npj | precision oncology

Article

Published in partnership with The Hormel Institute, University of Minnesota

https://doi.org/10.1038/s41698-024-00539-4

An interpretable machine learning system for colorectal cancer diagnosis from pathology slides

Check for updates

Pedro C. Neto[1,2,8] ✉, Diana Montezuma[3,4,5,8] ✉, Sara P. Oliveira[1,2,8] ✉, Domingos Oliveira[3], João Fraga[6], Ana Monteiro[3], João Monteiro[3], Liliana Ribeiro[3], Sofia Gonçalves[3], Stefan Reinhard[7], Inti Zlobec[7], Isabel M. Pinto[3] & Jaime S. Cardoso[1,2]

Considering the profound transformation affecting pathology practice, we aimed to develop a scalable artificial intelligence (AI) system to diagnose colorectal cancer from whole-slide images (WSI). For this, we propose a deep learning (DL) system that learns from weak labels, a sampling strategy that reduces the number of training samples by a factor of six without compromising performance, an approach to leverage a small subset of fully annotated samples, and a prototype with explainable predictions, active learning features and parallelisation. Noting some problems in the literature, this study is conducted with one of the largest WSI colorectal samples dataset with approximately 10,500 WSIs. Of these samples, 900 are testing samples. Furthermore, the robustness of the proposed method is assessed with two additional external datasets (TCGA and PAIP) and a dataset of samples collected directly from the proposed prototype. Our proposed method predicts, for the patch-based tiles, a class based on the severity of the dysplasia and uses that information to classify the whole slide. It is trained with an interpretable mixed-supervision scheme to leverage the domain knowledge introduced by pathologists through spatial annotations. The mixed-supervision scheme allowed for an intelligent sampling strategy effectively evaluated in several different scenarios without compromising the performance. On the internal dataset, the method shows an accuracy of 93.44% and a sensitivity between positive (low-grade and high-grade dysplasia) and non-neoplastic samples of 0.996. On the external test samples varied with TCGA being the most challenging dataset with an overall accuracy of 84.91% and a sensitivity of 0.996.

Colorectal cancer (CRC) incidence and mortality are increasing, with projections indicating continued growth until at least 2040, according to estimations of the International Agency for Research on Cancer[1]. Nowadays, it is the third most incident (10.7% of all cancer diagnoses) and the second most deadly type of cancer[1]. Despite the pessimist predictions, CRC is preventable and curable when detected in its earlier stages. Thus, effective screening through medical examination, imaging techniques and colonoscopy are of utmost importance[2,3]. Notwithstanding the CRC detection capabilities shown by imaging/endoscopic techniques, the definite diagnosis of cancer is always based on the pathologist's evaluation of the histological

[1]Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), R. Dr. Roberto Frias, Porto 4200-465 Porto, Portugal. [2]Faculty of Engineering, University of Porto (FEUP), R. Dr. Roberto Frias, Porto 4200-465 Porto, Portugal. [3]IMP Diagnostics, Praça do Bom Sucesso, 61, sala 808, Porto 4150-146 Porto, Portugal. [4]Cancer Biology and Epigenetics Group, Research Center of IPO Porto (CI-IPOP) / RISE@CI-IPOP (Health Research Network), Portuguese Oncology Institute of Porto (IPO Porto) / Porto Comprehensive Cancer Center (Porto.CCC), R. Dr. António Bernardino de Almeida 865, Porto 4200-072 Porto, Portugal. [5]Doctoral Programme in Medical Sciences, School of Medicine and Biomedical Sciences - University of Porto (ICBAS-UP), R. Jorge de Viterbo Ferreira 228, Porto 4050-313 Porto, Portugal. [6]Department of Pathology, IPO-Porto, R. Dr. António Bernardino de Almeida 865, Porto 4200-072 Porto, Portugal. [7]Institute of Pathology, University of Bern, Uni Bern, Murtenstrasse 31, Bern 3008 Bern, Switzerland. [8]These authors contributed equally: Pedro C. Neto, Diana Montezuma and Sara P. Oliveira. ✉e-mail: pedro.d.carneiro@inesctec.pt; diana.felizardo@impdiagnostics.com; s.oliveira@nki.nl

samples. The stratification of neoplasia development stages consists of non-neoplastic (NNeo), low-grade dysplasia (LGD), high-grade dysplasia (HGD, including intramucosal carcinomas), and invasive carcinomas, from the initial to the latest stage of cancer progression, respectively. In spite of the inherent subjectivity of the dysplasia grading system[4], recent guidelines from the European Society of Gastrointestinal Endoscopy, as well as those from the US multi-society task force on CRC, consistently recommend shorter surveillance intervals for patients with polyps with high-grade dysplasia, regardless of their dimension[3,5]. Hence, grading dysplasia is still routinely performed by pathologists worldwide when assessing colorectal tissue samples.

Private datasets of digitised slides are becoming widely available, in the form of whole-slide images (WSI), with an increase in the adoption of digital workflows[6–8]. WSI eases the revision of old cases, data sharing and peer-review[9,10]. It has also created several research opportunities within the computer vision domain, especially due to the complexity of the problem and the high dimensions of WSIs[11–14]. Robust and high-performance systems can be valuable assets to the digital workflow of a laboratory, especially if they are transparent and interpretable[9,10]. However, some limitations still affect the applicability of such solutions into practice[15].

The analysis of CRC samples from WSI is divided into two different branches: classification of regions of interest, and classification of WSI. On the latter topic, despite the limitations, researchers have been improving the state of the art on the classification of the slide from individual tile classification, or aggregation methods[15–18]. In 2020, ref. 19 used a recurrent neural network to aggregate the predictions of individual tiles processed by an Inception-v3 network into non-neoplastic, adenoma (AD) and adenocarcinoma (ADC). Due to the large dimensions of WSI related to their pyramidal format (with several magnification levels)[20], usually over 50,000 × 50,000 pixels, it is usual to use a scheme consisting of a grid of tiles. This scheme permits the acceleration of the processing steps since the tiles are small enough to fit in the memory of the graphics processing units (GPU), popular units for the training of deep learning (DL). Reference 21 studied the usage of an ensemble of five distinct ResNet networks, in order to distinguish the types of CRC adenomas H&E stained slides. Reference 22 experimented with a modified DeepLab-v2 network for tile classification, and proposed pixel probability thresholding to detect CRC adenomas. Both refs. 23–25 looked into the performance of the Inception-v3 architecture to detect CRC, with the latter also retrieving a cluster-based slide classification and a map of predictions. The MuSTMIL[26] method classifies five colon-tissue findings: normal glands, hyperplastic polyps, low-grade dysplasias, high-grade dysplasias and carcinomas. This classification originates from a multitask architecture that leverages several levels of magnification of a slide. Reference 27 extended the experiments with multitask learning, but instead of leveraging the magnification, its model aims to jointly segment glands, detect tumour areas and sort the slides into low-risk (benign, inflammation or reactive changes) and high-risk (adenocarcinoma or dysplasia) categories. The architecture of this model is considerably more complex, with regard to the number of parameters, and is known as Faster-RCNN with a ResNet-101 backbone network for the segmentation task. Further to this task, a gradient-boosted decision tree completes the pipeline that results in the final grade. More recently, ref. 28 presented an DL-based method to segment multiple colorectal tissue compartments and then used the best performing model classify biopsies as either (1) high-risk (tumour and high-grade dysplasia), (2) low-grade dysplasia, (3) hyperplasia and (4) benign; achieving an one-vs-all AUC of 0.87 for the high-risk category. Notably, ref. 29 have developed a graph neural network, Interpretable Gland-Graphs using a Neural Aggregator (IGUANA), to distinguish colorectal samples in normal vs. abnormal (non-neoplastic and neoplastic), achieving a sensitivity threshold of 99%, proposing, with their model, to reduce the number of normal slides to be reviewed by pathologists by 55%.

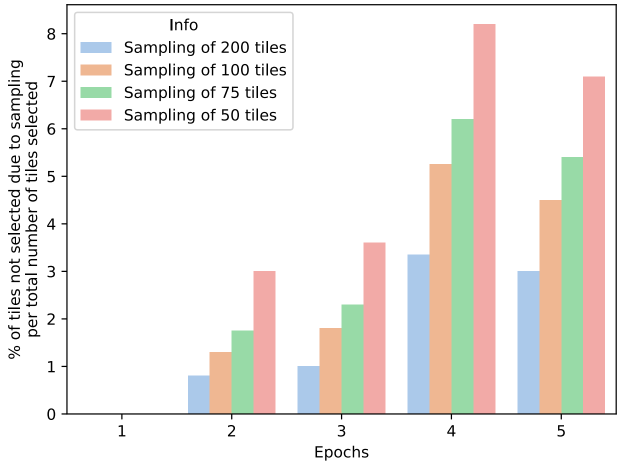


Fig. 1 | Tile sampling impact on information loss. Percentage of tiles not selected due to sampling with different thresholds, over the first four inference epochs. The blue bar represents a sampling strategy that retains 200 tiles per slide, the orange bar is for a strategy that retains 100 tiles, the green bar represents a strategy that retains 75 tiles and finally the strategy represented by the red line retains 50 tiles per slide.

Our work aims to further contribute to the landscape of computer-aided diagnosis (CAD) systems for colorectal pathology, addressing current hurdles and limitations: - The high volume of data needed, in addition to the massive resolution of the images, creates a significant bottleneck of DL approaches that extract patches from the whole slides. Hence, we introduce an efficient sampling approach that is performed once without sacrificing predictive performance on the classification. Leveraging the domain knowledge introduced in the data, by the expert pathologists, in the form of annotations at the pixel level, the model is capable of predicting pseudo-labels for the non-annotated samples. Leveraging these new pseudo-labels, we can discard tiles with the least meaningful pseudo-labels, resulting in 6 × less tiles while retaining most of the important information. This process is preceded by a supervised learning step using the pixel level annotations where the model learns how to create the pseudo-labels for the sampling step. After, the sampling is followed by a weakly-supervised approach on the reduced set of slides and using only slide labels. Our dataset contains, approximately, 10,500 high-quality slides from IMP Diagnostics. A large part of this dataset is publicly available[30], with corresponding case diagnostic labels (making it one of the largest colorectal samples (CRS) datasets available to date). We validate our proposed model in two different external datasets that vary in quality, country of origin and laboratory, ensuring its generalisation capability and robustness. Importantly, in order to bring this CAD system into production, and to infer its usefulness within clinical practice, we developed a prototype, with explainable predictions (visual maps), that was tested and evaluated by pathologists.

To summarise, in this paper we propose a novel dataset with more than thirteen million tiles, a sampling approach to reduce the difficulty of using large datasets, an accurate DL model that is trained with mixed supervision, is evaluated on four datasets, and finally incorporated in a prototype that provides a simple integration in clinical practice and visual explanations of the model's predictions. This way, we are a step closer to making CAD tools a reality for colorectal diagnosis.

Results

In this section, the results are organised to first demonstrate the effectiveness of sampling, followed by an evaluation of the model in the two internal datasets (CRS10K and the prototype dataset), and in the external datasets.

On the effectiveness of sampling

To find the most suitable threshold for sampling the tiles used in the weakly supervised training, we evaluated the percentage of relevant tiles that would be left out of the selection, if the original set was reduced to 75, 100, 150 or 200 tiles, over the first five inference epochs. A tile is considered relevant if it shares the same label as the slide, or if it would take part in the learning process in the weakly-supervised stage. As it is possible to see in Fig. 1, if we set the maximum number of tiles to 200 after the second loop of inference,

Table 1 | Model performance comparison with and without tile sampling of the top 200 tiles from the first inference iteration

Sampling	Best ACC at 5th epoch	Best ACC at 10th epoch	Best QWK at 5th epoch	Best QWK at 10th epoch
No	84.94% ± 2.20	86.42% ± 2.11	0.809 ± 0.024	**0.829 ± 0.023**
Train	85.43% ± 2.18	86.82% ± 2.08	0.817 ± 0.024	0.828 ± 0.023
Train and Val.	**86.12%** ± **2.13**	**86.92%** ± **2.08**	**0.824 ± 0.023**	**0.829 ± 0.023**

Compared the best epoch of the initial five epochs and of the initial ten epochs. Validation is represented as Val and the best results are in bold.

Table 2 | Model performance evaluation on the CRS10K test set

Method	ACC	Binary ACC	Sensitivity
iMIL4Path	91.33% ± 1.84	97.00% ± 1.11	**0.997 ± 0.004**
Ours (CRS4K)	89.44% ± 2.01	96.11% ± 1.26	**0.997 ± 0.004**
Ours (CRS10K) wo/ Agg	**93.44%** ± **1.62**	**97.78%** ± **0.96**	0.996 ± 0.005
Ours (CRS10K) w/ Agg	90.67% ± 1.90	97.55% ± 1.01	0.985 ± 0.009

The binary accuracy is calculated as NNeo vs all. Accuracy is represented as (ACC). In bold are the best results per column.

Table 3 | $\mathcal{X}^2$ and *p* value computed using the McNemar's test for the models evaluated on the Test set

Method	iMIL4Path	Ours (CRS4K)	Ours (CRS10K) wo/ Agg
iMIL4Path	–	1.82 (0.177)	1.92 (0.166)
Ours (CRS4K)	1.82 (0.177)	–	**6.94 (0.008)**
Ours (CRS10K) wo/ Agg	1.92 (0.166)	**6.94 (0.008)**	–

If $\mathcal{X}^2$ is > than 3.84 the difference between two methods is statistically significant. These statistically significant differences are highlighted in bold. *P* value, under parentheses, is computed by calculating the area under the PDF of the chi squared distribution to the right of $\mathcal{X}^2$.

we would discard only 3.5% of the potentially informative tiles, in the worst-case scenario. On the other side of the spectrum, a more radical sampling of only 50 tiles would lead to a cut of up to 8%.

Moreover, to assess the impact of this sampling on the model's performance, we also evaluated the accuracy and the QWK with and without sampling the top 200 tiles after the first inference iteration (Table 1). This evaluation considered sampling applied only to the training tile set, and to both the training and validation tile sets. As can be noticed, the performance is not degraded and the model is trained in a much faster way. In fact, using the setup previously mentioned, the first epoch of inference, with the full set of tiles takes 28h to be completed, while from the second loop the training time decreases to only 5h per epoch. Without sampling, training the model for 50 epochs would take around 50 days, whereas with sampling it takes around 10.

CRS10K and prototype

CRS10K test set and the prototype dataset were collected through different procedures. The first followed the same data collection process as the complete dataset, whereas the second originated from routine samples. Thus, the evaluation of both these sets is done separately.

The first experiment was conducted on the CRS10K test set. As expected, the steep increase in the number of training samples led to a significantly better algorithm in this test set. Initially, the model trained on the CRS10K correctly predicted the class of 819 out of 900 samples. For the wrong 81 cases, the pathologists performed a blind review and found that the algorithm was indeed correct in 22 of them. This led to a correction in the labels of the test set, and the appropriate adjustment of the metrics. In Table 2, the performance of the different algorithms is presented. CRS10K outperforms the other approaches by a reasonable margin.

Using the McNemar's test, it was shown that there were significant different performances between the proposed model trained on CRS10K data and the model trained on CRS4K with a *p* value of 0.008 (Table 3). The differences between the proposed methods trained on CSR10K and CSR4K, and iMIL4Path are not statistically significant with *p* value of 0.166 and 0.177 respectively. We further applied the aggregation proposed by ref. 31 to the proposed method trained on CRS10K, but without gains in performance. Despite being trained on the same dataset, iMIL4Path and the proposed methodology trained on CRS4K, they utilise different splits for training and validation, as well as different optimisation techniques due to the deterministic approach.

A more in-depth inspection of the performance considering the different errors is shown in Fig. 2, where the precision-recall curves for the three models is shown. Moreover, the F1-Score is also included, which shows that the most balanced model is the one that we proposed.

In addition to examining quantitative metrics, such as the accuracy of the model, we extended our study to include an analysis of the confidence in the model when it correctly predicts a class and when it makes an incorrect prediction. To this end, we recorded the confidence of the model for the predicted class and divided it into the set of correct and incorrect predictions. These were then used to fit a kernel density estimator. Figure 3 shows the density estimation of the confidence values for the three different models. It is worth noting that, when correct, the model trained on the CRS10K, returns higher confidence levels as shown by the shift of its mean towards values close to one. On the other hand, the confidence values of its incorrect predictions decrease significantly, and although it does not present the lowest values, it shows the largest gap between correct and incorrect means.

When tested on the prototype data ($n = 100$), the importance of a higher volume of data remains visible (Table 4). Nonetheless, the performance of iMIL4Path[31] approach is comparable to the proposed approach trained on CRS10K. It is worth noting that the latter achieves better performance on the binary accuracy at the cost of a decrease in sensitivity. In other words, the capability to detect negatives increases significantly.

The McNemar's test did show significant differences between any of the methods (Table 5). Similar performance drops were linked with the introduction of aggregation.

Despite similar results, the confidence of the model in its predictions is distinct in all three approaches, as seen in Fig. 4. The proposed approach when trained on the CRS10K dataset has a larger density on values close to one when the predictions are correct, and the mean confidence of those predictions is, once more, higher than the other approaches. However, especially when compared to the proposed approach trained on the CRS4K, the confidence of wrong predictions is also higher. It can be a result of a larger set of wrong predictions available on the latter approach. Nonetheless, the steep increase in the density of values closer to one further indicates that there is room to explore other effects of extending the number of training samples, besides benefits in quantitative metrics.

Domain generalisation evaluation

To ensure the generalisation of the proposed approach across external datasets, we have evaluated their performance on TCGA and PAIP datasets. Moreover, we conducted a similar analysis to both of these datasets, as the one done for the internal datasets.

From the two datasets, PAIP is arguably the closest to CRS10K. It contains similar tissue, despite its colour differences. The performances of the proposed approaches were expected to match the performance of iMIL4Path in this dataset. However, it did not happen for the version trained on the CRS4K dataset, as seen in Table 6. A possible explanation concerns

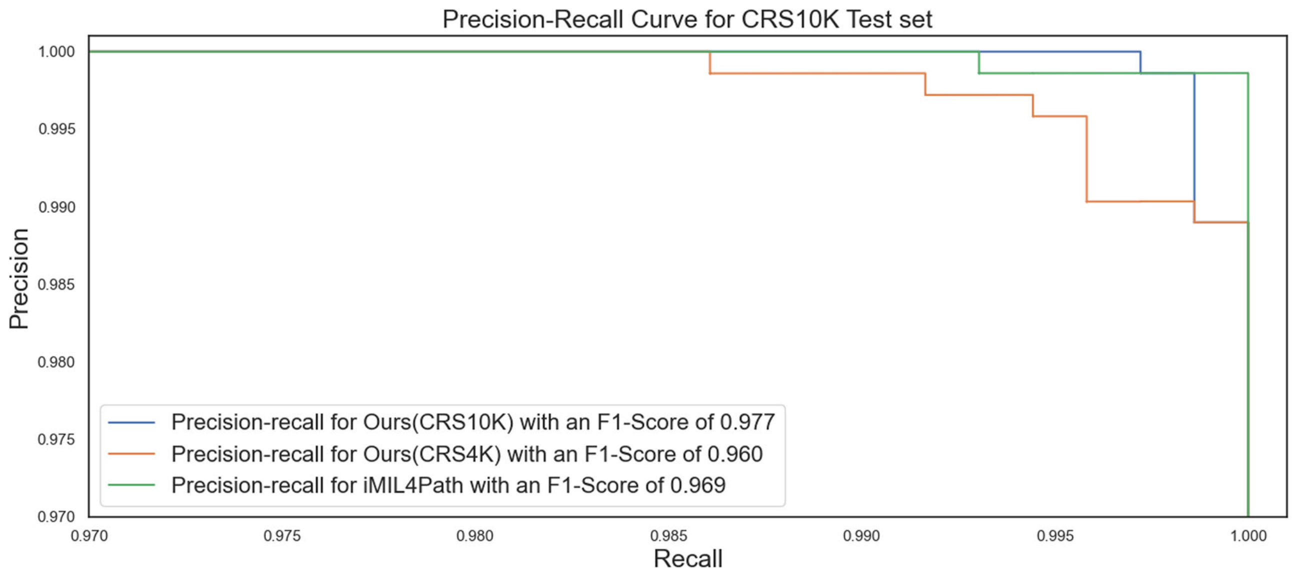


Fig. 2 | Precision-recall curve on the on the CRS10K test set. For the three distinct models, we have calculated the Precision-recall curve on this dataset. Includes an indication of the F1-Score for each of the different models. The blue line represents the curve of Our method when trained on CRS10K, while the orange line shows the same method when trained on CRS4K. The green line is the curve of iMIL4Path.

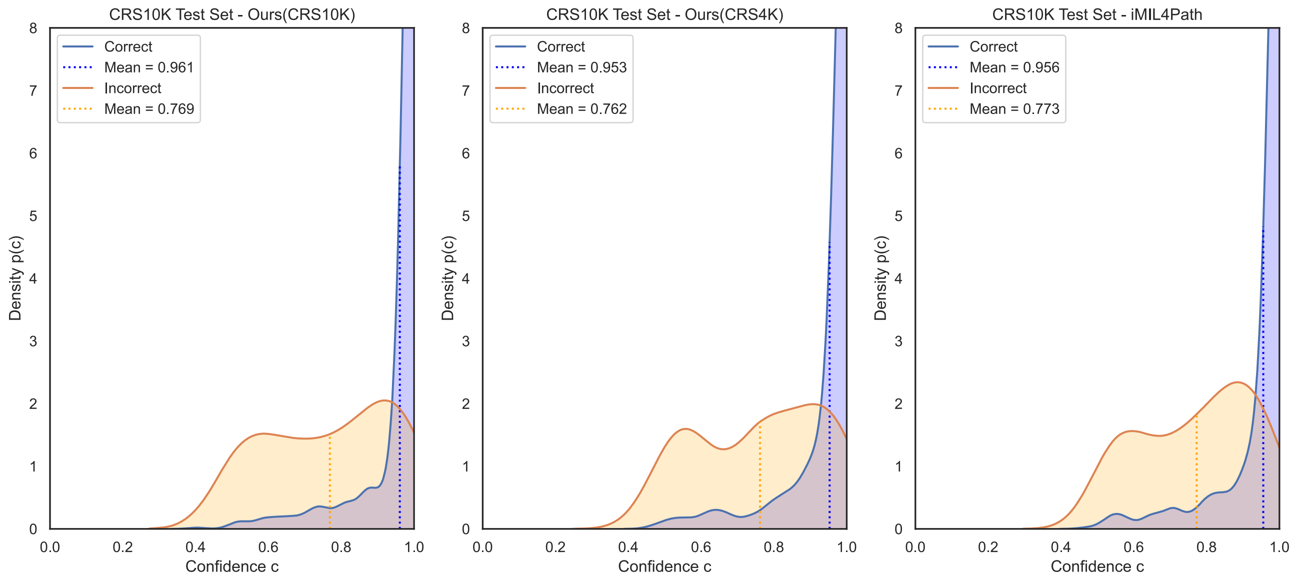


Fig. 3 | Confidence analysis for correct and incorrect predictions on the CRS10K test set. Kernel density estimation of the confidences of correct and incorrect predictions performed on the three-class classification problem by three distinct models on the CRS10K test set. The plots represent, from left to right, the proposed method trained on CRS10K, the proposed method trained on CRS4K and iMIL4Path. In each plot, the blue line defines the density function of the correct samples and the blue dashed line the mean confidence of those samples. On the other hand, the orange solid and dashed lines represent the same for incorrect predictions.

potential overfitting to the training data potentiated by an increase in the number of epochs of fully and weakly supervised training, a slight decrease in the tile variability in the latter approach, and a smaller number of samples when compared to the version trained on CRS10K. This version, trained on the larger set, mitigates the problems of the other method due to a significant increase in the training samples. Moreover, it is worth noting that in all three approaches, the errors corresponded only to a divergence between low and high-grade cases, with no non-neoplastic cases being classified as high-grade or vice-versa. As in previous sets, the version trained on the CRS10K dataset outperforms the remaining approaches. Using aggregation in this dataset leads to a discriminative power to distinguish between high- and low-grade lesions that is close to random.

The McNemar's test indicated a significant difference in the performance difference between the model trained on CRS10K and the one trained on the CRS4k (*p* value of 0.00), and between the latter and iMIL4Path (*p* value of 0.00). However, there was no significant difference between iMIL4Path and the former with a *p* value of 1.00 (Table 7) The confidence of the model was also calculated for this dataset (Supplementary Fig. 2), showing a visible shift towards higher values of confidence in the proposed approach trained on the CRS10K when compared to the method of iMIL4Path. The version trained on CRS4K showed very little separability between the confidence of correct and incorrect predictions.

The TCGA dataset has established itself as the most challenging for the proposed approaches. Besides the expected differences in colour and other elements, this dataset is mostly composed of resection samples, which are not present in the training dataset. As such, this presents itself as an excellent dataset to assess the capability of the model to handle these different types of samples. Both iMIL4Path and the proposed method trained on CRS4K have shown substantial problems in correctly classifying TCGA slides, as shown in Table 8. This can be explained as in the TCGA dataset the majority of the high-grade lesions exhibit an invasive component, and the morphology of the tumoral lesions is altered with the invasiveness. Also, features like an abundance of desmoplastic stroma tend to manifest more prominently in the deeper regions of the tumour, as opposed to the superficial sections typically obtained through biopsy/polypectomy. These aspects also hold relevance in explaining the comparatively inferior outcomes observed in the TCGA dataset. Despite having a lower performance on the general accuracy, the binary accuracy shows that our proposed method trained on CRS4K has much lower misclassification errors regarding the classification of high-grade samples as normal, demonstrating higher robustness of the new training approach against errors with a gap of two classes. As with other datasets, the proposed approach trained on CRS10K shows better results, this time by a significant margin with no overlapping between the confidence intervals.

This was further confirmed by the McNemar's test which once more highlighting the better performance of the proposed model with p values of 0.00 when compared to either iMIL4Path or the same model trained on

Table 4 | Model performance evaluation on the prototype test set

Method	ACC	Binary ACC	Sensitivity
iMIL4Path	**89.00%** ± **6.13**	96.00% ± 3.84	**1.000** ± **0.000**
Ours (CRS4K)	85.00% ± 6.99	93.00% ± 5.00	**1.000** ± **0.000**
Ours (CRS10K) wo/ Agg	**89.00%** ± **6.13**	**98.00%** ± **2.74**	0.986 ± 0.026
Ours (CRS10K) w/ Agg	85.00% ± 6.99	**98.00%** ± **2.74**	0.986 ± 0.026

Accuracy is represented as (ACC). The binary accuracy is calculated as NNeo vs all. In bold are the best results per column.

Table 5 | $\mathcal{X}^2$ and p value computed using the McNemar's test for the models evaluated on the Prototype set

Method	iMIL4Path	Ours (CRS4K)	Ours (CRS10K) wo/ Agg
iMIL4Path	–	0.13 (0.718)	0.00 (1.000)
Ours (CRS4K)	0.13 (0.718)	–	0.30 (0.584)
Ours (CRS10K) wo/ Agg	0.00 (1.000)	0.30 (0.584)	–

If $\mathcal{X}^2$ is > than 3.84 the difference between two methods is statistically significant. These statistically significant differences are highlighted in bold. P value, under parentheses, is computed by calculating the area under the PDF of the chi squared distribution to the right of $\mathcal{X}^2$.

Table 6 | Model performance evaluation on the PAIP test set

Method	ACC	Binary ACC	Sensitivity
iMIL4Path	99.00% ± 1.95	**100.00%** ± **0.00**	**1.000** ± **0.000**
Ours (CRS4K)	69.00% ± 9.06	**100.00%** ± **0.00**	**1.000** ± **0.000**
Ours (CRS10K) wo/ Agg	**100.00%** ± **0.00**	**100.00%** ± **0.00**	**1.000** ± **0.000**
Ours (CRS10K) w/ Agg	52.00 ± 9.79	**100.00%** ± **0.00**	**1.000** ± **0.000**

The binary accuracy is calculated as NNeo vs all. Accuracy is represented as (ACC). In bold are the best results per column.

Table 7 | $\mathcal{X}^2$ and p value computed using the McNemar's test for the models evaluated on the TCGA set

Method	iMIL4Path	Ours (CRS4K)	Ours (CRS10K) wo/ Agg
iMIL4Path	–	0.04 (0.839)	**26.26 (0.000)**
Ours (CRS4K)	0.04 (0.839)	–	**31.03 (0.000)**
Ours (CRS10K) wo/ Agg	**26.26 (0.000)**	**31.03 (0.000)**	–

If $\mathcal{X}^2$ is > than 3.84 the difference between two methods is statistically significant. These statistically significant differences are highlighted in bold. P value, under parentheses, is computed by calculating the area under the PDF of the chi squared distribution to the right of $\mathcal{X}^2$.

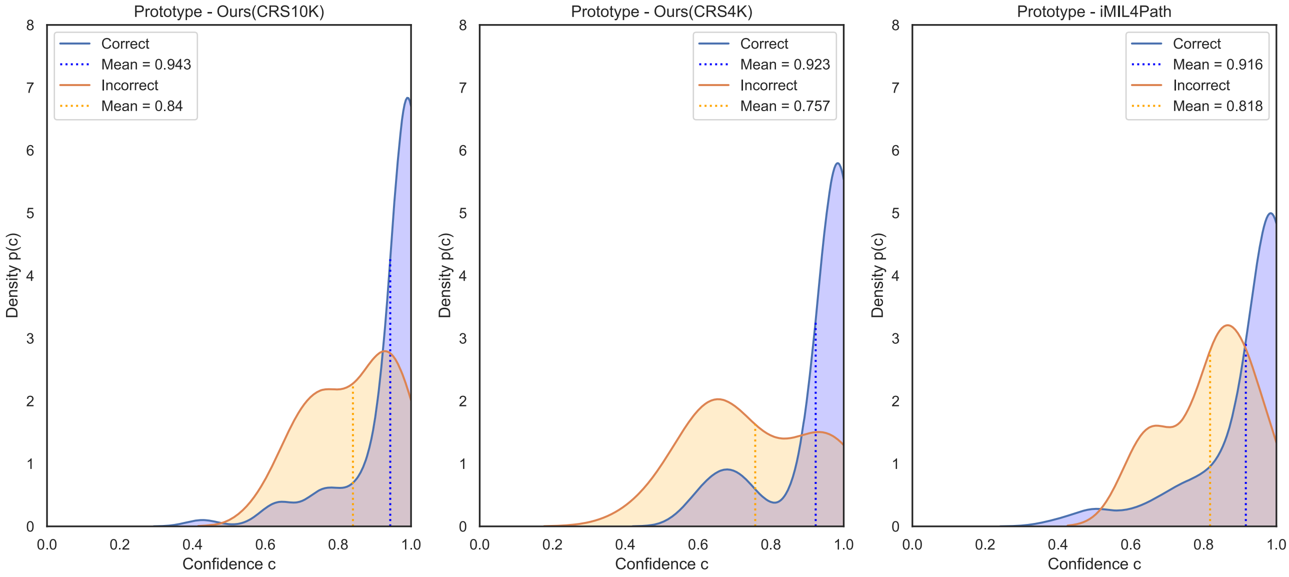


Fig. 4 | Confidence analysis for correct and incorrect predictions on the Prototype set. Kernel density estimation of the confidences of correct and incorrect predictions performed on the three-class classification problem by three distinct models on the prototype set. The plots represent, from left to right, the proposed method trained on CRS10K, the proposed method trained on CRS4K and iMIL4Path. In each plot, the blue line defines the density function of the correct samples and the blue dashed line the mean confidence of those samples. On the other hand, the orange solid and dashed lines represent the same for incorrect predictions.

CRS4K. The lack of significance between the differences of iMIL4Path and the model trained on CRS4K (p value of 0.839) further emphasises the capability of the sampling strategy to retain the results (Table 9). The confidence predictions for the three models were also assessed (Supplementary Fig. 3), indicating a behaviour in line with the accuracy-based performance. Also, the model trained on CRS10K showed a shift of wrong predictions' confidence towards smaller values, indicating that it is possible to quantify the uncertainty of the model and avoid the majority of the wrong predictions. In other words, when the uncertainty is above a learnt threshold, then the model refuses to make any prediction which is extremely useful in models designed as a second opinion system.

Table 8 | Model performance evaluation on the TCGA test set

Method	ACC	Binary ACC	Sensitivity
iMIL4Path	71.55% ± 5.80	80.60% ± 5.05	0.805 ± 0.051
Ours (CRS4K) wo/ Agg	70.69% ± 5.86	98.71% ± 1.45	0.991 ± 0.012
Ours (CRS10K) wo/ Agg	**84.91% ± 4.61**	**99.13% ± 1.19**	**0.996 ± 0.008**
Ours (CRS10K) w/ Agg	69.83% ± 5.91	97.41% ± 2.04	0.983 ± 0.017

The binary accuracy is calculated as NNeo vs all. Accuracy is represented as (ACC). In bold are the best results per column.

Table 9 | $\mathcal{X}^2$ and p value computed using the McNemar's test for the models evaluated on the PAIP set

Method	iMIL4Path	Ours (CRS4K)	Ours (CRS10K) wo/ Agg
iMIL4Path	–	**26.28 (0.000)**	0.00 (1.000)
Ours (CRS4K)	**26.28 (0.000)**	–	**29.03 (0.000)**
Ours (CRS10K) wo/ Agg	0.00 (1.000)	**29.03 (0.000)**	–

If $\mathcal{X}^2$ is > than 3.84 the difference between two methods is statistically significant. These statistically significant differences are highlighted in bold. P value, under parentheses, is computed by calculating the area under the PDF of the chi squared distribution to the right of $\mathcal{X}^2$.

Reject option

Following the confidence analysis previously introduced, we further explore the possibility of rejecting some samples that represent lower levels of confidence.

On the CRS10K test set, the reject rate correlates with an improved performance on all the algorithms displayed on Fig. 5. On our proposed algorithm we achieve 1.5% points improvement at a rejection rate of 4% resulting in a accuracy of approximately 95%. Moreover, if we reject 16% of the samples (i.e., still reducing the pathologist workload by 84%) the accuracy of the model is of 97.27%. With a rejection rate of 50%, which is less beneficial to pathologists, the accuracy would rise to 99.48%. The possibility of a reject option was also explored for the prototype dataset and TCGA dataset (Supplementary Figs. 4 and 5). We have not conducted this study on the PAIP dataset because the performance was already around 100% in two of the main algorithms evaluated.

Prototype usability in clinical practice

As it is currently designed, the algorithm works preferentially as a "second opinion", allowing the assessment of difficult and troublesome cases, without the immediate need for the intervention of a second pathologist. Due to its "user-friendly" interface, the cases can be easily introduced into the system and results are rapidly shown and accessed. Also, by presenting visualisation maps, the pathologist is able to compare his own remarks to those of the algorithm itself, towards a future "AI-assisted diagnosis", where the pathologist has a pivotal role. Further, the prototype allows for user feedback (agreeing or not with the model's proposed result), which can be integrated into further updates of the software and could be leveraged in the future to feature active learning. Also interesting, would be the possibility of using such a prototype as a triage system on a pathologist's daily workflow by running upfront, before the pathologist checks the cases. Signalling the cases that would need to be more urgently observed (namely high-risk lesions) would allow the pathologists to prioritise their workflow. Further, by providing a previous assessment of the cases, it could also contribute to enhancing the

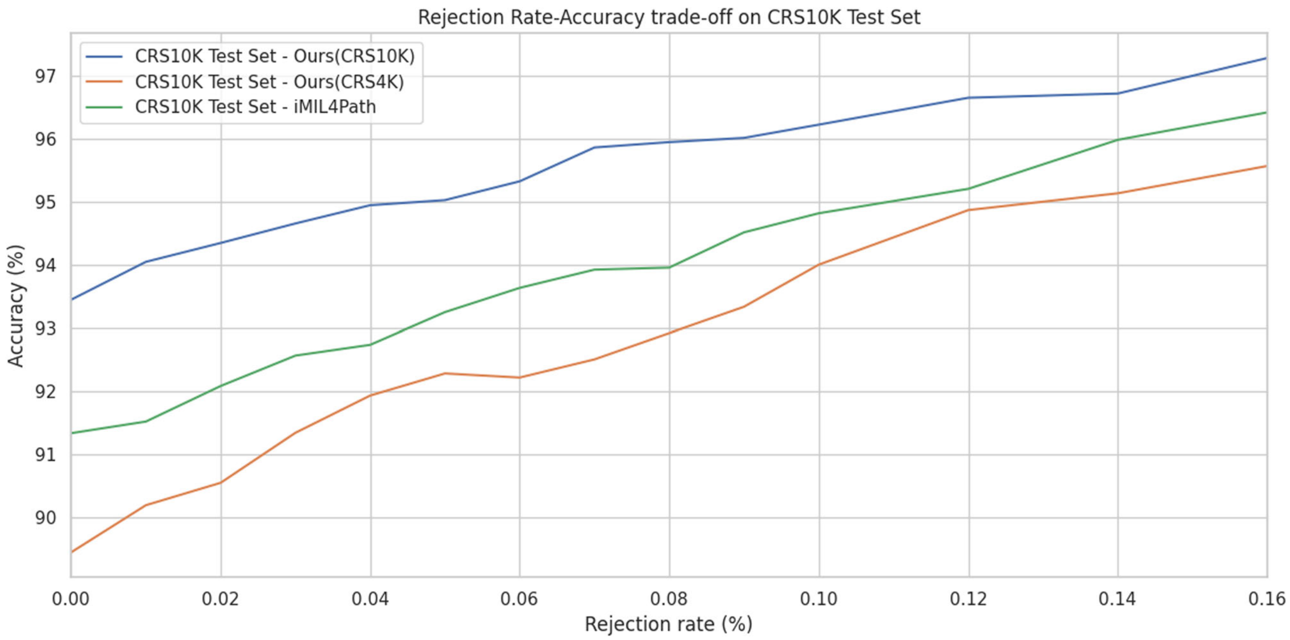


Fig. 5 | Accuracy-vs-Rejection-rate for the models evaluated on the CRS10K test set. Relation between the accuracy and the percentage of samples not classified by the model. Both axes are in percentage. The blue line represents Our method when trained on CRS10K, while the orange line shows the same method when trained on CRS4K. The green line is for iMIL4Path.

pathologists' efficiency. As such, this is one of our future work objectives. Presently, there is no recommendation for dual independent diagnosis of colorectal biopsies (contrary to gastric biopsies, where, in cases that surgical treatment is considered, it is recommended to obtain a pre-treatment diagnostic second opinion[32]), but, in the future, this can also become a requirement for colorectal samples. As such, CAD systems to assist pathologists in colorectal diagnosis can become even more important, being their relevance further amplified due to the worldwide shortage of pathologists.

Discussion

In this work, we have proposed a redesig of the previous MIL methodology applied to CRC diagnosis. We aimed to develop a scalable, efficient and interpretable solution for this task. For this, we have worked on a mixed supervision approach to design a sampling strategy, which utilises the knowledge from the full supervision training as a proxy to tile utility. Secondly, we studied the confidence that the model shows in its predictions. Our target in this latter part was to infer the possibility of using a reject option based on the confidence of the model. The results have shown that this confidence has the potential to be a resource to quantify uncertainty and avoid wrong predictions on low-certainty scenarios. The model was entirely integrated within a web-based prototype to assist pathologists in their routine work.

The proposed methodology was evaluated on several datasets, including two external sets. Through this evaluation, it was possible to infer that the performance of the proposed methodology benefits from a larger dataset and surpasses the performance of previous state-of-the-art models that were evaluated on this benchmark. As such, and given the excelling results that originated from the increase in the dataset, we are also publicly releasing the majority of the CRS10K dataset WSIs and case diagnostic labels, one of the largest publicly available colorectal datasets composed of H&E images in the literature, including the test set for the benchmark of distinct approaches across the literature[30].

Our findings have several noteworthy elements. First, we have shown that despite the ability to lead to better models, increasing the dataset size can be a double-edged sword due to the computational requirements of MIL solutions. With this in mind, and while conducting this study on one of the largest datasets for CRS, we have devised a sampling strategy that seems to minimise the information lost during training, leading to a comparable performance at 6x less processing time. Our method has also demonstrated the power of having a small portion of the dataset annotated to initialise the weights of the MIL model. We have further shown, that models trained on larger datasets seem to approximate more stable confidence distributions, leading to the possibility of using a reject option to comply with clinical requirements on the performance of the model. Finally, we have highlighted an interpretability method that is integrated into our prototype and supports the decision process of pathologists.

Within the evaluation of datasets collected on similar configurations to the training data, the performance of the proposed model represents a step towards better algorithms for colorectal pathology. The high sensitivity did not compromise the overall accuracy. On datasets that originate from other centres and scanners, the performance was around 100% of accuracy in one and around 85% on the other, with the latter coming from completely different tissue samples. The comparison with other studies is highly limited by the test data. In our scenario, we have tested in a pool of 1332 samples, which is larger than several studies' train sets. As we are releasing our test dataset, further research methods can be easily compared through it.

We can note that the strong performance of the model, aligned with the prototype and the prediction maps, supports the utilisation of such a system as a second opinion within the routine process in a pathology lab, assisting pathologists in their daily routine, ensuring higher quality and, thus, better patient care.

Nonetheless, the proposed algorithm still has potential for improvement. We aim to include the recognition of serrated lesions, to distinguish normal mucosa from significant inflammatory alterations/diseases, to stratify high-risk lesions into high-grade dysplasia and invasive carcinomas and to identify other neoplasia subtypes. This will enable the prototype to be used upfront in the future. Further, we would like to leverage the model to be able to evaluate also surgical specimens. Another relevant step will be the merge of our dataset and external ones for training, besides only testing it on external samples. This will enhance its generalisation capabilities and provide a more robust system. Lastly, we intend to measure the "user experience" and feedback from the pathologists, by its gradual implementation in general laboratory routine work. The following goals comprise a more extensive evaluation of the model across more scanner brands and labs. We also want to promote certain mechanisms that would allow for more direct and integrated uncertainty estimation. We have also been looking towards aggregation methods, but, since in the majority of them there is an increased risk of false negatives, we have work to do in that research direction.

Methods

In this section, after defining the problem at hand, we introduce the proposed dataset used to train, validate and test the model, the external datasets to evaluate the generalisation capabilities of the model and the preprocessing pipeline. After, we describe in detail the methodology followed to create the deep learning model and to design the experiments. Finally, we also detail the clinical assessment and evaluation of the model. This section also includes a description of the two main bottlenecks that can affect this type of systems. The first is the difficulty of collecting data and having large amounts of data annotated by experts. The second, which becomes apparent only after the first bottleneck is overcame, is the difficulty of scaling these systems as we increase the size of the training data. Without solving the latter problem, it would be impossible to take full advantage of the benefits of collecting large amounts of data.

Problem definition

Digitised CRC histological samples have large dimensions, which are far bigger than the traditional images used in medical or general computer vision problems. Labelling such images is expensive and highly dependent on the availability of expert knowledge. This limits the availability of WSIs, and, in scenarios where these are available, meaningful annotations are usually lacking. On the other hand, it is easier to label the dataset at the slide level. The inclusion of detailed spatial annotations on approximately 10% of the dataset has been shown to positively impact the performance of deep learning algorithms[15,31].

To fully leverage the potential of spatial and slide labels, we propose a deep learning pipeline, based on previous approaches[15,31], using mixed supervision. Each slide, $\mathcal{S}$ is composed of a set of tiles $\mathcal{T}_{s,n}$, where s represents the index of the slide and $n \in \{1, \cdots, n_s\}$ the tile number. Furthermore, there is an inherent order in the grading used to classify the tiles into one of the $C_{s,n}^{(k)}$ classes, which represents a variation in severity. We define $C_s^{(k)}$ and $C_{s,n}^{(k)} \in$ {"Non-Neoplastic", "Low-Grade Lesion", "High-Grade Lesion"}, and the label of each slide s corresponds to the index (k) of their class. We further define the output of the model as $\hat{y}_{s,n}$ where $\hat{y}_{s,n}(k)$ is the model estimation for $P(C_{s,n}^{(k)})$. The final prediction of the model is defined as $\hat{C}_s \in \{1, .., K\}$ for a slide prediction and $\hat{C}_{s,n} \in \{1, .., K\}$ for a tile prediction, where $K = 3$. The latter derives from $\arg\max_k P(C_{s,n}^{(k)})$. For fully supervised learning, only strongly annotated slides are useful, and for those, the class of each tile $C_{s,n}^{(k)}$ is known. The remaining slides are deprived of these detailed labels, hence, they can only be leveraged by training algorithms with weakly supervision using slide level labels as $C_s^{(k)}$. To be used by these algorithms, the weakly annotated slides have only a single label for the entire bag (set) of tiles, as seen in Fig. 6. Following the order of the labels and the clinical knowledge, we assume that the predicted slide label $\hat{C}_s$ is the most severe case of the tile labels:

$$\hat{C}_s = \max_n\{\hat{C}_{s,n}\}.$$

In other words, if there is at least one tile classified as containing high-grade dysplasia, then the entire slide that contains the tile is labelled/

annotated accordingly. On the other end of the spectrum, if the worst tile is labeled as non-neoplastic, then it is assumed that there is no dysplasia in the entire set of tiles. This is a generalisation of multiple-instance learning (MIL) to an ordinal classification problem, as proposed by ref. 15.

Datasets

The spectrum of large-scale CRC/CRS datasets is increasing due to the contributions of several researchers[30]. Two datasets that have been recently introduced in the literature are the CRS1K[15] and CRS4K[31] datasets from our research group. Since the latter is an extension of the former with roughly four times more slides, it will be the baseline dataset for the remaining of this document. We further extend this with the CRS10K dataset, which contains 9.26x and 2.36x more slides than CRS1K and CRS4K, respectively. We gathered our data retrospectively from IMP Diagnostics' archive, sequentially selecting all cases that matched the study's diagnostic categories. Thus, we performed consecutive sampling and our cases are a representative sample of the study's population, namely regarding pathology distribution across sex and age in the population. Similarly, the number of tiles is multiplied by a factor of 12.2 and 2.58 (Table 10). This volume of slides is translated into an increase in the flexibility to design experiments and infer the robustness of the model. Thus, the inclusion of a test set separated from the validation set is now facilitated. All procedures were in accordance with the ethical standards of the 1964 Helsinki declaration and its later amendments and comparable ethical standards. All data was anonymized and data collection and usage were performed in accordance with the General Data Protection Regulation (GDPR) and national laws and regulations. Ethical approval was waived by the local Ethics Committee of INESC TEC in view of the retrospective nature of the study and all the procedures being performed were part of the routine care.

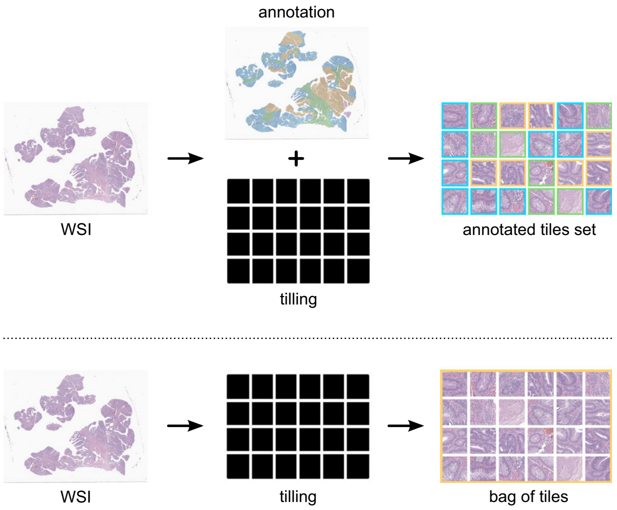


Fig. 6 | Overview of the proposed problem definition. Problem definition as a fully supervised task (on top), and as a weakly-supervised task (bottom).

The set is composed of colorectal biopsies and polypectomies (excluding surgical specimens). CRS10K slides are labelled according to three main categories: non-neoplastic (NNeo), low-grade lesions (LG), and high-grade lesions (HG). The first, contains normal colorectal mucosa, hyperplasia and non-specific inflammation. LG lesions correspond to conventional adenomas with low-grade dysplasia. Finally, HG lesions are composed high-grade dysplasia adenomas (including intramucosal carcinomas) and invasive adenocarcinomas. Slides with suspicion or known history of inflammatory bowel disease, infectious diseases, serrated lesions or other polyp types were not included in the dataset.

The slides were digitised with Leica GT450 WSI scanners, at 0.26 µm/pixel at 40 × magnification. The cases were initially seen and classified (labelled) by one of three pathologists. The pathologist revised and classified the slides, and then compared the result with the initial report diagnosis (which served as a second-grader). If there was a match between both, no further steps were taken. In discordant cases, a third pathologist served as a tie-breaker. Roughly 9% of the dataset (967 slides and over a million tiles) were manually annotated by a pathologist and rechecked by the other, in turn, using the Sedeen Viewer software[33]. For complex cases, or when the agreement for a joint decision could not be reached, a third pathologist reevaluated the annotation.

The CRS10K dataset was divided into train, validation and test sets. The training set includes all the strongly annotated slides, for fully-supervised learning, and a random selection of weakly-annotated samples. The validation set, on the other hand, consists of only weakly-annotated slides. Finally, the test set was selected from the new data added to extend the previous datasets and only includes weakly-annotated slides. Thus, it is completely separated from the training and validation sets of previous works. The test set, is publicly available, so that future research can directly compare their results and use this set as a benchmark.

Table 10 | Dataset summary, with the number of slides (annotated samples are detailed in parentheses) and tiles distributed by class for all the datasets used in this study

		NNeo	LG	HG	Total
	# slides	300 (6)	552 (35)	281 (59)	1133 (100)
CRS1K dataset[15]	# annotated tiles	49,640	77,946	83,649	211,235
	# non-annotated tiles	–	–	–	1,111,361
	# slides	663 (12)	2394 (207)	1376 (181)	4433 (400)
CRS4K dataset[31]	# annotated tiles	145,898	196,116	163,603	505,617
	# non-annotated tiles	–	–	–	5,265,362
	# slides	1740 (12)	5387 (534)	3369 (421)	10,496 (967)
CRS10K dataset	# annotated tiles	338,979	371,587	341,268	1,051,834
	# non-annotated tiles	–	–	–	13,571,871
CRS Prototype	# slides	28	44	28	100
	# non-annotated tiles	–	–	–	244,160
PAIP[38]	# slides	–	–	100	100
	# non-annotated tiles	–	–	–	97,392
TCGA[37]	# slides	1	1	230	232
	# non-annotated tiles	–	–	–	1,568,584

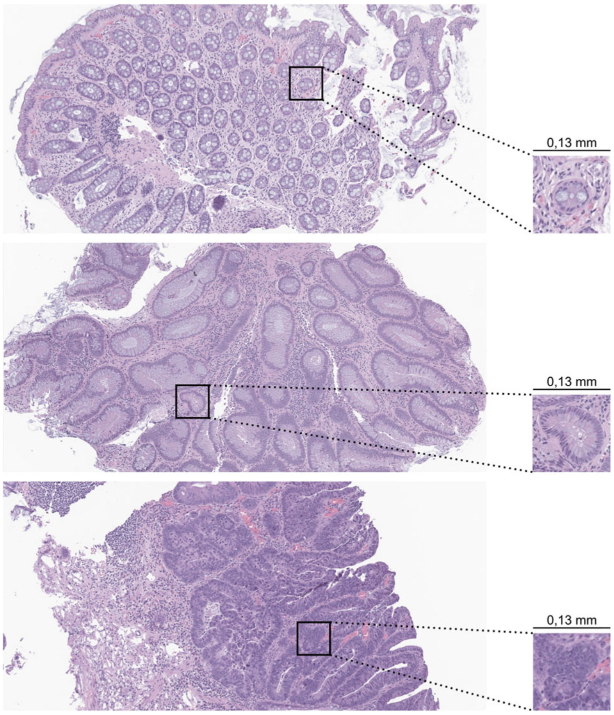


Fig. 7 | Examples of regions and a sample tile with 512 × 512 pixels (40 × magnification). The represented classes are: non-neoplastic (on top), low-grade dysplasia (on the middle) and high-grade dysplasia (on the bottom) with a width and height of 0.13 milimeters (mm) each.

Of note, when collected from routine archives, the slides can be digitised with duplicated tissue areas. Hence, the workflow for the automatic diagnostic also included an automatic fragment detection and counting system, to avoid repeated and lower quality fragments[34].

Furthermore, as detailed in the following sections, this work comprises the development of a fully-functional prototype to be used in clinical practice. Leveraging this prototype, it was possible to further collect a new set with 100 slides. It differs from the CRS10K dataset, as these cases were actively collected from the current year's routine exams. We argue that this might better reflect the real-world data distribution. Hence, we introduce this set as a distinct dataset to evaluate the robustness of the presented methodology. Differently from the datasets discussed below, the CRS Prototype dataset has a more balanced distribution of the slide labels. Although useful, using the fragment counting and selection algorithm for the evaluation could potentiate the propagation of errors from one system to another. Thus, in this evaluation, we did not use the fragment selection algorithm, and as shown in Table 10, the number of tiles per slide doubles when compared to CRS10K, which had its fragments carefully selected.

To evaluate the domain generalisation of the proposed approach, two external datasets, publicly available, were used. The first dataset is composed of samples of the TCGA-COAD[35] and TCGA-READ[36] collections from The Cancer Imaging Archive[37], which are composed in general by resection samples (in contrast to our dataset, composed only of biopsies and polypectomies). Samples containing pen markers, large air bubbles over tissue, tissue folds, and other artefacts affecting large areas of the slide were excluded. The final selection includeded 232 WSIs reviewed and validated by the same pathologists that reviewed our in-house datasets. 230 of those samples were diagnosed as high-grade lesions, whereas the remaining two have been diagnosed as low-grade and non-neoplastic. For this dataset, the specific model of the scanner used to digitise the images is unknown, but the file type (".svs") matches the file type of the training data. The second external dataset contains 100H&E slides from the Pathology AI Platform[38] colorectal cohort. All included cases had a more superficial sampling of the lesions, better comparing with our datasets. All the WSIs in this dataset were digitised with an Aperio AT2 at 20X magnification. Finally, the pathologists' team followed the same guidelines to review and validate all the WSI, which were all classified as high-grade lesions. It is interesting to note that while the PAIP contains significantly fewer tiles per slide, around 973, than the CRS10K dataset, around 1293, the TCGA dataset shows the largest amount of tissue per slide with an average of 6761 tiles as seen in Table 10.

Data pre-processing

H&E slides are composed of two distinct elements, white background and colourful tissue. Since the former is not meaningful for the diagnostic, the preprocessing of these slides incorporates an automatic tissue segmentation with Otsu's thresholding[39] on the saturation (S) channel of the HSV colour space, resulting in a separation between the tissue regions and the background. The result of this step, which receives as input a 32 × downsampled slide, is the mask used for the following steps. Leveraging this previous output, tiles with a dimension of 512 × 512 pixels (Fig. 7) were extracted from the original slide (without any downsampling) at its maximum magnification (40 ×), if they did not include any portion of background (i.e., a 100% tissue threshold was used). Following previous experiments in the literature, our empirical assessment, and the confirmation that smaller tiles would significantly increase the number of tiles and the complexity of the task, 512 × 512 was chosen as the tile size. Moreover, it is believed that 512 × 512 is the smallest tile size that still incorporates enough information to make a good diagnostic with the possibility of visually explaining the decision[15]. The selected threshold of 100% further reduces the number of tiles by not including the tissue at the edges and decreases the complexity of the task, since the model does not see the background at any moment. Due to tissue variations in different images, there is also a different number of tiles extracted per image.

Methodology

The massive size of images, which translates to thousands of tiles per image, allied to a large number of samples in the CRS10K dataset, bottlenecks the training of weakly-supervised models based on multiple instance learning (MIL). Hence, in this document, we propose a mix-supervision approach with self-contained tile sampling to diagnose CRC samples from WSIs. This subsection comprises the methodology, which includes supervised training, sampling and weakly-supervised learning.

Supervised training.

As mentioned in previous sections, spatial annotations are rare in large quantities. However, these include domain information, given by the expert annotator, concerning the most meaningful areas and what are the most and less severe tiles. Thus, they can facilitate the initial optimisation of a deep neural network. As shown in the literature, there has been some research on the impact of starting the training with a few iterations of fully-supervised training[15,40]. We further explore this in three different ways. First, we have 967 annotated slides resulting in more than one million annotated tiles for supervised training. Secondly, attending to the size of our dataset and the need for a stronger initial supervised training, the models are trained for 50 epochs, and their performance was monitored over specific checkpoint epochs. Finally, we explore this pre-trained model as the main tool to sample useful tiles for the weakly-supervised task.

Tile sampling.

Our scenario presents a particularly difficult condition for scaling the training data. First, let's consider the structure of the data, which consists of, on average, more than one thousand tiles per slide. If we carefully analyse Table 10, we can see that the CRS10K dataset and the CRS Prototype contain, when combined, ≈ 13.8 million tiles. These tiles come after preprocessing, as such, tiles containing background are not included. If we further analyse this data, and considering that each tile is of dimension 512 × 512 × 3, then we have ≈ 3.6 trillion pixels per colour channel, or ≈ 10.9 trillion pixels in total. Reference 41 described the difficulties of processing 399 WSI in a single GPU. With the following strategy we processed all the 10928 WSI described in Table 10 utilising a single GPU.

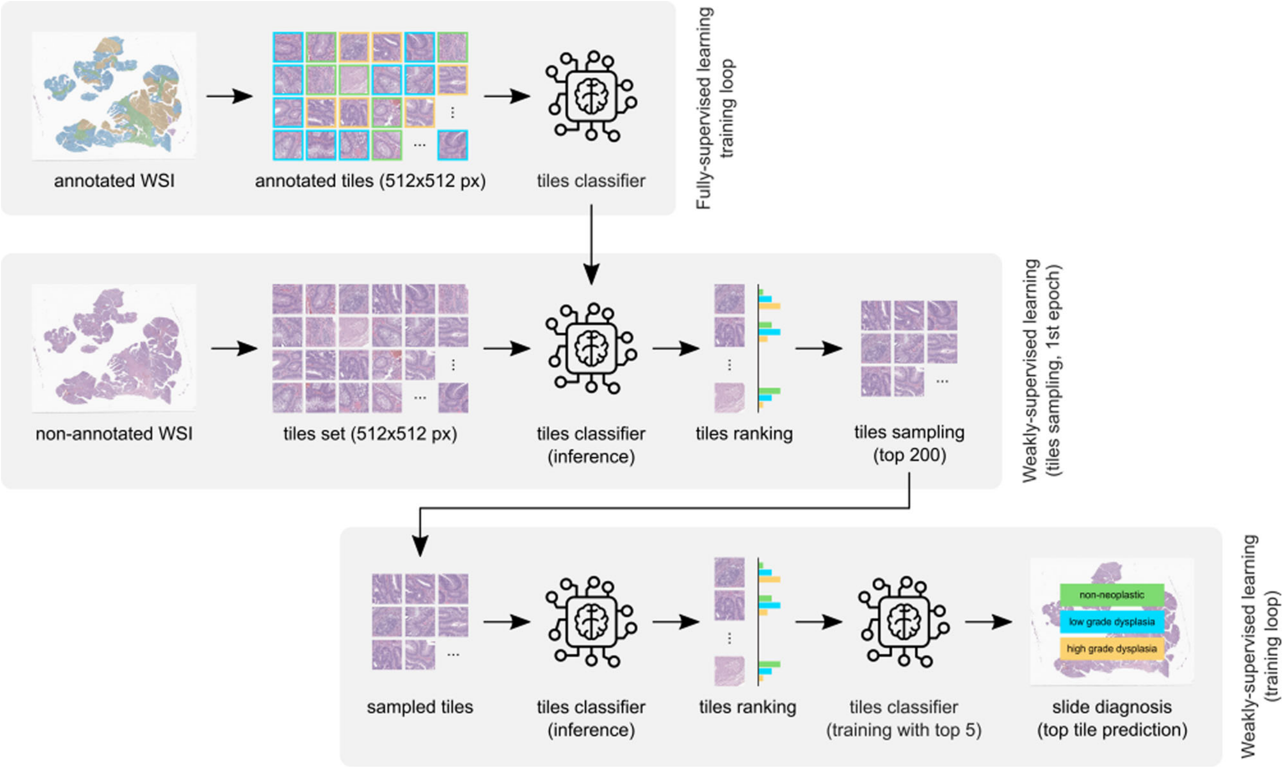

annotated WSI
annotated tiles (512x512 px)
tiles classifier
Fully-supervised learning training loop
non-annotated WSI
tiles set (512x512 px)
tiles classifier (inference)
tiles ranking
tiles sampling (top 200)
Weakly-supervised learning (tiles sampling, 1st epoch)
sampled tiles
tiles classifier (inference)
tiles ranking
tiles classifier (training with top 5)
non-neoplastic
low grade dysplasia
high grade dysplasia
slide diagnosis (top tile prediction)
Weakly-supervised learning (training loop)

Fig. 8 | Scheme for the proposed mixed precision workflow. Overall scheme of the proposed methodology containing the mix-supervision framework that is responsible for diagnosing colorectal samples from WSI. The top layer consists of the fully-supervised stage, the middle layer consists of the sampling strategy and the bottom layer represents the weakly supervised training stage. Tile sizes are in pixels (px).

Within the set of tiles from a slide, some tiles provide meaningful value for the prediction, and others do not add extra information. In other words, for the CRS10K dataset, the extensive, time and energy-consuming process of going through 13 million tiles every epoch can be avoided, and, as result, these models can be trained for more epochs. Nowadays, there is an increasing concern regarding energy and electricity consumption. Thus, these concerns, together with the sustainability goals, further support the importance of more efficient training processes.

Let $\mathcal{T}$ be the original set of tiles, and $\mathcal{T}_s$ be the original set of tiles from the slide *s*, the former is composed by a union of the latter of all the slides (Eq. (1)). We propose to map $\mathcal{T}$ to a smaller set of tiles $\mathcal{M}$ without affecting the overall performance and behaviour of the trained algorithm.

$$\mathcal{T} = \bigcup_{s=1}^{S} \mathcal{T}_s \tag{1}$$

The model trained in a fully supervised task, previously described, provides a good estimation of the utility of each tile. Hence, we utilise the function (Φ) learned by the model to compute the predicted severity of each tile. In other words, Φ represents the supervised model already trained. We select M tiles per slide ($M = 200$ in our experimental setup) utilising a Top-k function (with k set to 200) to be retained for the weakly-supervised training. As indicated by the results presented in the following sections, the value of M was selected in accordance with a trade-off between information lost and training time. This is formalised in Eq. (2).

$$\mathcal{M}_s = \text{Top-k}\,(\Phi(\mathcal{T}_s)) \tag{2}$$

For instance, in the CRS10K dataset, the total number of tiles after sampling would be at most 2,099,200, which represents a reduction of 6.46 × when compared to the total number of slides. Despite this upper bound on the number of tiles, there are WSI samples that contain less than M tiles, and as such, they remain unsampled and the actual total number of tiles after sampling is potentially lower. During the evaluation and test time, there is no sampling.

Weakly-supervised learning. The weakly-supervised learning approach designed for our methodology follows the same principles of recent work[31]. It is divided into two distinct stages, tile severity analysis and training. The former utilises the pre-trained model to evaluate the severity of every tile in a set of tiles. In the first epoch, $\mathcal{T}$, the set of all the tiles in the complete dataset is used. This is possible since the model used to assess the severity in this epoch is the same one used for sampling. Hence, both tasks are integrated with the initial epoch. The following epochs utilise the sampled tile set $\mathcal{M}$ instead of the original set. In other words, the bags (i.e., the representation of the slides) are all truncated to size 200. This overall structure is represented in Fig. 8. Moreover, the weakly-supervised approach leverages only slide labels.

The link between both stages is guided by a slide-wise tile ranking approach based on the expected severity as proposed in[15]. For tile $\mathcal{T}_{s,n}$, the expected severity is defined as

$$\mathbb{E}(\hat{C}_{s,n}) = \sum_{i=1}^{K} i \times \hat{y}_{s,n}(i) \tag{3}$$

where $\hat{y}_{s,n}(i)$ is a random vector of size *K*, which represents the $P(C_{s,n}^{(i)})$ for the tile *n* of the slide *s*. After this analysis, the five most severe tiles are selected from each bag of 200 tiles for training. The number of selected tiles was chosen in accordance with previous studies[31]. These five tiles per bag are used to train the proposed model for one more epoch. An epoch is composed of both stages, which means that the tiles used for training vary across epochs. The slide label is used as the ground truth of all five tiles of that same slide used for network optimisation. For validation and evaluation, only the most severe tile is used for diagnostics. Although it might lead to an increase in false positives, it shall significantly reduce false negatives. Furthermore, we argue that increasing the variability and quantity of data available leads to a better balance between the reduction of these two types of errors.

Reject option

Automatic systems designed to assist pathologists should be high-performing and achieve outstanding values in evaluation metrics. However, it is equally important for these systems to recognise their limitations and defer to expert pathologists in challenging cases. Recognising the importance of this feature, we introduce a reject option to our model. Pathologists can further tune the expected rate of rejection and the performance on a set of metrics to better suit the model to their needs.

The adopted strategy creates the possibility to reject a sample based on the predicted probability of the predicted label. Then, the desired rejection rate is calculated from the percentiles of all confidence values. This approach magnifies the innate capabilities of deep learning systems to be used as a second/third opinion system.

Confidence interval

In order to quantify the uncertainty of a result, it is common to compute the 95 percent confidence interval. In this way, two different models can be easily understood and compared based on the overlap of their confidence intervals. The standard approach to calculating these intervals requires several runs of a single experiment. As we increase the number of runs, our interval becomes narrower. However, this procedure is impractical for the computationally intensive experiments presented in this document. Hence, we use an independent test set to approximate the confidence interval as a Gaussian function[42]. To do so, we compute the standard error (*SE*) of an evaluation metric *m*, which is dependent on the number of samples (*n*), as seen in Equation (4).

$$SE = \sqrt{\frac{1}{n} \times m \times (1 - m)} \tag{4}$$

For the SE computation to be mathematically correct, the metric *m* must originate from a set of Bernoulli trials. In other words, if each prediction is considered a Bernoulli trial, then the metric should classify them as correct or incorrect. The number of correct samples is then given by a Binomial distribution $X \sim (n, p)$, where p is the probability of correctly predicting a label, and *n* is the number of samples. For instance, the accuracy is a metric that fits all these constraints.

Following the definition and the properties of a Normal distribution, we compute the number of standard deviations (*z*), known as a standard score, that can be translated to the desired confidence (*c*) set to 95% of the area under a normal distribution. This is a well-studied value, which is approximately $z \approx 1.96$. This value *z* is then used to calculate the confidence interval, calculated as the product of *z* and *SE* as seen in Equation (5).

$$M \pm z * \sqrt{\frac{1}{n} \times m \times (1 - m)} \tag{5}$$

To infer the statistical significance of the different performance of different classifiers the McNemar's test[43] was used. These statistical tests have been used in the literature for comparison of independent systems and there are several variations[44]. For the McNemar's test the classifiers must be compared in pairs. For each pair, it is necessary to build a contigency table containing four entries: (a) samples misclassified by both; (b) samples misclassified only by the second classifier; (c) samples misclassified only by the first classifier; (d) samples correctly classified by both. The null hypothesis of this test states that the second (b) and third (c) entries have equal probability. $\mathcal{X}^2$ corrected for continuity[45]is calculated as follows:

$$\mathcal{X}^2 = \frac{(|b - c| - 1)^2}{b + c} \tag{6}$$

Using a significance value of 0.95, we can reject the null hypothesis if $\mathcal{X}^2 > 3.841$, which corresponds to the area between 0.05 and $+\infty$ for a Chi-Squared distribution with 1 degree of freedom.

Label correction

The complex process of labelling thousands of WSI with CRC diagnostic grades is a task of increased difficulty. It should also be noted and taken into account that grading colorectal dysplasia is hurdled by considerable subjectivity, so it is to be expected that some borderline cases will be classified by some pathologists as low-grade and others as high-grade. Moreover, as the number of cases increases, it becomes increasingly difficult to maintain perfect criteria and avoid mislabelling. For this reason, we have extended the analysis of the model's performance to understand its errors and its capability to detect mislabelled slides.

After training the proposed model, it was evaluated on the test data. Following this evaluation, we identified the misclassified slides and conducted a second round of labelling. These cases were all blindly reviewed by two pathologists, and discordant cases from the initial ground truth were discussed and classified by both pathologists (and in case of doubt/complexity, a third pathologist was also consulted). We tried to maintain similar criteria between the graders and always followed the same guidelines. These new labels were used to rectify the performance of all the algorithms evaluated in the test set. We argue that the information regarding the strength/confidence of predictions of a model used as a second opinion is of utter importance. A correct integration of this feature can be shown as extremely insightful for the pathologists using the developed tool.

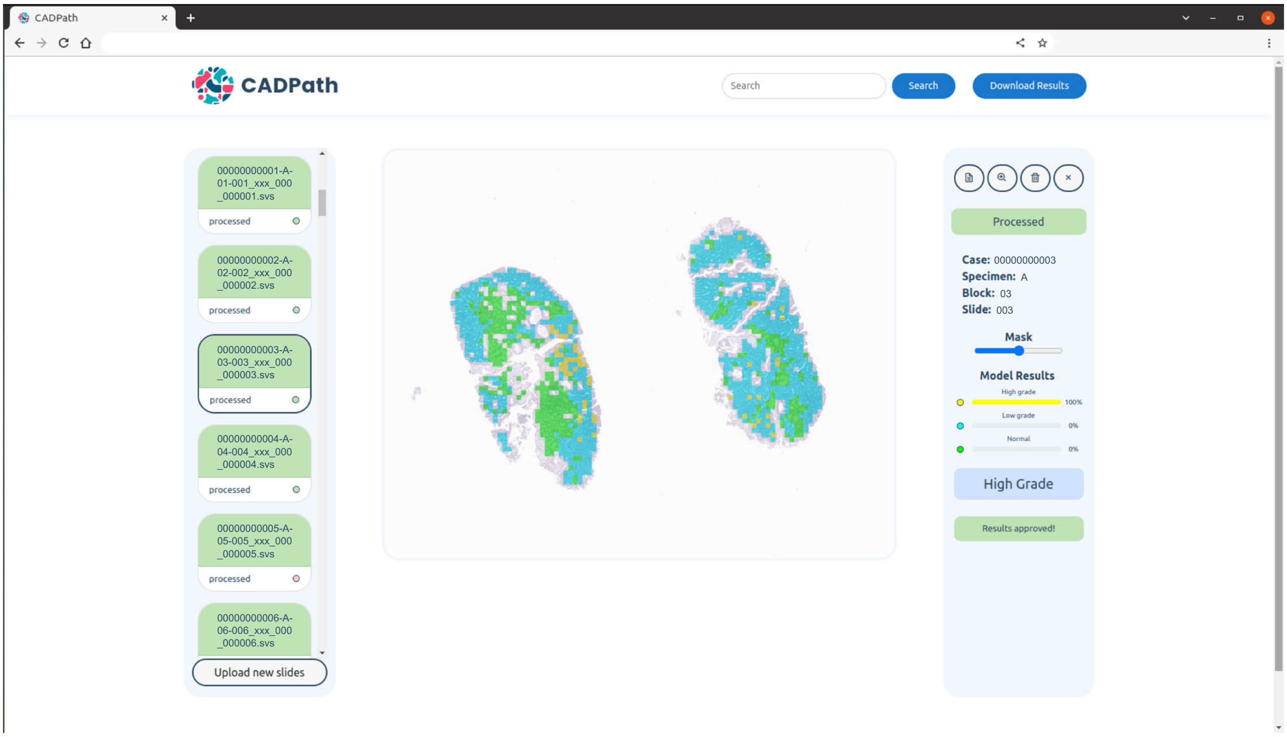
Experimental setup

For our experimental setup, we divide our data into training and validation sets. Besides, we further evaluate the performance of the former in our test set composed of slides never seen by any of the methods presented or in the literature. Following the split of these three sets, we have 8587, 1009 and 900 stratified non-overlapping samples in the training, validation and test set, respectively.

In an attempt to also contribute to reproducible research, the training of all the versions of the proposed algorithm uses the deterministic constraints available on PyTorch. The usage of deterministic constraints implies a trade-off between performance, either in terms of algorithmic efficiency or on its predictive power, and the complement with reproducible research guidelines. As such, due to the current progress in the field, we have chosen to comply with the reproducible research guidelines.

All the trained backbone networks were ResNet-34[46] with ImageNet weights. PyTorch was used to train these networks with the Adaptive Moment Estimation (Adam)[47] optimiser, a learning rate of 6×10^{-6} and a weight decay of 3×10^{-4}. The training batch size was set to 32 for both fully and weakly supervised training, while the test and inference batch size was 256. The performance of the model was verified on the validation set used for model selection in terms of the best accuracies and quadratic weighted kappa (QWK). The training was accelerated by an Nvidia Tesla V100 (32GB) GPU for 50 epochs of both weakly and fully supervised learning. In addition to the proposed methodology, we extended our experiments to include the aggregation approach proposed by Neto et al.[31] on top of our best-performing method. This strategy does not consider spatial location or context of the tiles, instead it select the seven most severe tiles, concatenates the output of the last convolutional layer for each of those tiles and feeds it to a multi-layer perceptron.

The number of epochs for training the fully- and weakly-supervised models was selected as follows: the fully supervised model was evaluated at every ten epochs for its performance on a weakly supervised scenario (using the non annotated samples), when its performance was stable the training was stopped; for the weakly-supervised model, several experiments were conducted on smaller versions of the training set and validated on the validation set. For the latter, besides training for 50 epochs, the best weights (with respect to the performance on the validation set) were selected.


CADPath
CADPath
Search
Search
Download Results
00000000001-A-01-001_xxx_000_000001.svs
processed
00000000002-A-02-002_xxx_000_000002.svs
processed
00000000003-A-03-003_xxx_000_000003.svs
processed
00000000004-A-04-004_xxx_000_000004.svs
processed
00000000005-A-05-005_xxx_000_000005.svs
processed
00000000006-A-06-006_xxx_000_000006.svs
Upload new slides
Processed
Mask
Model Results
High grade
100%
Low grade
0%
Normal
0%
High Grade
Results approved!

Case: 00000000003
Specimen: A
Block: 03
Slide: 003

Fig. 9 | **Main view of the CAD system prototype for CRS.** Slide identification, confidence per class, diagnostic, mask overlay controller, results download as csv and slide search are some of the features visible. Slide identification is anonymised.

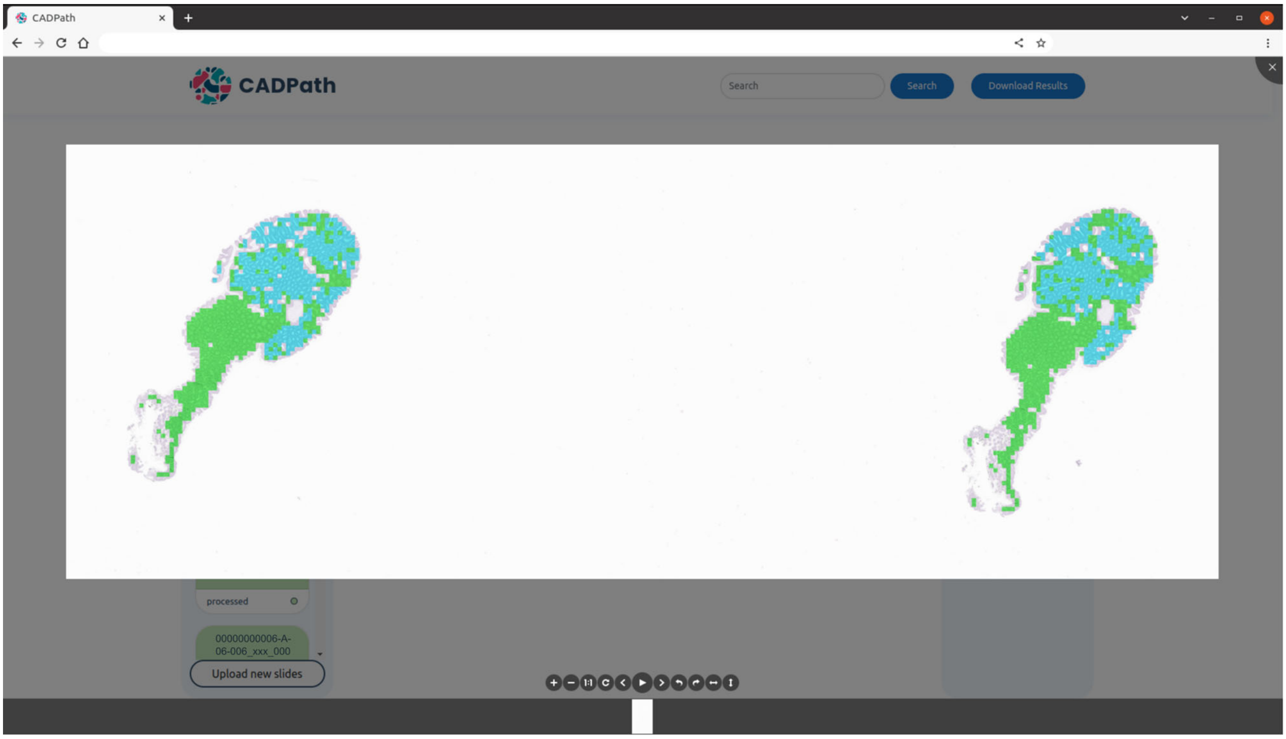
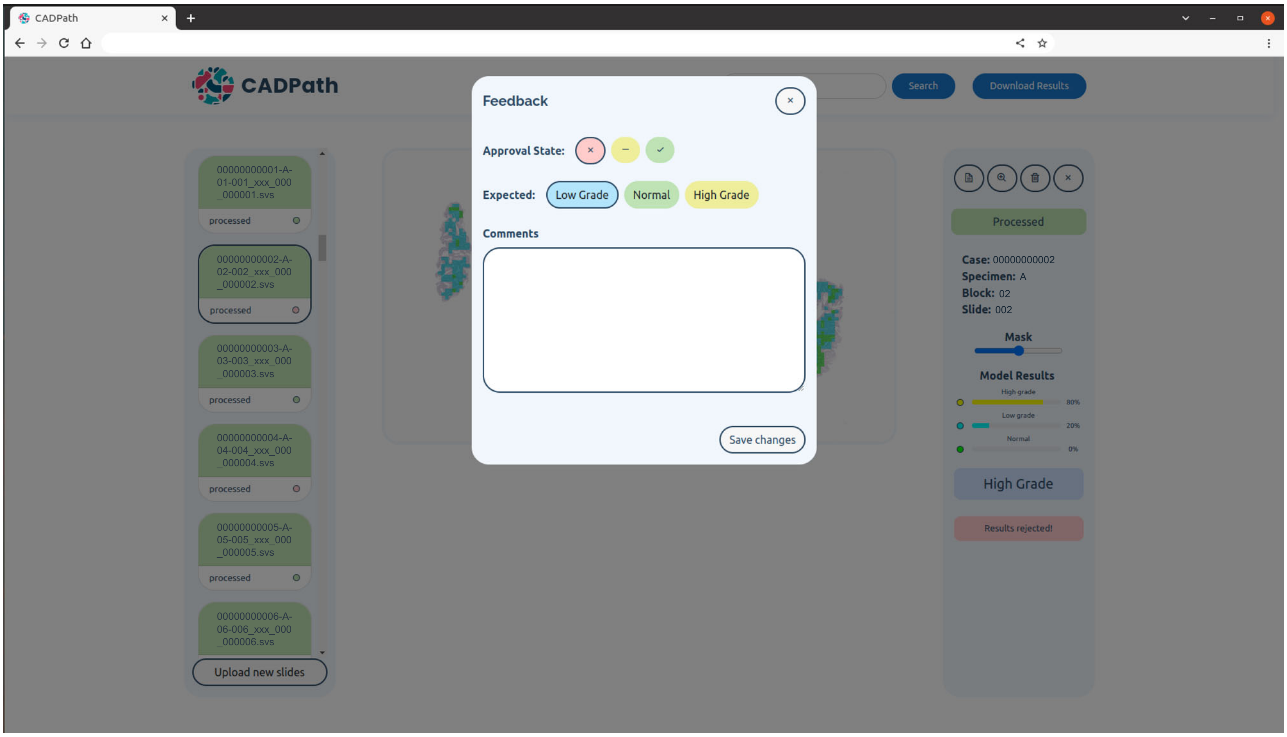


Fig. 10 | **Zoomed view of a slide from the CAD system prototype.** Further includes the predictions map with a small overlay threshold. Slide identification is anonymised.

Prototype and interpretability assessment

The proposed algorithm was integrated into a fully functional prototype to enable its use and validation in a real clinical workflow. This system was developed as a server-side web application that can be accessed by any pathologist in the lab. The system supports the evaluation of either a single slide or a batch of slides simultaneously and in real time. It also caches the most recent results, allowing re-evaluation without the need to re-upload slides. In addition to displaying the slide



00000000001-A-01-001_xxx_000_000001.svs
00000000002-A-02-002_xxx_000_000002.svs
00000000003-A-03-003_xxx_000_000003.svs
00000000004-A-04-004_xxx_000_000004.svs
00000000005-A-05-005_xxx_000_000005.svs
00000000006-A-06-006_xxx_000_000006.svs

Case: 00000000002
Specimen: A
Block: 02
Slide: 002

High grade 80%
Low grade 20%
Normal 0%

High Grade

Results rejected!

Fig. 11 | CAD system prototype report tool. The user can report if the result is either correct, wrong or inconclusive and leave a comment for each case individually. Slide identification is anonymised.

diagnosis, and confidence level for each class, a visual explanation map is also retrieved, to draw the pathologist's attention to key tissue areas within each slide (all seen in Fig. 9). The opaqueness of the map can be set to different thresholds, allowing the pathologist to control its overlay over the tissue. An example of the zoomed version of a slide with lower overlay of the map is shown in Fig. 10.

Furthermore, the prototype also allows user feedback where the user can accept/reject a result and provide a justification (Fig. 11), an important feature for software updates, research development and possible active learning frameworks that can be developed in the future. These results can be downloaded with the corrected labels to allow for further retraining of the model.

There are several advantages to developing such a system as a server-side web application. First, it does not require any specific installation or dedicated local storage in the user's device. Secondly, it can be accessed at the same time by several pathologists from different locations, allowing for a quick review of a case by multiple pathologists without data transference. Moreover, the lack of local storage of clinical data increases the privacy of patient data, which can only be accessed through a highly encrypted virtual private network (VPN). Finally, all the processing can be moved to an efficient GPU, thus reducing the processing time by several orders of magnitude. Similar behaviour on a local machine would require the installation of dedicated GPUs in the pathologists' personal devices. This platform is the first Pathology prototype for colorectal diagnosis developed in Portugal, and, as far as we know, one of the pioneers in the world. Its design was also carefully thought to be aligned with the needs of the pathologists.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

A large portion of the dataset has been publicly released[30] and is identified by the following https://doi.org/10.25747/fb1q-j507. This data composed of WSI and respective labels has been released under CC BY-NC. This release is part of the efforts of IMP Diagnostics and INESC TEC to advance science and share knowledge.

Received: 18 July 2023; Accepted: 8 February 2024;
Published online: 05 March 2024

References

1. International Agency for Research on Cancer (IARC). Global cancer observatory. https://gco.iarc.fr/ (2022).
2. Digestive Cancers Europe (DiCE). Colorectal screening in europe. https://bit.ly/3rFxSEL.
3. Hassan, C. et al. Post-polypectomy colonoscopy surveillance: European society of gastrointestinal endoscopy guideline - update 2020. *Endoscopy* **52**, 687–700 (2020).
4. Mahajan, D. et al. Reproducibility of the villous component and high-grade dysplasia in colorectal adenomas < 1 cm: Implications for endoscopic surveillance. *Am. J. Surg. Pathol.* **37**, 427–433 (2013).
5. Gupta, S. et al. Recommendations for follow-up after colonoscopy and polypectomy: a consensus update by the us multi-society task force on colorectal cancer. *Gastrointest. Endosc.* **115**, 415–434 (2020).
6. Eloy, C. et al. Digital pathology workflow implementation at ipatimup. *Diagnostics* **11**. https://www.mdpi.com/2075-4418/11/11/2111 (2021).
7. Fraggetta, F. et al. A survival guide for the rapid transition to a fully digital workflow: the "caltagirone example". *Diagnostics* **11**. https://www.mdpi.com/2075-4418/11/10/1916 (2021).
8. Montezuma, D. et al. Digital pathology implementation in private practice: specific challenges and opportunities. *Diagnostics* **12**, 529 (2022).
9. Madabhushi, A. & Lee, G. Image analysis and machine learning in digital pathology: challenges and opportunities. *Medical Image Analysis* **33**, 170–175 (2016).

10. Rakha, E. A. et al. Current and future applications of artificial intelligence in pathology: a clinical perspective. *J. Clin. Pathol.* **74**, 409–414 (2021).
11. Veta, M. et al. Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Med. Image Anal.* **20**, 237–248 (2015).
12. Campanella, G. et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, 1301–1309 (2019).
13. Oliveira, S. P. et al. Weakly-supervised classification of HER2 expression in breast cancer haematoxylin and eosin stained slides. *App. Sci.* **10**, 4728 (2020).
14. Albuquerque, T., Moreira, A. & Cardoso, J. S. Deep ordinal focus assessment for whole slide images. In *Proc. IEEE/CVF International Conference on Computer Vision*, 657–663 (2021).
15. Oliveira, S. P. et al. CAD systems for colorectal cancer from WSI are still not ready for clinical acceptance. *Sci. Rep.* **11**, 14358 (2021).
16. Thakur, N., Yoon, H. & Chong, Y. Current trends of artificial intelligence for colorectal cancer pathology image analysis: a systematic review. *Cancers* 12, 1884 (2020).
17. Wang, Y. et al. Application of artificial intelligence to the diagnosis and therapy of colorectal cancer. *Am. J. Cancer Res.* **10**, 3575–3598 (2020).
18. Davri, A. et al. Deep learning on histopathological images for colorectal cancer diagnosis: a systematic review. *Diagnostics* **12**, 837 (2022).
19. Iizuka, O. et al. Deep learning models for histopathological classification of gastric and colonic epithelial tumours. *Sci. Rep*.**10**, 1504 (2020).
20. Tizhoosh, H. & Pantanowitz, L. Artificial intelligence and digital pathology: challenges and opportunities. *J. Pathol. Inform.* **9**, 38 (2018).
21. Wei, J. W. et al. Evaluation of a deep neural network for automated classification of colorectal polyps on histopathologic slides. *JAMA Network Open* **3**, e203398 (2020).
22. Song, Z. et al. Automatic deep learning-based colorectal adenoma detection system and its similarities with pathologists. *BMJ Open* **10**, e036423 (2020).
23. Xu, L. et al. Colorectal cancer detection based on deep learning. *J. Pathol. Inf*. **11**, 28 (2020).
24. Wang, K.-S. et al. Accurate diagnosis of colorectal cancer based on histopathology images using artificial intelligence. *BMC Med.* **19**, 1–12 (2021).
25. Yu, G. et al. Accurate recognition of colorectal cancer with semi-supervised deep learning on pathological images. *Nat. Commun.* **12**, 1–13 (2021).
26. Marini, N. et al. Multi-scale task multiple instance learning for the classification of digital pathology images with global annotations. In *Proc. of the MICCAI Workshop on Computational Pathalogy of Proceedings of Machine Learning Research*, Vol. 156 (eds Atzori, M. et al.) 170–181 (PMLR, 2021).
27. Ho, C. et al. A promising deep learning-assistive algorithm for histopathological screening of colorectal cancer. *Sci. Rep.* **12**, 1–9 (2022).
28. Bokhorst, J.-M. et al. Deep learning for multi-class semantic segmentation enables colorectal cancer detection and classification in digital pathology images. *Sci. Rep.* **13**, 8398 (2023).
29. Graham, S. et al. Screening of normal endoscopic large bowel biopsies with interpretable graph learning: a retrospective study. *Gut* (2023). https://gut.bmj.com/content/early/2023/05/11/gutjnl-2023-329512.
30. Neto, P. C. et al. (2024). https://rdm.inesctec.pt/dataset/nis-2023-008.
31. Neto, P. C. et al. imil4path: a semi-supervised interpretable approach for colorectal whole-slide images. *Cancers* **14**, 2489 (2022).
32. WHO Classification of Tumours Editorial Board. *WHO classification of tumours of the digestive system* 5th edn (World Health Organization, 2019).
33. Pathcore. Sedeen viewer. https://pathcore.com/sedeen (2020).
34. Albuquerque, T. et al. Quality control in digital pathology: Automatic fragment detection and counting. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 588–593 (2022).
35. Kirk, S. et al. The Cancer Genome Atlas Colon Adenocarcinoma Collection (TCGA-COAD) (Version 3) [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2016.HJJHBOXZ (2016).
36. Kirk, S., Lee, Y., Sadow, C. A., & Levine, S. The Cancer Genome Atlas Rectum Adenocarcinoma Collection (TCGA-READ) (Version 3) [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2016.F7PPNPNU (2016).
37. Clark, K. et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. In *Journal of Digital Imaging*. Vol. 26, 1045–1057 (Springer Science and Business Media LLC 2013).
38. Platform, P. A. Paip (2020). http://www.wisepaip.org, accessed 20 January 2022.
39. Otsu, N. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics* **9**, 62–66 (1979).
40. Božič, J., Tabernik, D. & Skočaj, D. Mixed supervision for surface-defect detection: from weakly to fully supervised learning. *Comput. Industry* **129**, 103459 (2021).
41. Li, W., Mikailov, M. & Chen, W. Scaling the inference of digital pathology deep learning models using cpu-based high-performance computing. *IEEE Transactions on Artificial Intelligence,* **4***,*1691–1704 (2023).
42. Raschka, S. Model evaluation, model selection, and algorithm selection in machine learning. Preprint at https://arxiv.org/abs/1811.12808 (2018).
43. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**, 153–157 (1947).
44. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Machine Learn. Res.* **7**, 1–30 (2006).
45. Edwards, A. L. Note on the “correction for continuity” in testing the significance of the difference between correlated proportions. *Psychometrika* **13**, 185–187 (1948).
46. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
47. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. In *ICLR (Poster)* (2015).

Acknowledgements

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020 and the PhD Grants SFRH/BD/139108/2018 and 2021.06872.BD. Additionally, data collection, infrastructure and resources have been supported by the European Regional Development Fund (ERDF), through the Operational Programme for Competitiveness and Internationalisation (COMPETE 2020) Programme, within the project POCI-01-0247-FEDER-045413 and by IMP Diagnostics.

Author contributions

P.C.N., S.P.O. and D.M. contributed equally to this work; P.C.N., S.P.O. and J.S.C. designed the experiments; P.C.N. and S.P.O. conducted the experiments and the analysis of the results; J.M, L.R. and S.G. prepared the histopathological samples; J.F., D.O. and D.M. collected, reviewed and annotated the histopathological cases; P.C.N., S.P.O. and D.M. wrote the manuscript; A.M. and J.M. performed data curation and project management; I.M.P., I.Z. and S.R. clinically supervised the project; J.S.C., I.Z. and S.R. technically supervised the project. All authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare the following competing interests: D.M. and D.O. are employees at IMP Diagnostics and I.M.P. is an owner at IMP Diagnostics.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41698-024-00539-4.

Correspondence and requests for materials should be addressed to Pedro C. Neto, Diana Montezuma or Sara P. Oliveira.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.