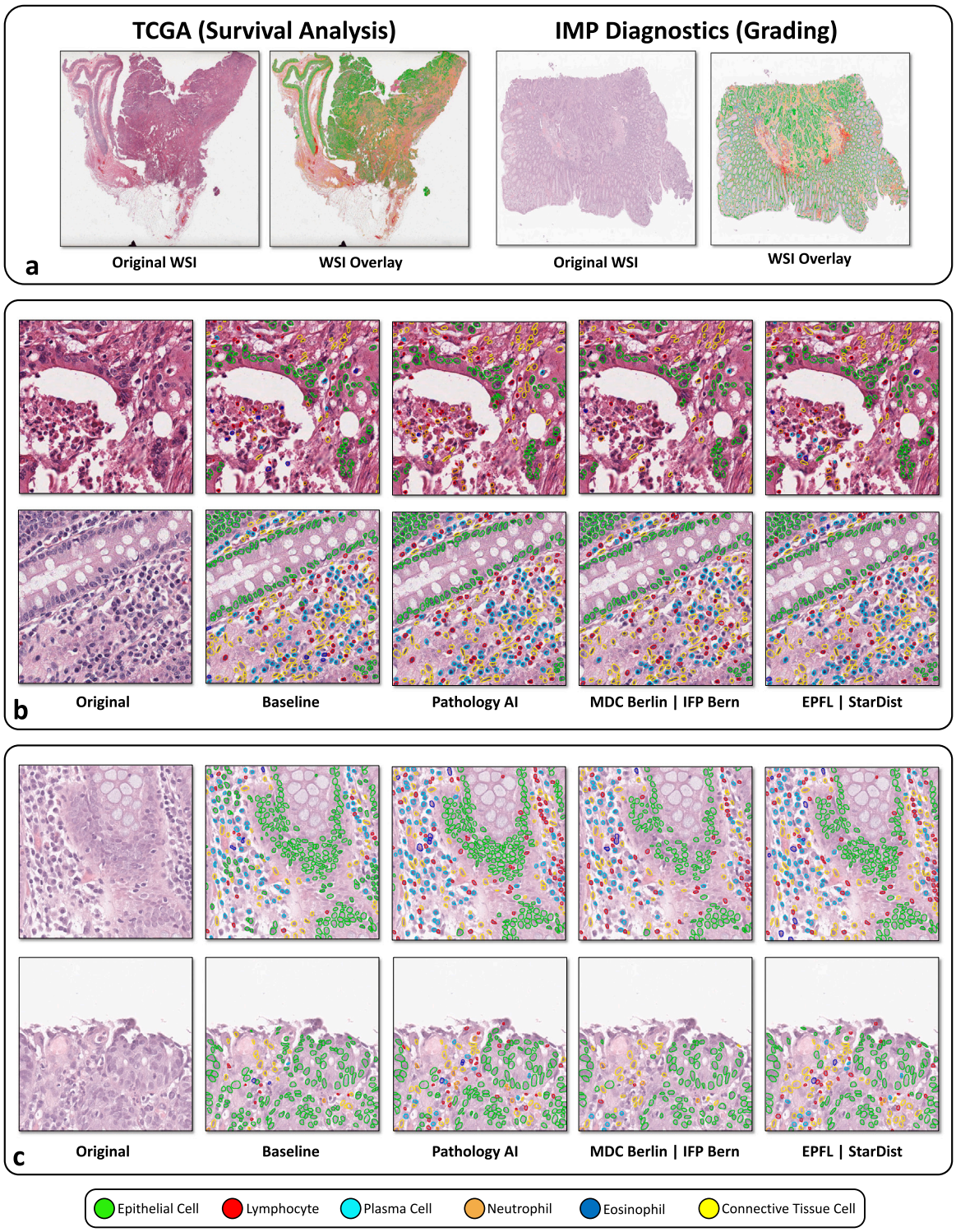

Fig. 7. WSI-level visual results from the top submissions on the segmentation and classification task trained on a single pre-defined split of the data. **a**, Original WSI along with an example nuclear segmentation overlay. **b**, Zoomed-in predictions of the top three teams compared to the baseline for the example TCGA WSI. **c**, Zoomed-in predictions of the top three teams compared to the baseline for the example IMP Diagnostics WSI. For b and c, each row shows a different region from the same WSI.

utilisation of deep learning models featuring an encoder–decoder architecture. In particular, successful submissions adopted specific tactics to address a considerable class imbalance present in the dataset, including techniques like patch oversampling or by incorporating weighted loss functions. This appears to have proved pivotal in securing a prominent position on the leaderboard. A significant observation was that many of the successful submissions introduced subtle refinements to HoVer-Net, indicating that the provided baseline served as a strong foundation for participants to build upon. Furthermore, it became evident that enhancing the baseline HoVer-Net approach could be achieved by considering a more advanced backbone or the incorporation of alternate instance segmentation targets, such as additional directional distance maps. We provide a visual summary of the algorithms submitted by the participants, along with the training details in Fig. S15. We also provide a more detailed description of each of the approaches in Section S1 of the supplementary material.

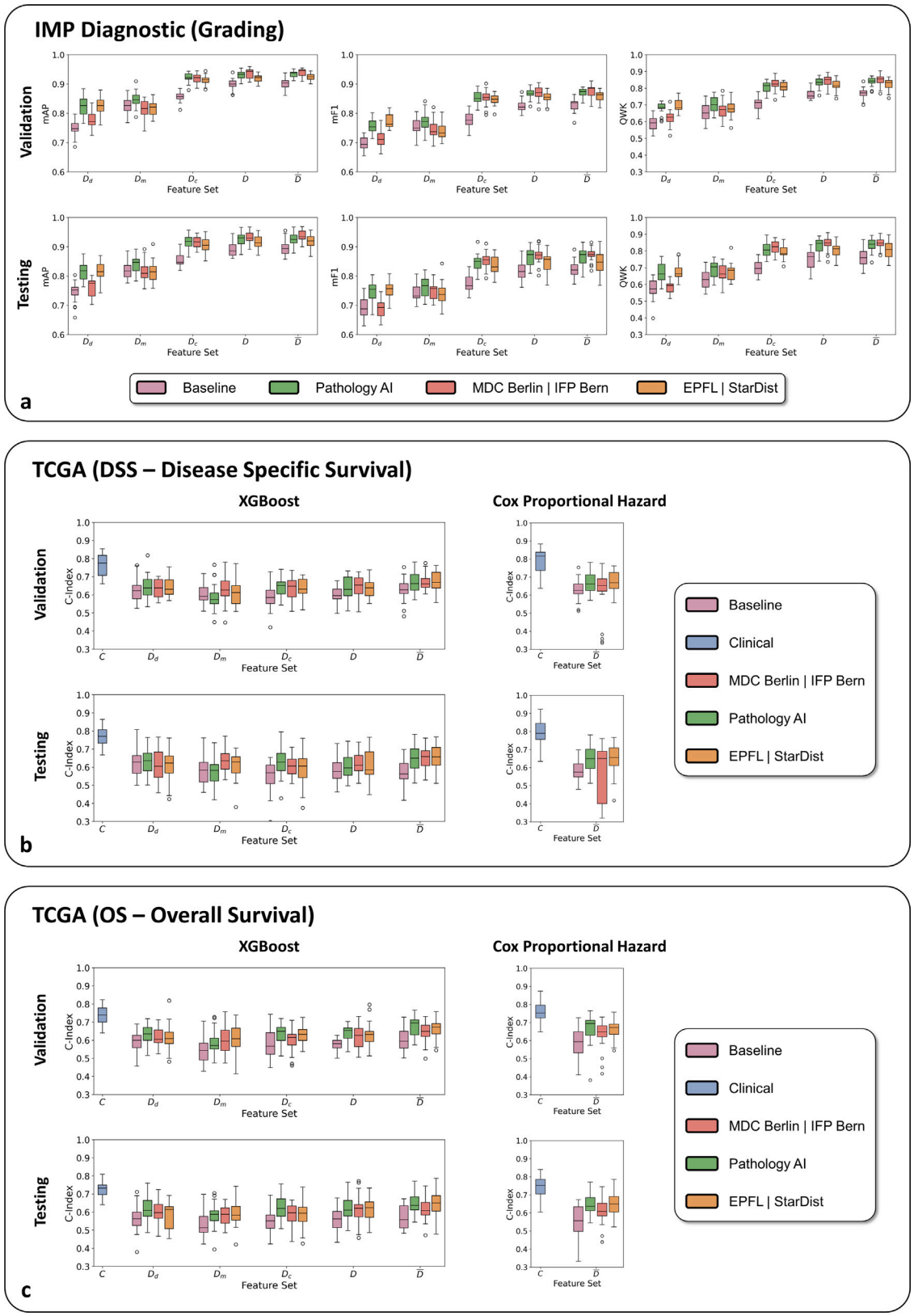

Fig. 8. Results when using nucleus features computed based on the predictions of the top three teams in segmentation and classification task for patient-level grading and survival analysis. **a**, Results for grading, **b**, results for disease specific survival analysis and **c**, results for overall survival analysis. mF_1 and mAP denote the mean F_1 and mean Average Precision scores, respectively. C-Index is the concordance index, which is a commonly used metric for survival analysis. D_d, D_m and D_c refer to density-based, morphological and colocalization features, respectively. D is the entire feature set and $\overline{D}$ is the set after feature selection. C is the set of clinical features.

Despite us allowing participants to treat each task independently, nearly all submissions inferred the cellular composition from the segmentation and classification output. However, those teams that predicted the cellular composition directly from the original image did not achieve a high ranking on the final leaderboard. Determination of the cellular composition allows us to effectively model the tumor microenvironment (TME) of the tissue, which has been shown to be particularly powerful as a prognostic indicator. With the challenge acting as a facilitator to improve automatic nuclear profiling, we hope that it will stimulate the development of advanced methods correlating the TME to patient outcome.

Overall, we found that participants were able to achieve a strong performance on both tasks on our developed dataset. However, despite significant advancements being made within the challenge, additional work is required to further boost the ability to recognise minority classes, such as neutrophils and eosinophils. This may be achieved with the help of additional data and the development of new strategies for dealing with the class imbalance. A particularly interesting technique that was used in the challenge (Arontier and Aman) is copy-and-paste augmentation, which can be used to artificially increase the number of under-represented nuclei in the dataset. Another strategy that appears promising is the utilisation of generative methods to create synthetic images containing minority classes, while preserving the expected spatial configuration of nuclei within the tissue (Deshpande et al., 2022).

The dataset that we introduced as part of the challenge is the largest existing dataset of nuclear segmentation and classification in CPath. Despite this, our dataset is currently from a single tissue type and so we cannot guarantee that models developed during the challenge will generalise to unseen tissues, despite various inflammatory cells appearing the same across different organs. In future work, we may extend the current dataset to other major tissue types, such as breast, prostate and lung to increase the range of downstream applications that the models can be applied to. Also, we currently group endothelial cells, fibroblasts and muscle cells into a single category. Explicit separation of these classes will enable the consideration of features such as cancer-associated fibroblasts and endothelial cell morphology, which can be prognostically informative (Sahai et al., 2020; Tommelein et al., 2015; Hida et al., 2016). We are also aware of the limitations that exist as a result of labelling nuclei using only routine H&E slides. In future work, perhaps it would be advantageous to instead rely on co-registered H&E and IHC slides to provide more accurate ground truth, especially for immune cell subtyping.

We ensured that our dataset was sufficiently large to provide a good indication for how models perform across a range of scanner types and lab preparation methods. However, this does not guarantee that developed models will work out-of-the-box when deployed in a clinical setting. Future work may include a thorough investigation into the robustness of AI models for nuclear identification (Vu et al., 2022; Foote et al., 2022), where lessons learned can help reduce the likelihood of unexpected model behaviour in the wild.

In addition to the main challenge, we utilised the baseline and three best performing models from the segmentation and classification task and processed 1,658 WSIs from two datasets, with the intention of understanding how the performance of automated nuclear identification affects downstream clinical tasks and identifying whether state-of-the-art (SoTA) methods can improve upon the baseline. In particular, we extracted patient-level features from the WSI results and used them as input to perform automatic dysplasia grading and survival analysis. From Fig. 8, the digital features based on the SoTA methods are all significantly more predictive than the baseline features ($p \ll 0.001$, student-t test). This suggests that accurate models for nuclear recognition may in fact be essential when using associated features for downstream tasks. Although there is an apparent relationship between CoNIC results and the performance of subsequent tasks, further work is required to understand the implications of further significant boosts in nuclear recognition. We found that despite small differences in the best results obtained within the challenge, there was notable variation in the importance of features when applied to downstream tasks. Given that many studies typically utilise predictions from a single nucleus segmentation and classification method, our results raise questions on the validity of identified digital features using just a single approach, especially for survival analysis problems.

Specifically, when grading dysplasia, while all three SoTA methods achieved similar performance, MDC Berlin | IFP Bern disproportionately utilised statistics concerning plasma nuclei morphology while neglecting those from eosinophils (Fig. S7a and Fig. S8a). On the other hand, neutrophil morphology was determined to be more important for EPFL | StarDist than any other method. In spite of these differences, all three methods considered the morphology of epithelial nuclei as important features for accurately predicting dysplasia grade (Fig. S9 to S11). This finding aligns with existing clinical observations (Shia et al., 2017)

For survival analyses, from Fig. 7b and Fig. 7c, we observe that SoTA methods have significant variation in performance. Not only that, but from Fig. S7b and Fig. S7c, we observe that they have a different set of identified features for a given task. Despite this, we found that statistics concerning the TME of neutrophils were consistently identified as important features, as shown in Fig. S8b and Fig. S8c as well as from Fig. S12 to S14. This observation aligns with existing clinical studies (Schmitt and Greten, 2021) which perhaps encourages further investigation into neutrophils and their impact and emphasises the need to further improve neutrophil detection performance. Also, from those same figures, while not as informative as neutrophil-based features, all teams also agree that a higher cellular composition of eosinophils relate to better patient outcome. This observation confirms a recent clinical finding (Reichman et al., 2019) which states that eosinophils have anti-tumourigenic properties in colorectal cancers.

As a result of its competitive nature, the utilisation of the largest dataset of its kind and the existence of a rigorous evaluation protocol, we believe that the CoNIC Challenge has been largely influential in helping to further push forward the state-of-the-art for automatic nuclear recognition in CPath. To foster the development of future approaches for cell-based biomarker exploration, we are releasing the WSI-level results using the best methods from the challenge. We are also accepting post-challenge submissions using the same evaluation framework as the original competition.

4.1. Code availability

Evaluation code used within the challenge, along with example notebooks can be found at the following repository: https://github.com/TissueImageAnalytics/CoNIC. A template for making code-based challenge submissions can be found in a separate branch of the same repository. Code used for running the baseline can be found in a separate branch of the original HoVer-Net repository at https://github.com/vqdang/hover_net/tree/conic.

4.2. Ethics approval

The additional test data collected for this challenge was performed under Health Research Authority National Research Ethics approval 15/NW/0843; IRAS 189095 and the Pathology image data Lake for Analytics, Knowledge and Education (PathLAKE) research ethics committee approval (REC reference 19/SC/0363, IRAS project ID 257932, South Central—Oxford C Research Ethics Committee). The study was conducted on retrospective data from histopathology archives relating to samples taken in the course of clinical care, and for which consent for research had not been taken. Gathering consent retrospectively was not feasible and deemed not necessary by the research ethics committee, as referenced above.

CRediT authorship contribution statement

Simon Graham: Designed and conducted the study, Curated the challenge dataset, Performed analysis and interpretation of the results, Analysed the impact of WSI-level nuclear recognition results on downstream applications, Wrote the first draft of the paper. **Quoc Dang Vu:** Designed and conducted the study, Set up the docker-based submission and evaluation framework for the challenge, Performed analysis and interpretation of the results, Processed all slides with the best challenge algorithms, Analysed the impact of WSI-level nuclear recognition results on downstream applications. **Mostafa Jahanifar:** Designed and conducted the study. **David Snead:** Provided technical and material support. **Shan E. Ahmed Raza:** Designed and conducted the study, Provided technical and material support. **Fayyaz Minhas:** Designed and conducted the study, Provided technical and material support. **Nasir M. Rajpoot:** Designed and conducted the study, Provided technical and material support.

Declaration of competing interest

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

Data availability

The challenge data can be downloaded at https://conic-challenge.grand-challenge.org.

Acknowledgements

SG, MJ, DS, SR, FM and NR would like to acknowledge the support from the PathLAKE digital pathology consortium which is funded by the Data to Early Diagnosis and Precision Medicine strand of the government's Industrial Strategy Challenge Fund, managed and delivered by UK Research and Innovation (UKRI). FM acknowledges funding from EPSRC, United Kingdom grant EP/W02909X/1. We thank Georgios Hadjigeorghiou and Thomas Leech for initial discussions regarding the challenge setup. All authors read and approved the final paper.

Appendix A. The CoNIC Challenge Consortium[ae]

Dagmar Kainmueller[1,2], Carola-Bibiane Schönlieb[3], Shuolin Liu[4], Dhairya Talsania[5,6], Yughender Meda[5,6], Prakash Mishra[5,6], Muhammad Ridzuan[7], Oliver Neumann[8], Marcel P. Schilling[8], Markus Reischl[8], Ralf Mikut[8], Banban Huang[9], Hsiang-Chin Chien[10], Ching-Ping Wang[10], Chia-Yen Lee[11], Hong-Kun Lin[12], Zaiyi Liu[13], Xipeng Pan[13], Chu Han[13], Jijun Cheng[14], Muhammad Dawood[15], Srijay Deshpande[15], Raja Muhammad Saad Bashir[15], Adam Shephard[15], Pedro Costa[16,17], João D. Nunes[16,17], Aurélio Campilho[16,17], Jaime S. Cardoso[16,17], Hrishikesh P S[18], Densen Puthussery[18], Devika R G[19], Jiji C V[19], Ye Zhang[20], Zijie Fang[21], Zhifan Lin[20], Yongbing Zhang[20], Chunhui Lin[21], Liukun Zhang[22], Lijian Mao[22], Min Wu[22], Vi Thi-Tuong Vo[23], Soo-Hyung Kim[23], Taebum Lee[24], Satoshi Kondo[25], Satoshi Kasai[26], Pranay Dumbhare[27], Vedant Phuse[27], Yash Dubey[27], Ankush Jamthikar[27], Trinh Thi Le Vuong[28], Jin Tae Kwak[28], Dorsa Ziaei[29], Hyun Jung[29], Tianyi Miao[29]

[1]Max-Delbrueck-Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany. [2]Humboldt University of Berlin, Faculty of Mathematics and Natural Sciences, Berlin, Germany. [3]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, United Kingdom [4]Department of Electrical Engineering and Automation, AnHui University, HeiFei, China. [5]Softsensor.ai, Bridgewater, New Jersey, United States of America [6]PRR.ai, Texas, United States of America. [7]Computer Vision Department, Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates. [8]Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany. [9]School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou, China. [10]Department of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan. [11]Department of Electrical Engineering, National United University, Miaoli, Taiwan. [12]Institute of Biomedical Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan. [13]Department of Radiology, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, Guangdong, China. [14]School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin, Guangxi, China. [15]Tissue Image Analytics Centre, University of Warwick, Coventry, United Kingdom [16]Faculty of Engineering, University of Porto, Porto, Portugal. [17]Institute for Systems and Computer Engineering Technology and Science, Porto, Portugal. [18]FMS Lab, Founding Minds Software, Cochin, India. [19]Department of Electronics and Communication, College of Engineering, Trivandrum, India. [20]Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China. [21]Tsinghua Shenzhen International Graduate School, Shenzhen, China. [22]Research and Development Center, Zhejiang Dahua Technology Co., Ltd, Hangzhou, Zhejiang, China. [23]Department of AI Convergence, Chonnam National University, Gwangju, South Korea. [24]Department of Pathology, Chonnam National University Medical School, Gwangju, South of Korea. [25]Muroran Institute of Technology, Hokkaido, Japan. [26]Niigata University of Healthcare and Welfare, Niigata, Japan. [27]Visvesvaraya National Institute of Technology, Nagpur, India. [28]School of Electrical Engineering, Korea University, Seoul, Republic of Korea. [29]Advanced Biomedical Computational Science, Frederick National Laboratory for Cancer Research, Frederick, United States of America.

[ae]The consortium comprises of an extended list of authors contributing towards the challenge.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2023.103047.

References

Altmann, A., Toloşi, L., Sander, O., Lengauer, T., 2010. Permutation importance: a corrected feature importance measure. Bioinformatics 26 (10), 1340–1347.

Awan, R., Sirinukunwattana, K., Epstein, D., Jefferyes, S., Qidwai, U., Aftab, Z., Mujeeb, I., Snead, D., Rajpoot, N., 2017. Glandular morphometrics for objective grading of colorectal adenocarcinoma histology images. Sci. Rep. 7 (1), 16852.

Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., et al., 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 318 (22), 2199–2210.

Berry, S., Giraldo, N.A., Green, B.F., Cottrell, T.R., Stein, J.E., Engle, E.L., Xu, H., Ogurtsova, A., Roberts, C., Wang, D., et al., 2021. Analysis of multispectral imaging with the AstroPath platform informs efficacy of PD-1 blockade. Science 372 (6547), eaba2609.

Bulten, W., Kartasalo, K., Chen, P.-H.C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D.F., van Boven, H., Vink, R., et al., 2022. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the PANDA challenge. Nat. Med. 28 (1), 154–163.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. pp. 785–794.

Da, Q., Huang, X., Li, Z., Zuo, Y., Zhang, C., Liu, J., Chen, W., Li, J., Xu, D., Hu, Z., et al., 2022. DigestPath: A benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system. Med. Image Anal. 80, 102485.

Dawood, M., Branson, K., Rajpoot, N.M., Minhas, F., 2021. Albrt: Cellular composition prediction in routine histology images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 664–673.

Deshpande, S., Dawood, M., Minhas, F., Rajpoot, N., 2022. SynCLay: Interactive synthesis of histology images from bespoke cellular layouts. arXiv preprint arXiv: 2212.13780.

Foote, A., Asif, A., Rajpoot, N., Minhas, F., 2022. REET: robustness evaluation and enhancement toolbox for computational pathology. Bioinformatics 38 (12), 3312–3314.

Gamper, J., Alemi Koohbanani, N., Benet, K., Khuram, A., Rajpoot, N., 2019. PanNuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In: Digital Pathology: 15th European Congress, ECDP 2019, Warwick, UK, April 10–13, 2019, Proceedings 15. Springer, pp. 11–19.

Gamper, J., Koohbanani, N.A., Benes, K., Graham, S., Jahanifar, M., Khurram, S.A., Azam, A., Hewitt, K., Rajpoot, N., 2020. Pannuke dataset extension, insights and baselines. arXiv preprint arXiv:2003.10778.

Graham, S., Jahanifar, M., Azam, A., Nimir, M., Tsang, Y.-W., Dodd, K., Hero, E., Sahota, H., Tank, A., Benes, K., et al., 2021. Lizard: A large-scale dataset for colonic nuclear instance segmentation and classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 684–693.

Graham, S., Vu, Q.D., Raza, S.E.A., Azam, A., Tsang, Y.W., Kwak, J.T., Rajpoot, N., 2019. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Med. Image Anal. 58, 101563.

Hida, K., Maishi, N., Torii, C., Hida, Y., 2016. Tumor angiogenesis—characteristics of tumor endothelial cells. Int. J. Clin. Oncol. 21 (2), 206–212.

Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.-A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D., et al., 2019. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS Med. 16 (1), e1002730.

Kira, K., Rendell, L.A., 1992. A practical approach to feature selection. In: Machine Learning Proceedings 1992. Elsevier, pp. 249–256.

Koohbanani, N.A., Jahanifar, M., Tajadin, N.Z., Rajpoot, N., 2020. NuClick: a deep learning framework for interactive segmentation of microscopic images. Med. Image Anal. 65, 101771.

Kumar, N., Verma, R., Anand, D., Zhou, Y., Onder, O.F., Tsougenis, E., Chen, H., Heng, P.-A., Li, J., Hu, Z., et al., 2019. A multi-organ nucleus segmentation challenge. IEEE Trans. Med. Imaging 39 (5), 1380–1391.

Kursa, M.B., Rudnicki, W.R., 2010. Feature selection with the boruta package. J. Stat. Softw. 36, 1–13.

Lennard-Jones, J., 1989. Classification of inflammatory bowel disease. Scand. J. Gastroenterol. 24 (sup170), 2–6.

Lu, M.Y., Chen, T.Y., Williamson, D.F., Zhao, M., Shady, M., Lipkova, J., Mahmood, F., 2021. AI-based pathology predicts origins for cancers of unknown primary. Nature 594 (7861), 106–110.

Magro, F., Langner, C., Driessen, A., Ensari, A., Geboes, K., Mantzaris, G., Villanacci, V., Becheanu, G., Nunes, P.B., Cathomas, G., et al., 2013. European consensus on the histopathology of inflammatory bowel disease. J. Crohn's Colitis 7 (10), 827–851.

Oliveira, S.P., Neto, P.C., Fraga, J., Montezuma, D., Monteiro, A., Monteiro, J., Ribeiro, L., Gonçalves, S., Pinto, I.M., Cardoso, J.S., 2021. CAD systems for colorectal cancer from WSI are still not ready for clinical acceptance. Sci. Rep. 11 (1), 1–15.

Rakha, E.A., El-Sayed, M.E., Lee, A.H., Elston, C.W., Grainge, M.J., Hodi, Z., Blamey, R.W., Ellis, I.O., 2008. Prognostic significance of nottingham histologic grade in invasive breast carcinoma. J. Clin. Oncol. 26 (19), 3153–3158.

Reichman, H., Itan, M., Rozenberg, P., Yarmolovski, T., Brazowski, E., Varol, C., Gluck, N., Shapira, S., Arber, N., Qimron, U., et al., 2019. Activated eosinophils exert antitumorigenic activities in colorectal CancerEosinophils in colorectal cancer. Cancer Immunol. Res. 7 (3), 388–400.

Ropponen, K.M., Eskelinen, M.J., Lipponen, P.K., Alhava, E., Kosma, V.-M., 1997. Prognostic value of tumour-infiltrating lymphocytes (TILs) in colorectal cancer. J. Pathol. J. Pathol. Soc. Great Britain Ireland 182 (3), 318–324.

Sahai, E., Astsaturov, I., Cukierman, E., DeNardo, D.G., Egeblad, M., Evans, R.M., Fearon, D., Greten, F.R., Hingorani, T., et al., 2020. A framework for advancing our understanding of cancer-associated fibroblasts. Nat. Rev. Cancer 20 (3), 174–186.

Salgado, R., Denkert, C., Demaria, S., Sirtaine, N., Klauschen, F., Pruneri, G., Wienert, S., Van den Eynden, G., Baehner, F.L., Pénault-Llorca, F., et al., 2015. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an international TILs working group 2014. Ann. Oncol. 26 (2), 259–271.

Schmidt, U., Weigert, M., Broaddus, C., Myers, G., 2018. Cell detection with star-convex polygons. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11. Springer, pp. 265–273.

Schmitt, M., Greten, F.R., 2021. The inflammatory pathogenesis of colorectal cancer. Nat. Rev. Immunol. 21 (10), 653–667.

Shaban, M., Awan, R., Fraz, M.M., Azam, A., Tsang, Y.-W., Snead, D., Rajpoot, N.M., 2020. Context-aware convolutional neural network for grading of colorectal cancer histology images. IEEE Trans. Med. Imaging 39 (7), 2395–2405.

Shia, J., Schultz, N., Kuk, D., Vakiani, E., Middha, S., Segal, N.H., Hechtman, J.F., Berger, M.F., Stadler, Z.K., Weiser, M.R., et al., 2017. Morphological characterization of colorectal cancers in the cancer genome atlas reveals distinct morphology–molecular associations: clinical and biological implications. Modern Pathol. 30 (4), 599–609.

Sirinukunwattana, K., Pluim, J.P., Chen, H., Qi, X., Heng, P.-A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., et al., 2017. Gland segmentation in colon histology images: The glas challenge contest. Med. Image Anal. 35, 489–502.

Tommelein, J., Verset, L., Boterberg, T., Demetter, P., Bracke, M., De Wever, O., 2015. Cancer-associated fibroblasts connect metastasis-promoting communication in colorectal cancer. Front. Oncology 5, 63.

Verma, R., Kumar, N., Patil, A., Kurian, N.C., Rane, S., Sethi, A., 2020. Multi-organ nuclei segmentation and classification challenge 2020. IEEE Trans. Med. Imaging 39 (1380–1391), 8.

Vu, Q.D., Graham, S., Kurc, T., To, M.N.N., Shaban, M., Qaiser, T., Koohbanani, N.A., Khurram, S.A., Kalpathy-Cramer, J., Zhao, T., et al., 2019. Methods for segmentation and classification of digital microscopy tissue images. Front. Bioeng. Biotechnol. 53.

Vu, Q.D., Jewsbury, R., Graham, S., Jahanifar, M., Raza, S.E.A., Minhas, F., Bhalerao, A., Rajpoot, N., 2022. Nuclear segmentation and classification: On color and compression generalization. In: Machine Learning in Medical Imaging: 13th International Workshop, MLMI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings. Springer, pp. 249–258.

Vu, Q.D., Rajpoot, K., Raza, S.E.A., Rajpoot, N., 2023. Handcrafted histological transformer (H2T): Unsupervised representation of whole slide images. Med. Image Anal. 102743.

Wahab, N., Miligy, I.M., Dodd, K., Sahota, H., Toss, M., Lu, W., Jahanifar, M., Bilal, M., Graham, S., Park, Y., et al., 2022. Semantic annotation for computational pathology: Multidisciplinary experience and best practice recommendations. J. Pathol. Clin. Res. 8 (2), 116–128.

Wulczyn, E., Steiner, D.F., Xu, Z., Sadhwani, A., Wang, H., Flament-Auvigne, I., Mermel, C.H., Chen, P.-H.C., Liu, Y., Stumpe, M.C., 2020. Deep learning-based survival prediction for multiple cancer types using histopathology images. PLoS One 15 (6), e0233678.

Zhu, W., Xie, L., Han, J., Guo, X., 2020. The application of deep learning in cancer prognosis prediction. Cancers 12 (3), 603.