Original research

Screening of normal endoscopic large bowel biopsies with interpretable graph learning: a retrospective study

Simon Graham [1,2] Fayyaz Minhas [1] Mohsin Bilal [1] Mahmoud Ali [3] Yee Wah Tsang,[3] Mark Eastwood [1] Noorul Wahab [1] Mostafa Jahanifar [1] Emily Hero [4] Katherine Dodd,[3] Harvir Sahota,[3] Shaobin Wu,[5] Wenqi Lu [1] Ayesha Azam [3] Ksenija Benes,[3,6] Mohammed Nimir [3] Katherine Hewitt [3] Abhir Bhalerao,[1] Andrew Robinson,[3] Hesham Eldaly,[3] Shan E Ahmed Raza [1] Kishore Gopalakrishnan [3] David Snead [2,3,7] Nasir Rajpoot [1,2,3]

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/gutjnl-2023-329512).

[1]Department of Computer Science, University of Warwick, Coventry, UK
[2]Histofy Ltd, Birmingham, UK
[3]Department of Pathology, University Hospitals Coventry and Warwickshire NHS Trust, Coventry, UK
[4]Department of Pathology, University Hospitals of Leicester NHS Trust, Leicester, UK
[5]Department of Pathology, East Suffolk and North Essex NHS Foundation Trust, Colchester, UK
[6]Department of Pathology, Royal Wolverhampton Hospitals NHS Trust, Wolverhampton, UK
[7]Division of Biomedical Sciences, University of Warwick Warwick Medical School, Coventry, UK

Correspondence to
Professor Nasir Rajpoot, Department of Computer Science, University of Warwick, Coventry CV4 7EZ, UK; n.m.rajpoot@warwick.ac.uk
Dr Simon Graham; simon.graham@warwick.ac.uk

DS and NR are joint senior authors.

Received 16 January 2023
Accepted 15 April 2023
Published Online First 12 May 2023

To cite: Graham S, Minhas F, Bilal M, *et al*. *Gut* 2023;**72**:1709–1721.

ABSTRACT

Objective To develop an interpretable artificial intelligence algorithm to rule out normal large bowel endoscopic biopsies, saving pathologist resources and helping with early diagnosis.

Design A graph neural network was developed incorporating pathologist domain knowledge to classify 6591 whole-slides images (WSIs) of endoscopic large bowel biopsies from 3291 patients (approximately 54% female, 46% male) as normal or abnormal (non-neoplastic and neoplastic) using clinically driven interpretable features. One UK National Health Service (NHS) site was used for model training and internal validation. External validation was conducted on data from two other NHS sites and one Portuguese site.

Results Model training and internal validation were performed on 5054 WSIs of 2080 patients resulting in an area under the curve-receiver operating characteristic (AUC-ROC) of 0.98 (SD=0.004) and AUC-precision-recall (PR) of 0.98 (SD=0.003). The performance of the model, named Interpretable Gland-Graphs using a Neural Aggregator (IGUANA), was consistent in testing over 1537 WSIs of 1211 patients from three independent external datasets with mean AUC-ROC=0.97 (SD=0.007) and AUC-PR=0.97 (SD=0.005). At a high sensitivity threshold of 99%, the proposed model can reduce the number of normal slides to be reviewed by a pathologist by approximately 55%. IGUANA also provides an explainable output highlighting potential abnormalities in a WSI in the form of a heatmap as well as numerical values associating the model prediction with various histological features.

Conclusion The model achieved consistently high accuracy showing its potential in optimising increasingly scarce pathologist resources. Explainable predictions can guide pathologists in their diagnostic decision-making and help boost their confidence in the algorithm, paving the way for its future clinical adoption.

WHAT IS ALREADY KNOWN ON THIS TOPIC

- ⇒ Increasing screening rates for early detection of colon cancer are placing significant pressure on already understaffed and overloaded histopathology resources worldwide and especially in the UK.
- ⇒ Approximately one-third of endoscopic colon biopsies are reported as normal, and therefore, require minimal intervention, yet the biopsy results can take up to 2–3 weeks.
- ⇒ Artificial intelligence (AI) models hold great promise for reducing the burden of diagnostics for cancer screening but require incorporation of pathologist domain knowledge and explainability.

WHAT THIS STUDY ADDS

- ⇒ This study presents the first AI algorithm for rule out of normal from abnormal large bowel endoscopic biopsies with high accuracy across different patient populations.
- ⇒ For colon biopsies predicted as abnormal, the model can highlight diagnostically important biopsy regions and provide a list of clinically meaningful features of those regions such as glandular architecture, inflammatory cell density and spatial relationships between inflammatory cells, glandular structures and the epithelium.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

- ⇒ The proposed tool can both screen out normal biopsies and act as a decision support tool for abnormal biopsies, therefore, offering a significant reduction in the pathologist workload and faster turnaround times.

INTRODUCTION

Histological examination is a vital component in ensuring accurate diagnosis and appropriate treatment of many diseases. In routine practice, it involves visual assessment of key histological and cellular patterns in the tissue, which is a major step in understanding the state of various conditions, such as cancer. Histopathology has been at the forefront of many advances in care including, but not limited to, cancer screening programmes, molecular pathology, tumour classification and companion diagnostic testing, resulting in a rapid rise in demand for histology-derived data.[1] This extra workload is placing tremendous pressure on pathologists,[2] with 78% of UK cellular pathology departments already facing significant staff shortages.[3] The surging demand and staffing challenges ultimately lead to delays in diagnosis,[4] negatively impacting patient care especially for those with abnormal conditions (eg, cancer or serious inflammation) where early intervention and treatment are critical.[5]

New National Institute for Health and Care Excellence guidelines for referral of suspected cancer forecast an unprecedented rise in demand for endoscopy, with more than 750 000 additional procedures performed per year by 2020,[6] leading to a breach in standard wait times in a quarter of National Health Service (NHS) hospitals.[7 8] Endoscopic large bowel biopsies constitute approximately 10% of all requests in the UK NHS pathology laboratories. During the examination process, the pathologist examines each biopsy slide searching for disease, typically working from low to high magnification, and analyses a set of predefined histological features, such as gland architecture, inflammation and nuclear atypia for signs of abnormality.[9 10] The resulting report indicates the presence of any disease process and categorises the abnormality into the most appropriate diagnosis.[11 12] An overview of the pathologist diagnostic decision process for reporting endoscopic colon biopsies is provided in online supplemental figure 1. Approximately one-third of colonic biopsy samples are reported as normal (online supplemental table 1), representing a substantial workload where the pathologist's expertise is not fully used. The underlying hypothesis of this study is that automated screening of normal biopsies may help address rising histopathology capacity challenges.

Since the advent of digital pathology,[13] there has been a sharp increase in the development of artificial intelligence (AI) tools that enable computational analysis of multi-gigapixel whole-slide images (WSIs). In particular, deep learning (DL) algorithms have achieved remarkable performance not only in routine diagnostic tasks, such as cancer grading[14] and finding metastasis in lymph nodes, but also in finding origins for cancers of unknown primary[15] and improved patient stratification.[16 17] Notably, Campanella *et al*[18] presented a seminal paper on clinical-grade WSI classification, while Ehteshami Bejnordi *et al*[19] demonstrated that AI models are capable of surpassing pathologist performance for breast cancer metastasis detection. These models can be leveraged to help reduce inevitable errors in diagnosis, given that humans are naturally prone to mistakes, especially when faced with fatigue or distractions.[20 21] Despite challenges associated with algorithm bias,[22 23] AI tools are not as susceptible to these kinds of errors and therefore may help mitigate oversight, reduce workload and increase reproducibility.

Differentiating between normal and neoplastic colorectal WSIs using DL has previously been addressed, with reports of excellent performance.[24–26] However, distinguishing normal from abnormal tissue samples required for large bowel biopsy screening remains a challenge, due to the difficulty in detecting various subtle conditions, such as mild inflammation. To the best of our knowledge, there are no existing multi-centric studies for normal versus abnormal classification of large bowel biopsies. Existing methods for colonic analysis operate on high power subimages (or image patches) and so do not explicitly model both the tissue microstructure and macrostructure, including glandular architecture, inflammatory cell density and spatial relationships between inflammatory cells, glandular structures and the epithelium. Relying solely on DL models to automatically detect histological patterns that are diagnostically relevant in small image regions may lead to suboptimal performance. Alternatively, explicitly incorporating histological features that are routinely used by pathologists during the colon biopsy diagnostic workflow may not only improve performance over conventional DL models but may also increase transparency and interpretability of the algorithm's decision-making to the pathologist—a key requirement for trustworthy AI-based medical decision models.[27 28]

To help reduce the burden of large bowel biopsy screening, we propose the first interpretable AI algorithm for large bowel slide classification employing a gland-graph network named IGUANA (Interpretable Gland-Graphs using a Neural Aggregator). In the proposed approach, a WSI is modelled as a graph with nodes,[29–33] each representing a gland associated with a set of 25 interpretable features capturing gland architecture, intra-gland nuclear morphology and inter-gland cell density. The interconnections between these nodes capture the spatial organisation of glands within the tissue. The node features were developed in collaboration with pathologists and in accordance with existing diagnostic pathways to boost predictive accuracy, interpretability and alignment with known histological characteristics of a wide range of colorectal pathologies. IGUANA identifies highly predictive regions in the biopsy tissue slide and provides an explanation as to why they may be highly predictive. Because of the use of biologically meaningful features, this explanation can easily be interpreted by a pathologist as the basis of the algorithm's diagnostic decision-making. We validate our algorithm on an internal dataset containing 5054 WSIs and an independent multi-centre dataset containing 1561 WSIs, achieving the best performance compared with recent top-performing approaches. In addition, we analyse predictive regions identified by IGUANA along with local and WSI-level explanations and show that our approach can identify areas of abnormality, such as inflammation and neoplasia. The code for IGUANA is available in the open-source domain for research purposes (https://github.com/TissueImageAnalytics/iguana) and example results can be visualised in an interactive demo available at https://iguana.dcs.warwick.ac.uk.

MATERIALS AND METHODS

Study design

A summary of the used datasets and our overall pipeline can be seen in figure 1, which consists of the following steps: (1) histological segmentation, (2) feature extraction and edge generation, (3) graph prediction and (4) graph explanation. An overview of the experiment design is provided in online supplemental figure 2 and an in-depth description of the used datasets is given in online supplemental section S4.1, including the disease and demographic breakdown (online supplemental figures 3 and 4 and online supplemental tables 2–4). In addition, we provide a detailed method description in online supplemental sections S4.1–S4.7.

Patient and public involvement

Lay members have made a valuable contribution to this project in ensuring that the patient is at the heart of this project. Three lay advisors have been working with us since the conception of this project. One of the advisors is part of the National Cancer Research Institute consumer network and Independent Cancer Patient's Voice group, who are both supportive of new technologies being brought into the NHS for patient benefit.

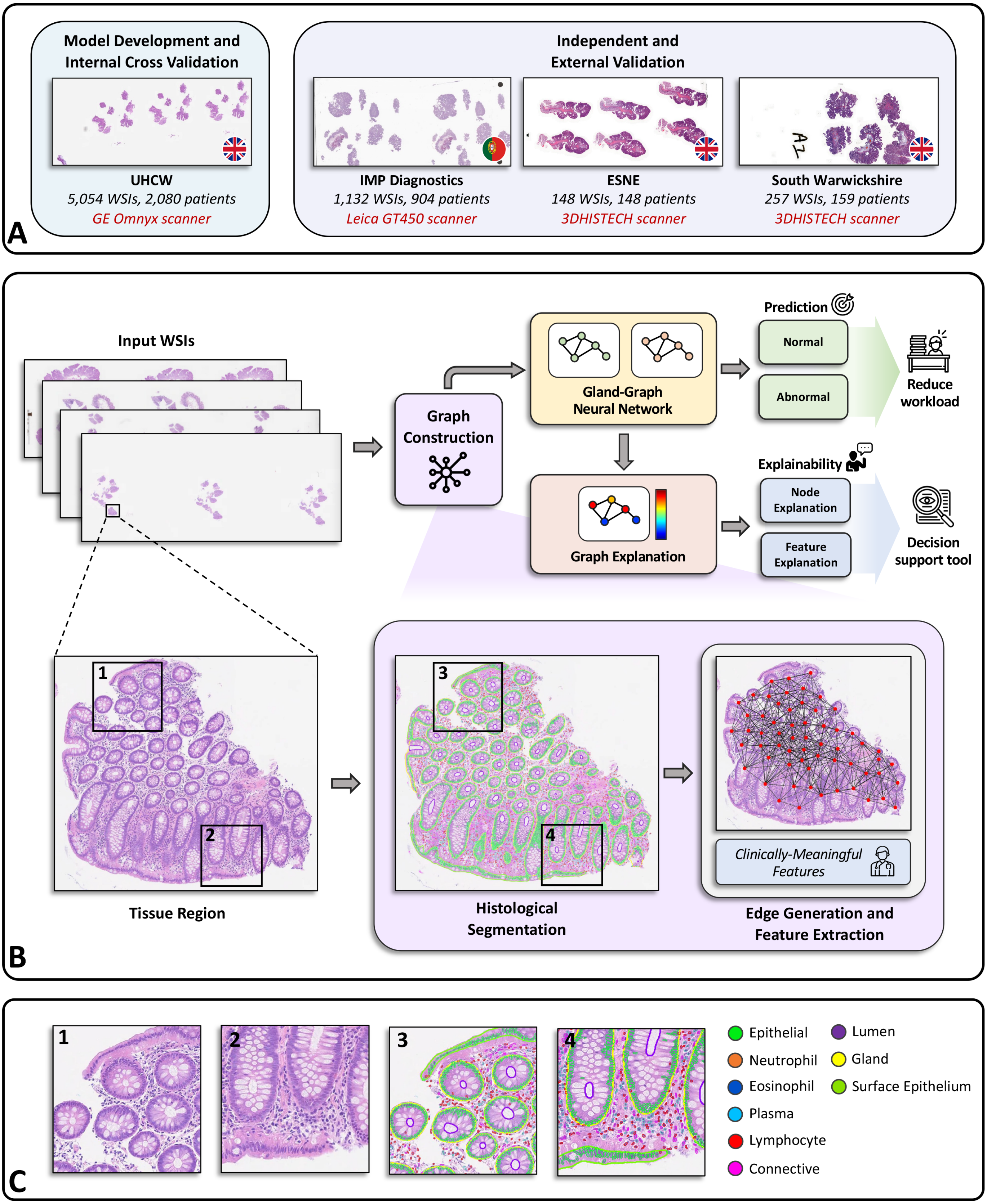


Figure 1 Illustration of the overall pipeline for colon tissue classification with gland-graph convolutional networks. (A) Overview of the data used in our experiments from four different centres using different scanners. (B) Summary of the pipeline, which involves graph construction, gland-graph inference and gland-graph explanation. (C) Zoomed-in image regions and corresponding results taken from the example in B. ESNE, East Suffolk and North Essex; UHCW, University Hospitals Coventry and Warwickshire; WSI, whole-slide images.

RESULTS

Large-scale cross-validation for colon biopsy screening

To rigorously evaluate our approach for colon biopsy screening, we performed 3-fold cross-validation using 5054 H&E-stained colon biopsy WSIs from University Hospitals Coventry and Warwickshire (UHCW), where each slide was labelled as either normal or abnormal. Interpretable screening of normal colon biopsies is a challenging problem due to a wide spectrum of large bowel abnormalities including a variety of neoplastic and inflammatory conditions. Figure 2 shows the results of IGUANA, achieving an average area under the receiver operating characteristic (AUC-ROC) curve of 0.9783 ± 0.0036 and an AUC precision-recall (AUC-PR) of 0.9798 ± 0.0031. We also include results obtained using other existing slide-level classification algorithms such as Iterative Draw and Rank Sampling (IDaRS)[34], Clustering-constrained Attention Multiple Instance Learning (CLAM)[35] and a random forest (RF) baseline classifier using our glandular features (denoted by Gland-RF). We observe that IGUANA achieves the best performance compared with both patch-based methods (IDaRS and CLAM), demonstrating its strong predictive ability given that it uses only 25 features per gland. We provide additional comparative results between IGUANA and IDaRS in online supplemental figure 5. Detailed statistical results are also provided in online supplemental tables 5–9. Note that despite IGUANA outperforming it, the Gland-RF model produces comparable performance—signifying the strength of our set of clinically derived features—although without the localised interpretability provided by IGUANA. Also, as opposed to the two patch-based methods, IGUANA provides concrete justification as to why a certain diagnostic class was predicted. We go into further detail on interpretability and explainability later in this section.

In addition, we assess differences in model performance across sex, age, ethnicity and anatomical site of the biopsy. For each subgroup-level analysis, we run 100 bootstrap runs to compute average AUC-ROC and its SD across subcategories (online supplemental table 10) and observe that our method is not biased towards any particular subgroup with only minor differences.

Model generalisation to independent cohorts

A true reflection of a model's clinical utility requires the assessment of its performance on completely unseen cohorts. For this, we used three additional cohorts of H&E-stained colon biopsy slides, providing a total of 1537 WSIs. These cohorts consisted of 1132 slides from IMP Diagnostics Laboratory in Portugal,[25] 148 slides from East Suffolk and North Essex (ESNE) NHS Foundation Trust and 257 slides and South Warwickshire NHS Foundation Trust, where slides were again categorised as either normal or abnormal. We observe from figure 2 that our model attains high performance for both the ESNE and South Warwickshire cohorts, reaching AUC-ROC scores of 0.9567 ± 0.0155, 0.9649 ± 0.0025 and 0.9789 ± 0.0023 and AUC-PR scores of 0.9731 ± 0.0105, 0.9466 ± 0.0034 and 0.9949 ± 0.0006 for ESNE, South Warwickshire and IMP datasets, respectively. It is evident that there is a large difference in performance between IGUANA and other approaches on the external cohorts, signifying that superior generalisation to unseen data is a strength of our model. At a sensitivity of 0.99, we obtain a percentage increase over IDaRS of 47.4%, 63.6% and 58.9% for IMP, ESNE and South Warwickshire cohorts, respectively. This may be partly due to the ability of our initial segmentation model to perform well across images with different staining protocols.[36] Example results obtained by this model across the four datasets are shown in figure 3.

Analysis of expected reduction in pathologist workload

The real-world value of our approach is determined by its ability to reduce pathologist workload. As our model is intended for screening, it must achieve high sensitivity. Therefore, assessment of the specificity at high sensitivity cut-off thresholds provides a good indication of its potential effectiveness as a screening tool. Here, the specificity is indicative of the percentage reduction in normal slides that require pathologist review. In the middle column of figure 2, we display the specificity of our model at sensitivities of 0.97, 0.98 and 0.99 on all datasets used in our experiments, where we see that IGUANA sustains the best performance at various cut-offs compared with other methods. During internal cross-validation, we obtain specificities of 0.7865 ± 0.0429, 0.6720 ± 0.1128 and 0.5409 ± 0.1210 for sensitivities of 0.97, 0.98 and 0.99, respectively. For independent validation, our method obtains average specificities across the three external datasets of 0.7513 ± 0.0919, 0.6679 ± 0.0779 and 0.5487 ± 0.1599 for sensitivities of 0.97, 0.98 and 0.99. Therefore, this indicates that at a sensitivity of 0.99, our method is able to screen around 54% of normal cases during both internal and external validation.

In online supplemental figure 6, we show the proportion of slides that require pathologist review to achieve a certain sensitivity.[18] In these plots, we consider a target sensitivity of 0.99, which is reasonable due to high levels of interobserver disagreement for conditions such as mild inflammation. We also show with a vertical dashed line the proportion of abnormal slides in each dataset, which indicates the minimum number of slides that need to be reviewed for screening. For each of the cohorts, we observe that for our target of 0.99 sensitivity our model can screen out 32%, 31%, 17% and 13% of slides from UHCW, South Warwickshire, ESNE and IMP datasets, respectively. If considering a sensitivity of 0.97, we can screen out 44% of slides from UHCW, 46% from South Warwickshire, 30% from ESNE and 19% from IMP.

Local feature explanations increase model transparency

A major component of IGUANA is the ability to provide an interpretable and explainable output. In figure 4, we display visual explanations of the most predictive nodes and features given by IGUANA. Node explanations are shown in the form of a heatmap, where relatively high values indicate glandular areas that contribute to the slide being classified as abnormal. Therefore, we should expect that all glands in a normal slide will have low values in the associated heatmap as shown in figure 4A, where no glands contribute to the slide being classified as abnormal. Figure 4B–D shows WSIs with hyperplastic polyps, inflammation and adenocarcinoma, respectively. Hyperplastic polyps are often characterised by intraluminal folds and lumen dilation. On the other hand, inflammatory conditions usually have an increased number of lymphocytes, plasma cells, eosinophils and neutrophils within the lamina propria and potentially within the glands. Other indicators of inflammation can include crypt branching and crypt dropout. Colon adenocarcinoma is often denoted by irregular glandular morphology, epithelial nuclear atypia and multiple lumina. High-grade cancers typically lose their glandular appearance and form solid sheets of tumour cells. It can be observed that IGUANA is able to pick up abnormal glands with features in line with the above descriptions. In particular, we see that the most predictive glands in figure 4B contain

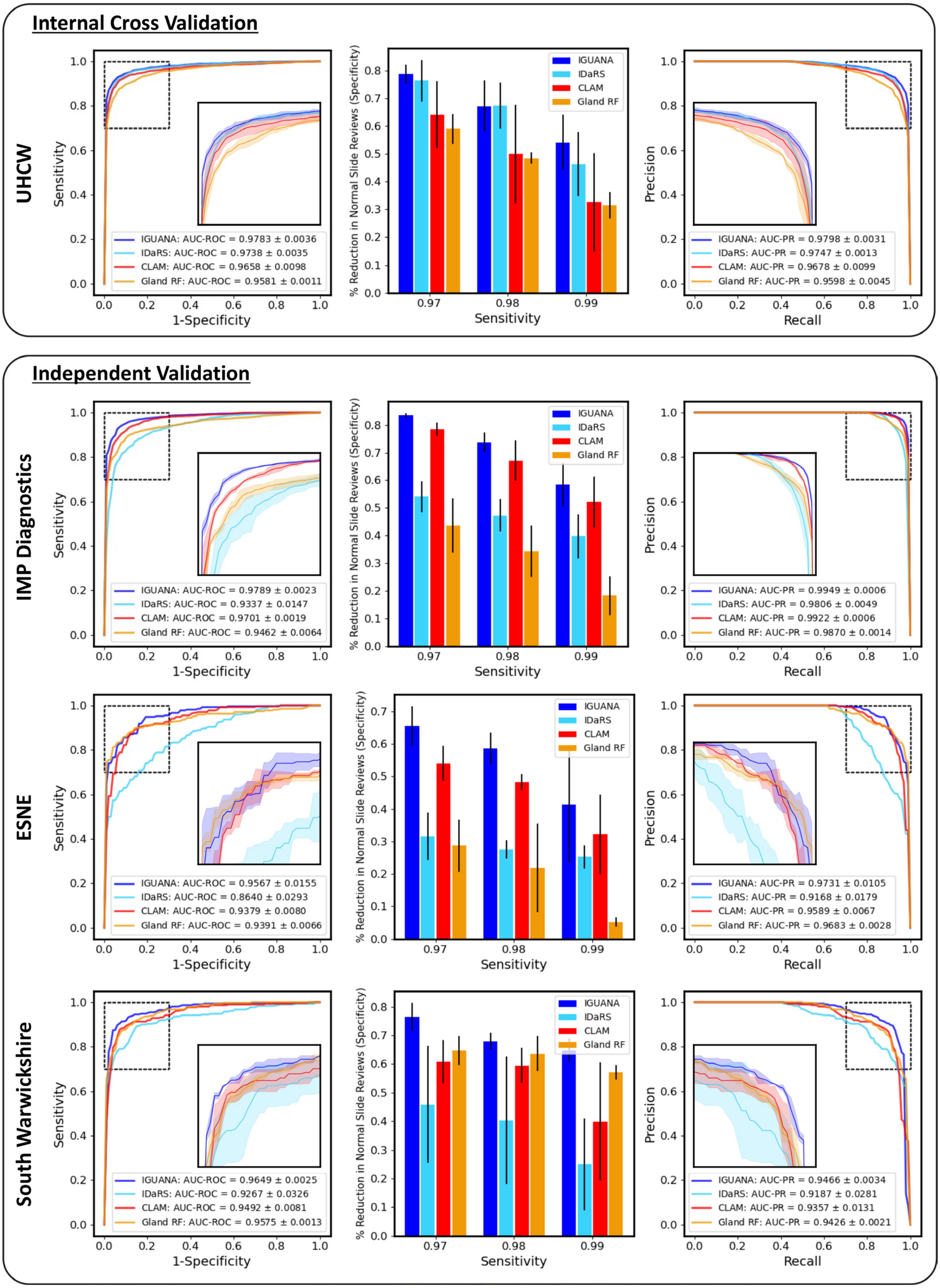


Figure 2 Results obtained across the four cohorts used in our experiments. Here, we display the ROC and PR curves along with the respective AUC scores of our approach compared with IDaRS, CLAM and Gland-RF (a random forest approach using the same handcrafted features with global aggregation). We also display the specificities obtained at sensitivity cut-offs of 0.97, 0.98 and 0.99. The shaded areas in the curves and the error bars in the bar plots show one SD from the results. AUC, area under the curve; CLAM, Clustering-constrained Attention Multiple Instance Learning; ESNE, East Suffolk and North Essex; IDaRS, Iterative Draw and Rank Sampling; PR, precision-recall; RF, random forest; ROC, receiver operating characteristic; UHCW, University Hospitals Coventry and Warwickshire.

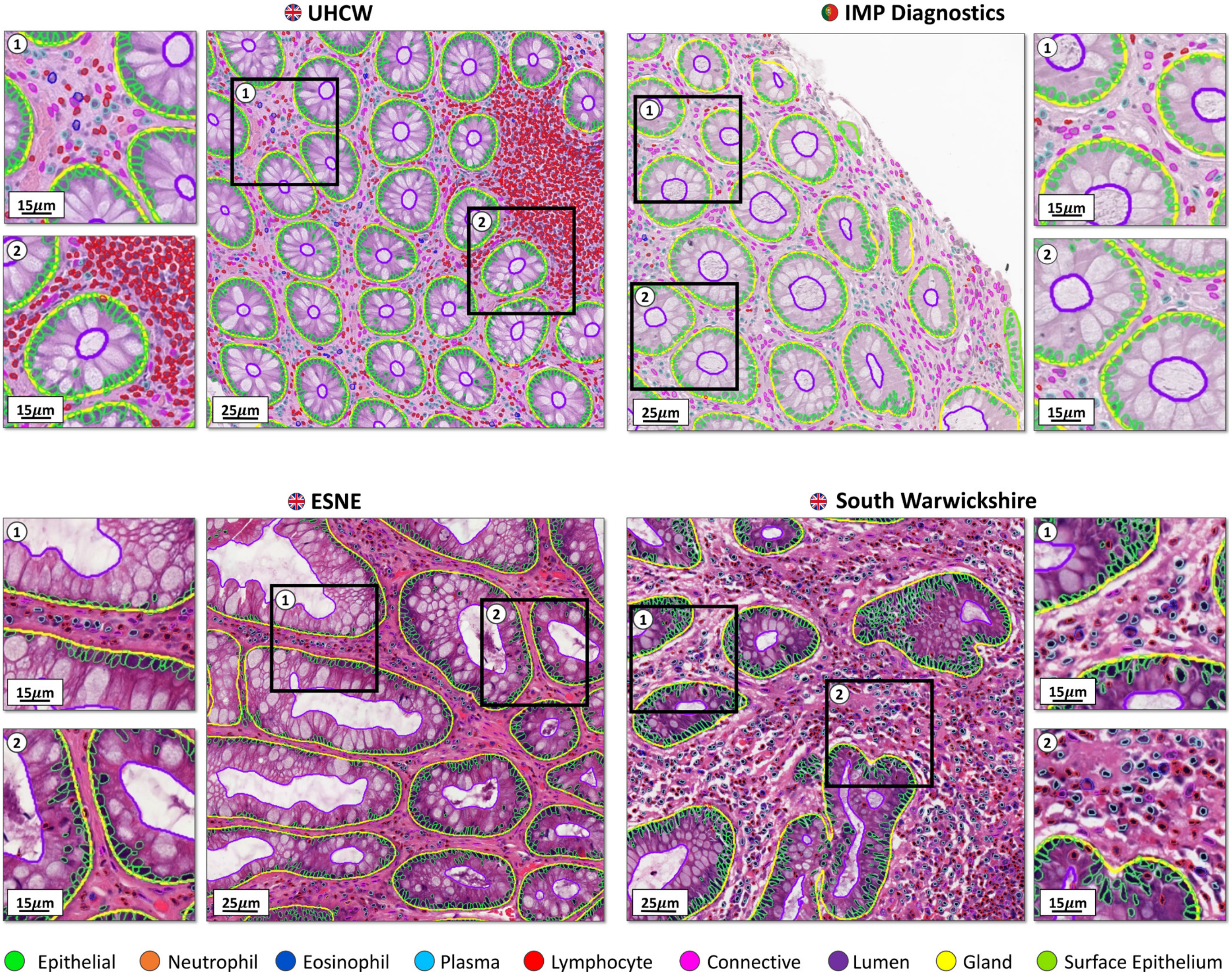


Figure 3 Example segmentation results obtained by our multi-task model across the four datasets used in our experiments. The top row shows normal examples, whereas the bottom row shows abnormal examples. In particular, the bottom-left example from ESNE shows a hyperplastic polyp and the bottom-right example from South Warwickshire shows inflammation. AUC-PR, area under the curve-precision-recall; CLAM, Clustering-constrained Attention Multiple Instance Learning; ESNE, East Suffolk and North Essex; IDaRS, Iterative Draw and Rank Sampling; IGUANA, Interpretable Gland-Graphs using a Neural Aggregator; RF, random forest; UHCW, University Hospitals Coventry and Warwickshire.

lumen with a clearly irregular morphology, whereas highlighted glands in figure 4C show areas with a high degree of inflammation. The adenocarcinoma heatmap in figure 4D highlights areas that have lost their conventional glandular appearance. Specifically, epithelial nuclei are no longer arranged at the gland boundary, cribriform architecture is observed and glands appear much larger, due to the formation of tumour cell sheets.

In addition to the node explanation heatmap, IGUANA indicates why certain glands are being identified as abnormal. This is useful because it can provide confirmation that the correct features are being identified by the model, giving researchers and clinicians confidence that it is performing as expected. This strategy can also be used to identify additional features within abnormal conditions. To show this, in figure 4, we display the most predictive glands in each slide and provide the corresponding feature explanations. Specifically, we display the top ten features in descending order of significance, along with their corresponding feature importance values between 0 and 1. Here, we expect that the feature explanations should align with what is observed in the associated cropped regions. In our hyperplastic polyp example, we see that the top glands (ie, 1, 2 and 3) contain lumen with abnormal morphology, whereas lumen dilation is observed in top gland 4. In line with this, lumen morphology and lumen composition are high-scoring features across the provided examples. We also observe that lumen size and organisation of epithelial nuclei within the glands are often found to be important features. In the example shown in figure 4C, we observe that top glands have a high degree of inflammation, which is matched by top features, such as inflammatory cell density, gland density and lamina propria neutrophil proportion. In the adenocarcinoma example, we see that the top four glands are all large, have irregular morphology and often display solid sheets of tumour cells with no obvious glandular structure. This is highlighted in the feature explanation, where gland morphology, gland size and epithelial organisation are consistently top-ranked features. Here, epithelial organisation describes how the epithelial nuclei are positioned at the gland boundary. Due to the presence of solid tumour patterns across the top glands,

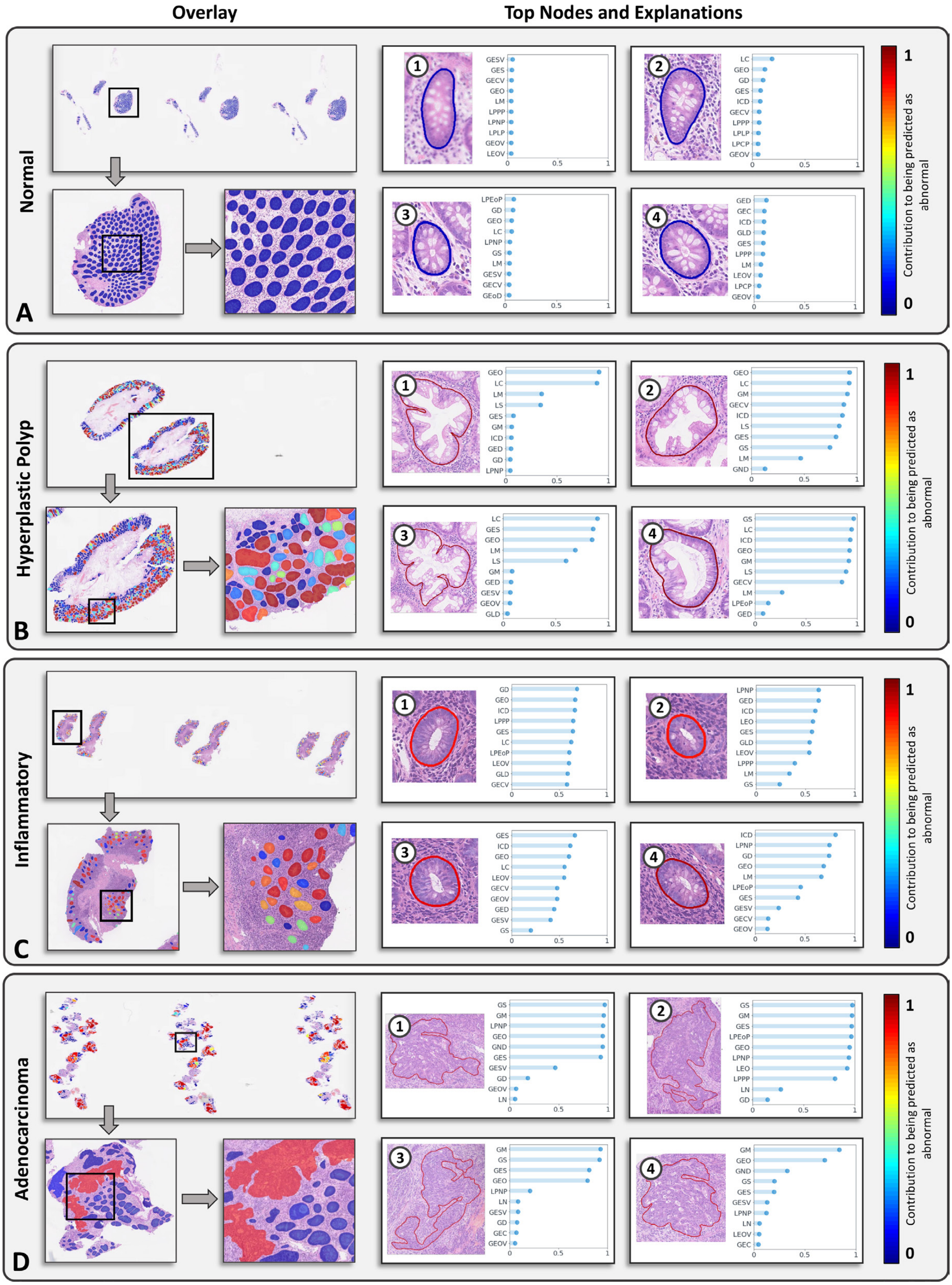


Figure 4 Visualisation of node and feature explainability. We display the overlay of the node-level explanations in the form of a heat map showing the most predictive nodes in the WSI. We also show cropped images of the four most predictive nodes within each WSI along with the associated ten most predictive features and their feature importance value. The colour of the boundary of the top nodes (glands) indicates the corresponding value in the node explanation heatmap. (A–D) Show example slides that are normal, hyperplastic, inflammatory or cancerous, respectively. GEC, gland epithelial clustering; GECV, gland epithelial clustering variation; GED, gland epithelial density; GEO, gland epithelial organisation; GEoD, gland eosinophil density; GEOV, GEO variation; GES, gland epithelial size; GESV, GES variation; GD, gland density; GLD, gland lymphocyte density; GM, gland morphology; GND, gland neutrophil density; GS, gland size; ICD, Inflammatory cell density; LEO, Lumen epithelial organisation; LEOV, Lumen epithelial organisation variation; LPCP, lamina propria connective proportion; LPEoP, lamina propria eosinophil proportion; LPLP, lamina propria lymphocyte proportion; LPNP, lamina propria neutrophil proportion; LPPP, lamina propria plasma proportion; WSIs, whole-slide images.

this feature is frequently highlighted in cancerous cases. We provide additional visual examples of the interpretability of our model output in figure 5.

WSI-level feature explanations are consistent with known histological patterns

In figure 6A, we show WSI-level explanations averaged over different subconditions in the UHCW and IMP cohorts. We focus on these datasets because they are the largest, with both containing over 1000 samples. Here, we plot top 10 features across the various subconditions for increased readability. These plots can be used both to confirm that the global explanations are as expected and to understand which features are particularly important for categorising a certain subcondition as abnormal. In both UHCW and IMP cohorts, the normal radar plots have a small radius, indicating that no feature contributes to the slide being classified as abnormal. For inflammatory cases, the UHCW and IMP radar plots show that a wide range of features can contribute to the slide being classified as abnormal, where there may be both cellular and architectural changes in the tissue. However, the most important features that can differentiate between other subconditions include inflammatory cell density, gland lymphocyte infiltration and gland density. Gland density can be indicative of gland dropout, which is a sign of inflammation. The UHCW radar plots for dysplasia and adenocarcinoma are similar, where the most important features are gland morphology, gland epithelial cell organisation, gland epithelial cell size and variation of gland epithelial cell size. This is in line with the key expected histological patterns observed within these tissue types. Likewise, these plots are similar to the low-grade and high-grade dysplasia plots for the IMP cohorts, indicating that the correct histological features are being highlighted when providing the WSI feature explanation. For hyperplastic polyps, we can see that lumen composition, lumen morphology and epithelial cell organisation have a large influence in the slide being classified as abnormal. Lumen composition is the ratio of lumen to gland size, and therefore, can identify glands with lumen dilation, which is a distinguishing feature of hyperplastic polyps. Conversely, lumen serrations, which are present in hyperplastic polyps, can lead to irregular lumen morphology, further validating the feature explanations output by our model.

WSI-level feature explanations identify population subgroups

In figure 6B, we perform hierarchical biclustering of all abnormal slides and WSI-level feature importance scores to help identify various subgroups that exist within the UHCW dataset. At the bottom of the plot, we identify various patient clusters which have varying histological appearance. These are numbered as follows: (1) general sign of inflammation, without neutrophil infiltration; (2) inflammation with a high degree of both lymphocytic and neutrophilic gland infiltration; (3) mainly neoplastic slides with irregular-shaped glands and large epithelial cells; (4) irregular gland morphology, with minimal inflammation; (5) abnormal lumen morphology and composition, with signs of inflammation in the lamina propria; (6) increased eosinophilic infiltration in the lamina propria and (7) neoplastic slides with gland epithelial clustering. Therefore, this gives us confidence that the network is learning key histological differences among the dataset to make an informed WSI-level prediction. More fine-grained clusters can be observed by referring to the associated dendrograms in the biclustering plot.

Interactive visualisation of results

We provide an interactive demo at https://iguana.dcs.warwick.ac.uk showing sample IGUANA results and highlighting the full output of our model at global and local levels, including the intermediate gland, lumen and nuclear segmentation results. In particular, we display the node explanations overlaid as a heat map on top of the glands and the local explanations by hovering over each node in the overlaid graph. Here, we provide the top five features to provide insight into what is contributing to certain glands being flagged as abnormal. It may also be of interest to assess the difference in features for nodes across the WSI. Therefore, we also enable visualisation of each of the 25 features overlaid on top of the glands as heat maps.

DISCUSSION

There has been a staffing crisis in pathology for many years,[37] which is being further exacerbated by the increased demand for histopathological examination. Embracing new technologies and AI in clinical practice may be necessary as hospitals seek to find new ways to improve patient care.[38] AI screening of large bowel endoscopic biopsies holds great promise in helping to reduce these escalating workloads by filtering out normal specimens. However, currently there does not exist a solution that can do this with a high predictive performance. Also, explainable AI is now recognised as a key requirement for trustworthy AI in human-centred decision-making,[28] yet is usually not considered in many healthcare applications. Therefore, in this study, we developed an AI model that can accurately differentiate normal from abnormal large bowel endoscopic biopsies, while providing an explanation for why a particular diagnosis was made.

We demonstrated that our proposed method for automatic colon biopsy screening could achieve a strong performance during both internal cross validation (mean AUC-ROC=0.98, mean AUC-PR=0.98) and on three independent external datasets (mean AUC-ROC=0.97, mean AUC-PR=0.97). Highly sensitive tools for screening are required to minimise the number of undetected abnormal conditions, since the false negative report is likely to lead to delayed diagnosis and potential patient harm. We believe a sensitivity of 0.99 is a reasonable target because the ground truth being used is the diagnosis provided by pathologists, which may have less than perfect sensitivity. This is also reflected in guidelines for breast biopsy screening in the UK, where sensitivities of 0.99 are expected.[39] Currently, we obtain promising specificities of 0.789 ± 0.043 at a sensitivity of 0.97 and 0.541 ± 0.121 at a sensitivity of 0.99, which could have a positive impact on reducing pathologist workloads. We also show in online supplemental figure 6 the expected reduction in clinical workload, where we report up to a 32% time saving by screening out normal biopsies that do not require assessment, while still maintaining a sensitivity of 0.99.

To understand misclassifications made by our model, we show six normal slides with the highest predicted abnormality scores in online supplemental figure 7. After inspection, we see that IGUANA correctly classifies these slides and therefore identifies mislabelling errors in the dataset. Here, the examples should have been labelled as either inflammatory or hyperplastic polyp. In the figure, we include sample image regions, as well as local and WSI-level feature explanations that are reflective of the true category of each slide. In addition, we performed a false negative analysis, where in online supplemental figure 8a we show the counts of various subconditions along with the corresponding number of false negatives. In online supplemental figure 8b, we show the false negative rate of each category. It can be observed

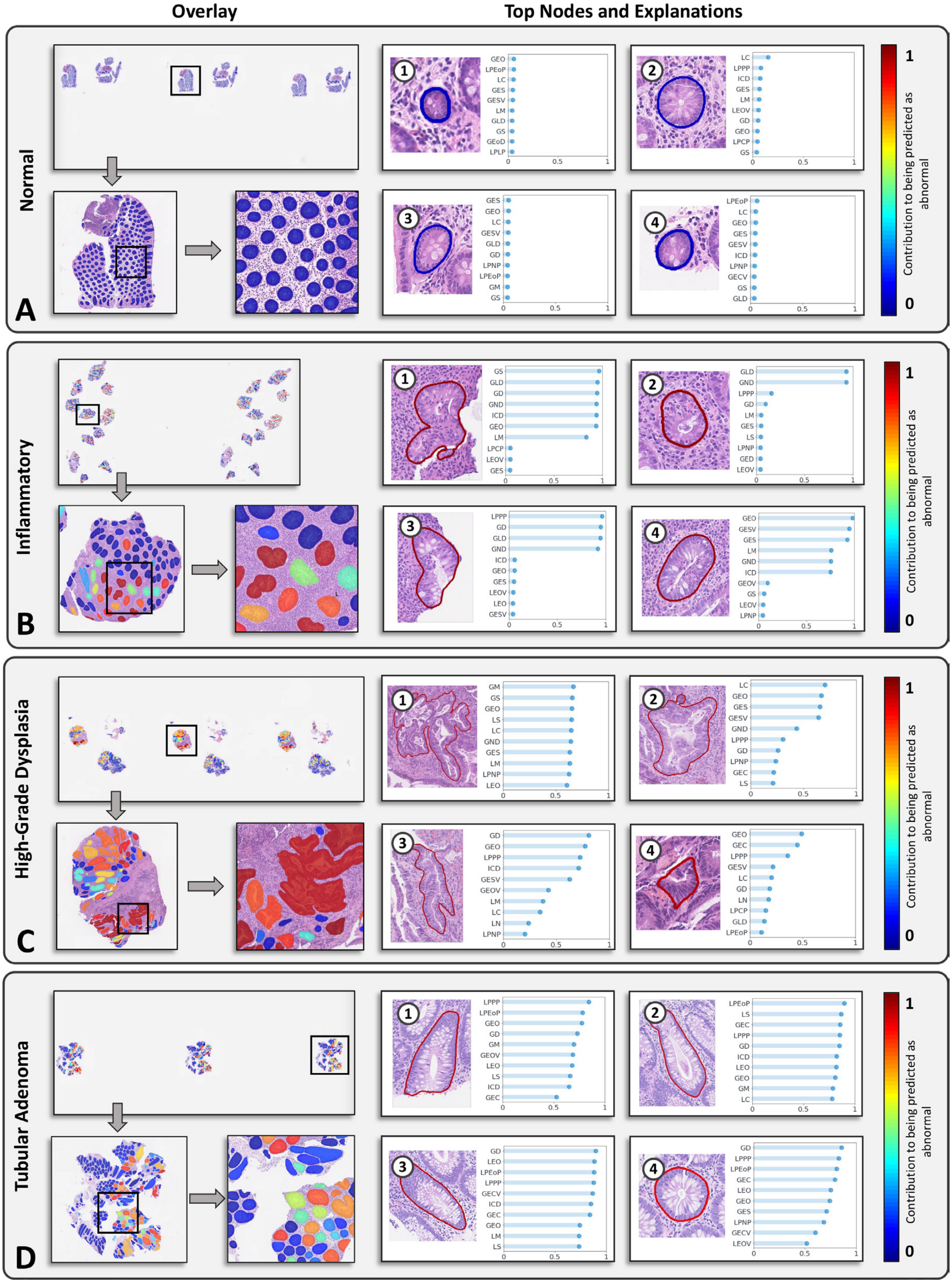


Figure 5 Additional visualisation of node and feature explainability. As before; we display the overlay of the node-level explanations in the form of a heat map showing the most predictive nodes in the WSI. We also show cropped images of the four most predictive nodes within each WSI along with the associated ten most predictive features and their feature importance value. (A–D) Show slides that are normal, inflammatory (with crypt abscesses), high-grade dysplasia or adenomatous polyps, respectively. GEC, gland epithelial clustering; GECV, gland epithelial clustering variation; GED, gland epithelial density; GEO, gland epithelial organisation; GEoD, gland eosinophil density; GEOV, GEO variation; GES, gland epithelial size; GESV, GES variation; GD, gland density; GLD, gland lymphocyte density; GM, gland morphology; GND, gland neutrophil density; GS, gland size; ICD, Inflammatory cell density; LEO, Lumen epithelial organisation; LEOV, Lumen epithelial organisation variation; LPCP, lamina propria connective proportion; LPEoP, lamina propria eosinophil proportion; LPLP, lamina propria lymphocyte proportion; LPNP, lamina propria neutrophil proportion; LPPP, lamina propria plasma proportion; WSIs, whole-slide images.

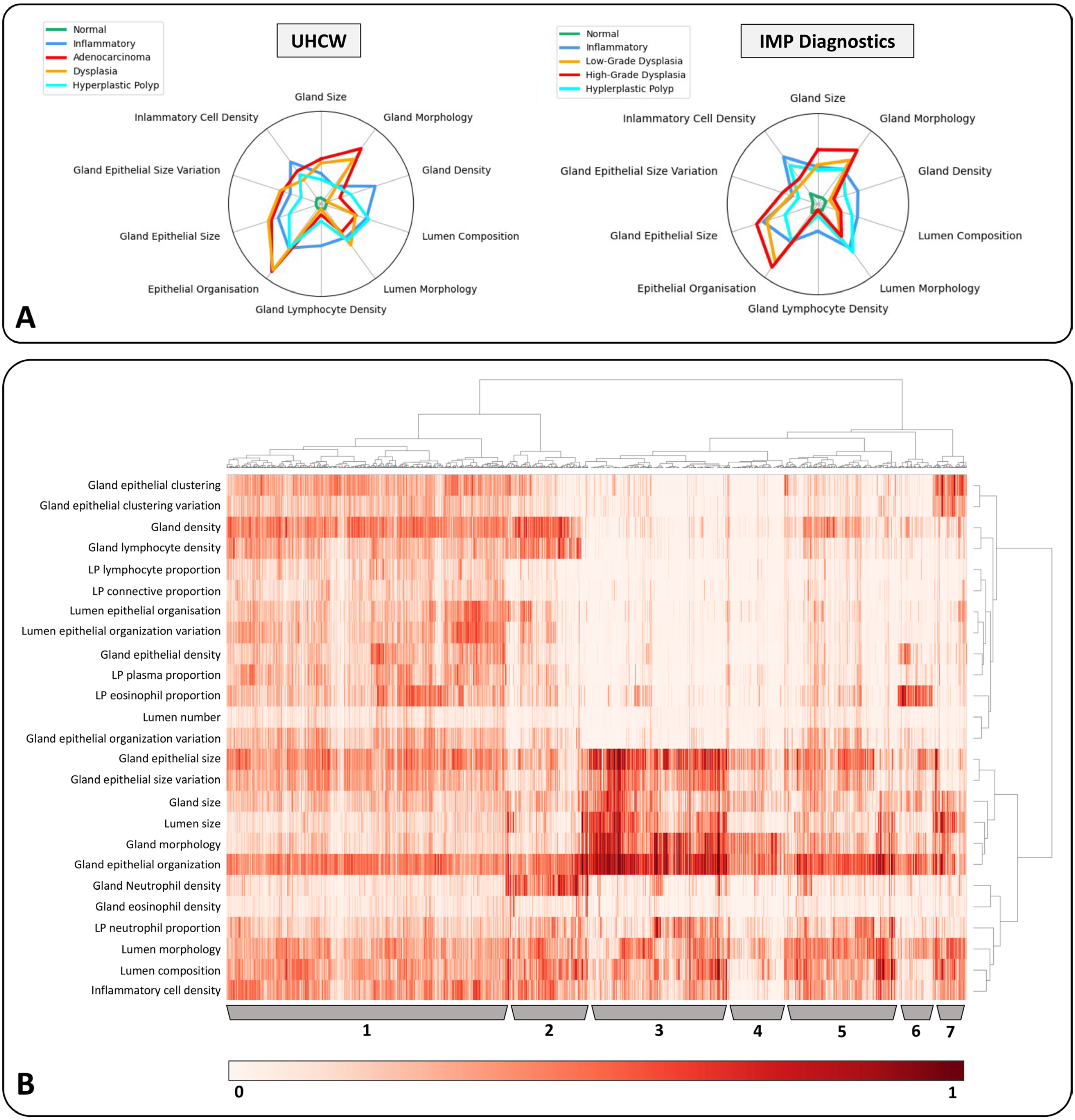


Figure 6 Analysis of global explanations. (A) Radar plots showing global feature importance for subconditions in the UHCW and IMP datasets. (B) Hierarchical biclustering of feature importance values. 1–7 denote prominent clusters after biclustering, with the following distinguishing histological characteristics: (1) inflammation, without gland neutrophil infiltration; (2) inflammation with both gland lymphocytic and neutrophilic infiltration; (3) neoplasia with irregular gland morphology and large epithelial cells; (4) irregular gland morphology with minimal inflammation; (5) hyperplasia with irregular lumen morphology and composition with inflammation in the lamina propria; (6) eosinophilic infiltration in the lamina propria and (7) neoplasia with gland epithelial cell clustering. UHCW, University Hospitals Coventry and Warwickshire.

that the model found slides with lymphocytic and collagenous colitis somewhat challenging, with false positive rates of 0.29 and 0.46, respectively. Explicit modelling of the subepithelial collagen band should enable us to better detect collagenous colitis. It may be worth noting that there was a relatively small number of collagenous colitis samples in all four cohorts and so they may not have a large impact on the overall performance. Also, a high false negative rate was observed in the mild inflammation category, but this is to be expected because they are visually similar to normal samples.

In online supplemental figure 9, we show that our model output is well calibrated and hence can be interpreted as a measure of confidence. To enable explainable predictions, our algorithm relies on an accurate intermediate segmentation step, which requires many pixel-level annotations. This can be a time-consuming step and can therefore act as a bottleneck in the development of similar methods. In addition, the type of features that can be incorporated into our AI algorithm are dependent on which kinds of histological objects are initially localised. For example, we do not currently detect goblet cells and so do not include features indicative of goblet cell-rich hyperplastic polyps. Other histological objects that could be added include giant cells, signet ring cells and mitotic figures. In addition, although we segment the surface epithelium, we do not extract any associated features that can help identify conditions such as collagenous colitis. Our method also does not assess surface abnormality to detect intestinal spirochaetosis or pigment to detect melanosis coli. These shortcomings will be addressed in future work. Visual examples of features used within our framework, along with examples from the 5th and 95th percentiles, are given in online supplemental figure 10. We also provide a more in-depth description of these features, along with what conditions they can detect in online supplemental table 11. In online supplemental figure 1b, we highlight diagnostic features (in a red colour) that are not currently modelled in our framework.

There have been recent AI approaches developed for cancer detection in colonic WSIs.[24 40 41] However, such approaches cannot be used for screening in clinical practice because they often fail to identify non-cancerous abnormalities such as inflammation. Similarly, AI models have been developed for detecting polyps,[42 43] inflammatory bowel disease[44] or grading dysplasia,[25] but again they do not address the problem of screening normal from all types of abnormality. Our approach uses retrospective biopsies from pathology archives, where data are accordingly labelled as normal or abnormal to reflect the clinical screening process. Therefore, unlike other approaches, our AI model can be directly implemented as a triaging tool and may therefore have a profound effect on reducing pathologist workloads. In addition, most recent automatic methods rely on weak supervision, where only the overall diagnosis is used to guide the algorithm. This strategy may be advantageous because it does not rely on the time-consuming task of collecting many annotations. However, this limits the interpretability of the output, which may hinder the acceptance of such models in hospitals.

Analysis of colon biopsy slides by visual examination, either under the microscope or more recently on the computer screen, is the current gold standard. However, the current practice is unsustainable with increasing numbers of specimens that require examination and due to staff shortages, where only 3% of NHS hospitals report adequate staffing.[3] With advances in cancer screening programmes and no immediate sign of the pathologist staffing crisis being resolved, additional measures to assist with reporting will be essential. Our proposed AI model addresses this unmet need by automatically filtering out normal colonic biopsies that require minimal intervention, yet make up a substantial proportion of all cases, with high degree of accuracy. As a result, our model significantly reduces the number of samples that require review by pathologists.

AI models are now starting to be used in clinical practice for prostate cancer detection, where a clear advantage for clinicians has been demonstrated in terms of reducing workloads and increasing reporting accuracy.[45 46] There is a growing evidence that automated methods for tissue diagnosis can transform pathologist workflows and help drive new policies in healthcare. However, no such tool currently exists for screening large bowel endoscopic biopsies, perhaps due to the fact that no automated tool has been able to accurately detect all kinds of abnormality, including inflammation, dysplasia, hyperplasia and neoplasia. With its triaging capability, the proposed model promises to have positive implications on patient treatment due to faster time to diagnosis, resulting in the potential for early intervention where it is needed the most.

The proposed model may be particularly advantageous in low-income countries, where there exists an even greater shortage of pathologists. Despite the obvious benefit of outsourcing tasks to AI in these countries, there still remains a lack of infrastructure for digital pathology, which is a requirement for our approach. A few options may be explored to overcome this challenge, such as using digital mobile phone cameras,[35 47] acquiring low-cost consumer-grade scanners and obtaining them via financing, leasing, philanthropic sources or non-profit organisations. Rather than investing in expensive hardware and performing full clinical integration, a cloud-based setup may be a more affordable option in low-resource settings, where scanned slides can be uploaded to the internet for processing. With AI models now appearing rapidly on the market, it is becoming increasingly important for initiatives to be put in place by policy-makers to help with the digitisation of pathology labs across the world, enabling the widespread adoption of computational pathology.

We have shown that IGUANA offers promise as an effective tool for AI-based colon biopsy screening with a strong emphasis on diagnostic interpretability providing concrete justification as to why a certain diagnostic class was predicted and making its predictions transparent and explainable. The proposed AI method can help alleviate current issues in pathologist shortages in the NHS and worldwide and reduce turnaround times of the screening results. Before deployment in clinical practice, a larger scale validation is required with further analysis of IGUANA's feature explanation output. In addition, considerable time needs to be invested into extending the current user interface so that it easily integrates with current pathologists' clinical workflows. This will involve a detailed study on the effectiveness of the decision support tool within abnormal biopsies and assessing its implications on time to report the diagnosis.

Twitter Simon Graham @simongraham73, Fayyaz Minhas @fayyazhere, Mohsin Bilal @bilal_mohsin, Ayesha Azam @AyeshaSAzam, Shan E Ahmed Raza @sea_raza and Nasir Rajpoot @nmrajpoot

Acknowledgements We acknowledge Dr SA Sanders and Dr Naresh Chachlani for their assistance in providing WSIs from South Warwickshire NHS Foundation Trust.

Contributors SG, DS and NR designed the study with support from all coauthors. SG led the development of the method with support from FM and NR. SG wrote the code and carried out the experiments. MB provided results using IDaRS and CLAM. MA, YWT, EH, KD, HS, AR, SW, AA, KB, MN, KH, HE, KG and DS provided diagnostic annotations of colonic biopsy slides. SG, FM, MA, KG, DS and NR performed analysis and interpretation of the results. MB, MJ, NW, WL, AB and SEAR provided technical and material support. SG, FM, DS and NR were all involved in the drafting of the paper. NR is the guarantor of the study. All authors read and approved the final paper. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. DS and NR are last authors.

Funding All authors 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 grant EP/W02909X/1.

Competing interests SG, DS and NR are co-founders of Histofy. DS reports personal fees from Royal Philips, outside the submitted work. NR and FM report research funding from GlaxoSmithKline.

Patient and public involvement Patients and/or the public were involved in the design, or conduct, or reporting, or dissemination plans of this research. Refer to the Methods section for further details.

Patient consent for publication Not applicable.

Ethics approval This study was conducted 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. Data collection and usage of the IMP Diagnostics dataset was performed in accordance with the Portuguese national legal and ethical standards applicable to that cohort.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement WSIs from University Hospitals Coventry and Warwickshire NHS Trust, East Suffolk and North Essex NHS Foundation Trust, and South Warwickshire NHS Foundation Trust will be made available upon successful application to the PathLAKE data access committee. Relevant information on obtaining the data from the IMP cohort can be found in the original publication.

ORCID iDs
Simon Graham http://orcid.org/0000-0002-2214-8212
Fayyaz Minhas http://orcid.org/0000-0001-9129-1189
Mohsin Bilal http://orcid.org/0000-0001-8632-2729
Mahmoud Ali http://orcid.org/0000-0001-7722-6196
Mark Eastwood http://orcid.org/0000-0003-3768-7953
Noorul Wahab http://orcid.org/0000-0002-1251-1559
Mostafa Jahanifar http://orcid.org/0000-0001-5842-0460
Emily Hero http://orcid.org/0000-0003-0863-8263
Wenqi Lu http://orcid.org/0000-0002-7838-0918
Ayesha Azam http://orcid.org/0000-0003-2681-8153
Mohammed Nimir http://orcid.org/0000-0002-9409-6685
Katherine Hewitt http://orcid.org/0000-0001-6602-0141
Shan E Ahmed Raza http://orcid.org/0000-0002-1097-1738
Kishore Gopalakrishnan http://orcid.org/0000-0003-0459-6967
David Snead http://orcid.org/0000-0002-0766-9650
Nasir Rajpoot http://orcid.org/0000-0002-4706-1308

REFERENCES

1 Cancer Research UK. Testing times to come? An evaluation of pathology capacity across the UK. 2016. Available: http://www.cancerresearchuk.org/sites/default/files/testing_times_to_come_nov_16_cruk.pdf
2 Written evidence submitted by the Royal College of Pathologist's digital pathology Committee (CRV0004). 2020. Available: https://committees.parliament.uk/writtenevidence/11168/pdf/
3 The Royal College of Pathologists. Meeting pathology demand. Histopathology workforce census. 2018. Available: https://www.rcpath.org/uploads/assets/952a934d-2ec3-48c9-a8e6e00fcdca700f/Meeting-Pathology-Demand-Histopathology-Workforce-Census-2018.pdf
4 NHS. Colonoscopy results. 2019. Available: https://www.nhs.uk/conditions/colonoscopy/results/
5 Hanna TP, King WD, Thibodeau S, *et al*. Mortality due to cancer treatment delay: systematic review and meta-analysis. *BMJ* 2020;371:m4087.
6 Cancer Research UK. Scoping the future. An evaluation of evaluation of evaluation of endoscopy endoscopy capacity across the capacity across the across the NHS in England. 2015. Available: https://www.cancerresearchuk.org/sites/default/files/scoping_the_future_-_final.pdf
7 Bowel Cancer UK. A quarter of England hospitals in breach of waiting time target for bowel cancer tests. 2017. Available: https://www.bowelcanceruk.org.uk/news-and-blogs/news/a-quarter-of-hospitals-in-breach-of-waiting-time-target-for-bowel-cancer-tests-as-services-reach-crisis-point/
8 Bowel Cancer UK. Unacceptable endoscopy waiting times put launch of new world-class screening programme at risk. 2018. Available: https://www.bowelcanceruk.org.uk/news-and-blogs/news/unacceptable-endoscopy-waiting-times-put-launch-of-new-world-class-screening-programme-at-risk/
9 Loughrey MB, Shepherd NA. The pathology of bowel cancer screening. *Histopathology* 2015;66:66–77.
10 Talbot I, Price A, Salto-Tellez M. *Biopsy pathology in colorectal disease*. CRC Press, 2006.
11 Cappell MS. Reducing the incidence and mortality of colon cancer: mass screening and colonoscopic polypectomy. *Gastroenterol Clin North Am* 2008;37:129–60,
12 Itzkowitz SH, Present DH. Consensus conference: colorectal cancer screening and surveillance in inflammatory bowel disease. *Inflamm Bowel Dis* 2005;11:314–21.
13 Snead DRJ, Tsang Y-W, Meskiri A, *et al*. Validation of digital pathology imaging for primary histopathological diagnosis. *Histopathology* 2016;68:1063–72.
14 Bulten W, Pinckaers H, van Boven H, *et al*. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol* 2020;21:233–41.
15 Lu MY, Chen TY, Williamson DFK, *et al*. AI-based pathology predicts origins for cancers of unknown primary. *Nature* 2021;594:106–10.
16 Bychkov D, Linder N, Turkki R, *et al*. Deep learning based tissue analysis predicts outcome in colorectal cancer. *Sci Rep* 2018;8:3395.
17 Kather JN, Krisam J, Charoentong P, *et al*. Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. *PLoS Med* 2019;16:e1002730.
18 Campanella G, Hanna MG, Geneslaw L, *et al*. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med* 2019;25:1301–9.
19 Ehteshami Bejnordi B, Veta M, Johannes van Diest P, *et al*. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017;318:2199.
20 Graber M, Gordon R, Franklin N. Reducing diagnostic errors in medicine: what's the goal? *Acad Med* 2002;77:981–92.
21 Nakhleh RE. Error reduction in surgical pathology. *Arch Pathol Lab Med* 2006;130:630–2.
22 Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. *PLoS Med* 2018;15:e1002689.
23 Mehrabi N, Morstatter F, Saxena N, *et al*. A survey on bias and fairness in machine learning. *ACM Comput Surv* 2022;54:1–35.
24 Iizuka O, Kanavati F, Kato K, *et al*. Deep learning models for histopathological classification of gastric and colonic epithelial tumours. *Sci Rep* 2020;10:1–11.
25 Oliveira SP, Neto PC, Fraga J, *et al*. Cad systems for colorectal cancer from WSI are still not ready for clinical acceptance. *Sci Rep* 2021;11:14358.
26 Tsuneki M, Kanavati F. Deep learning models for poorly differentiated colorectal adenocarcinoma classification in whole slide images using transfer learning. *Diagnostics (Basel)* 2021;11:2074.
27 Watson DS, Krutzinna J, Bruce IN, *et al*. Clinical applications of machine learning algorithms: beyond the black box. *BMJ* 2019;364:l886.
28 Leslie D. Understanding artificial intelligence ethics and safety. *ArXiv Preprint ArXiv* 2019:190605684.
29 Lu W, Graham S, Bilal N, *et al*. Capturing cellular topology in multi-gigapixel pathology images. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops; 2020
30 Yanning Z, Simson G, Navid Alemi K, *et al*. Cgc-net: cell graph convolutional network for grading of colorectal cancer histology images. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops; 2019
31 Pati P, Jaume G, Foncubierta-Rodríguez A, *et al*. Hierarchical graph representations in digital pathology. *Med Image Anal* 2022;75:102264.
32 Lu W, Toss M, Dawood M, *et al*. SlideGraph+: whole slide image level graphs to predict HER2 status in breast cancer. *Med Image Anal* 2022;80:102486.
33 Ahmedt-Aristizabal D, Armin MA, Denman S, *et al*. A survey on graph-based deep learning for computational histopathology. *Comput Med Imaging Graph* 2022;95:102027.
34 Bilal M, Raza SEA, Azam A, *et al*. Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study. *Lancet Digit Health* 2021;3:e763–72.
35 Lu MY, Williamson DFK, Chen TY, *et al*. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat Biomed Eng* 2021;5:555–70.
36 Graham S, Vu QD, Jahanifar M, *et al*. One model is all you need: multi-task learning enables simultaneous histology image segmentation and classification. *Med Image Anal* 2023;83:102685.
37 Anderson JR, Tighe JR. Staffing crisis in pathology. *Br Med J* 1980;281:1370–2.
38 McGovern PD. Embracing artificial intelligence is the only way to avoid obsolescence. *BMJ* 2018;363:k5208.
39 Public Health England. National breast screening pathology audit. 2021. Available: https://www.gov.uk/government/publications/national-breast-screening-pathology-audit/national-breast-screening-pathology-audit

40 Kalkan H, Nap M, Duin RPW, *et al*. Automated colorectal cancer diagnosis for whole-slice histopathology. *Med Image Comput Comput Assist Interv* 2012;15:550–7.

41 Xu L, Walker B, Liang P-I, *et al*. Colorectal cancer detection based on deep learning. *J Pathol Inform* 2020;11:28.

42 Korbar B, Olofson AM, Miraflor AP, *et al*. Deep learning for classification of colorectal polyps on whole-slide images. *J Pathol Inform* 2017;8:30.

43 Wei JW, Suriawinata AA, Vaickus LJ, *et al*. Evaluation of a deep neural network for automated classification of colorectal polyps on histopathologic slides. *JAMA Netw Open* 2020;3:e203398.

44 Najdawi F, Sucipto K, Mistry P, *et al*. Artificial intelligence enables quantitative assessment of ulcerative colitis histology. *Mod Pathol* 2023;36:100124.

45 Pantanowitz L, Quiroga-Garza GM, Bien L, *et al*. An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. *Lancet Digit Health* 2020;2:e407–16.

46 Raciti P, Sue J, Ceballos R, *et al*. Novel artificial intelligence system increases the detection of prostate cancer in whole slide images of core needle biopsies. *Mod Pathol* 2020;33:2058–66.

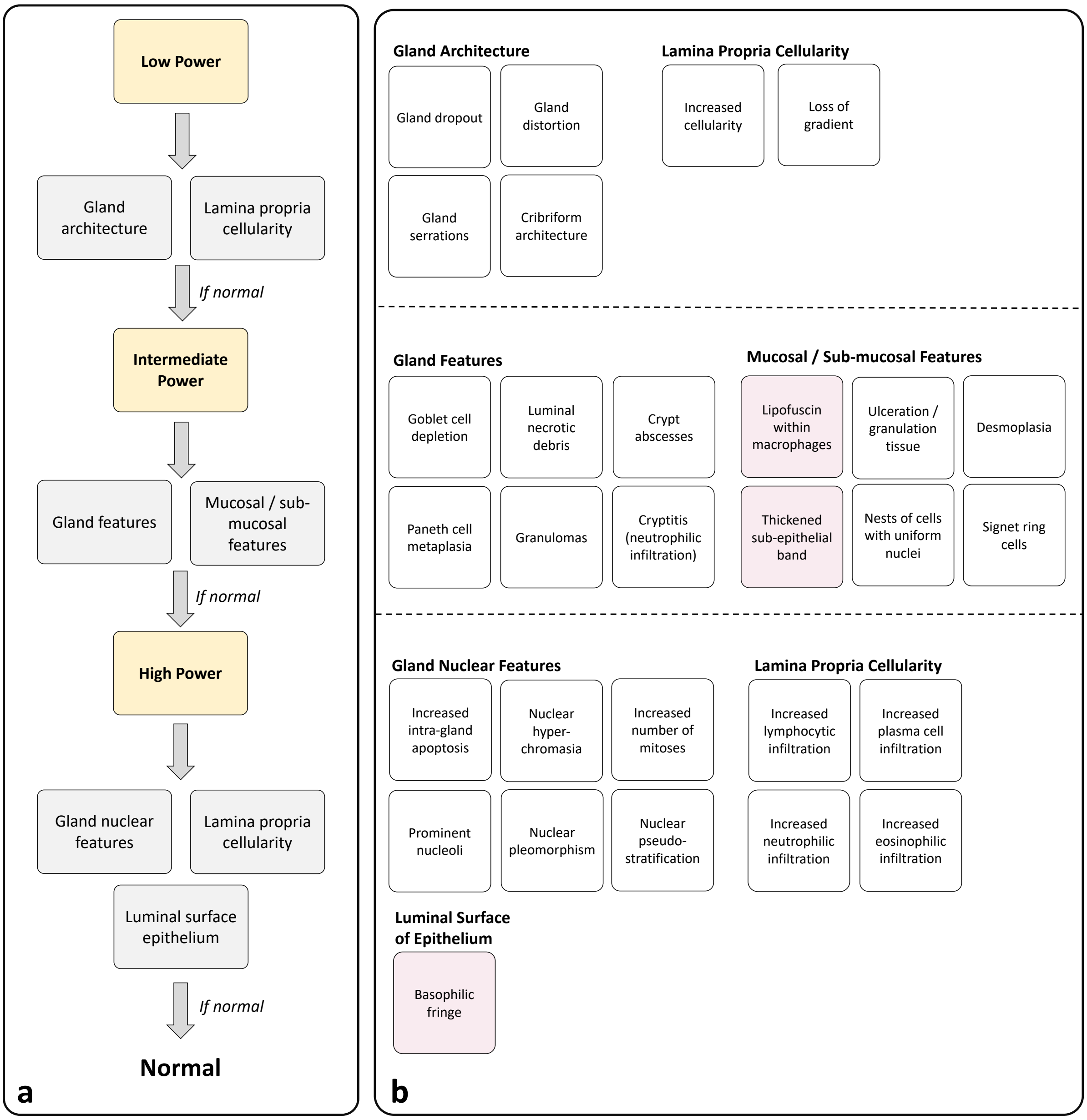
47 Wallis LA, Fleming J, Hasselberg M, *et al*. A smartphone APP and cloud-based consultation system for burn injury emergency care. *PLoS ONE* 2016;11:e0147253.

48 Norgeot B, Quer G, Beaulieu-Jones BK, *et al*. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med* 2020;26:1320–4.

49 Cohen JF, Korevaar DA, Altman DG, *et al*. Stard 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open* 2016;6:e012799.

Supplementary material

S1 Pathologist diagnostic algorithm


Low Power
Gland architecture
Lamina propria cellularity
If normal
Intermediate Power
Gland features
Mucosal / sub-mucosal features
If normal
High Power
Gland nuclear features
Lamina propria cellularity
Luminal surface epithelium
If normal
Normal
a
Gland Architecture
Gland dropout
Gland distortion
Gland serrations
Cribriform architecture
Lamina Propria Cellularity
Increased cellularity
Loss of gradient
Gland Features
Goblet cell depletion
Luminal necrotic debris
Crypt abscesses
Paneth cell metaplasia
Granulomas
Cryptitis (neutrophilic infiltration)
Mucosal / Sub-mucosal Features
Lipofuscin within macrophages
Ulceration / granulation tissue
Desmoplasia
Thickened sub-epithelial band
Nests of cells with uniform nuclei
Signet ring cells
Gland Nuclear Features
Increased intra-gland apoptosis
Nuclear hyper-chromasia
Increased number of mitoses
Prominent nucleoli
Nuclear pleomorphism
Nuclear pseudo-stratification
Lamina Propria Cellularity
Increased lymphocytic infiltration
Increased plasma cell infiltration
Increased neutrophilic infiltration
Increased eosinophilic infiltration
Luminal Surface of Epithelium
Basophilic fringe
b

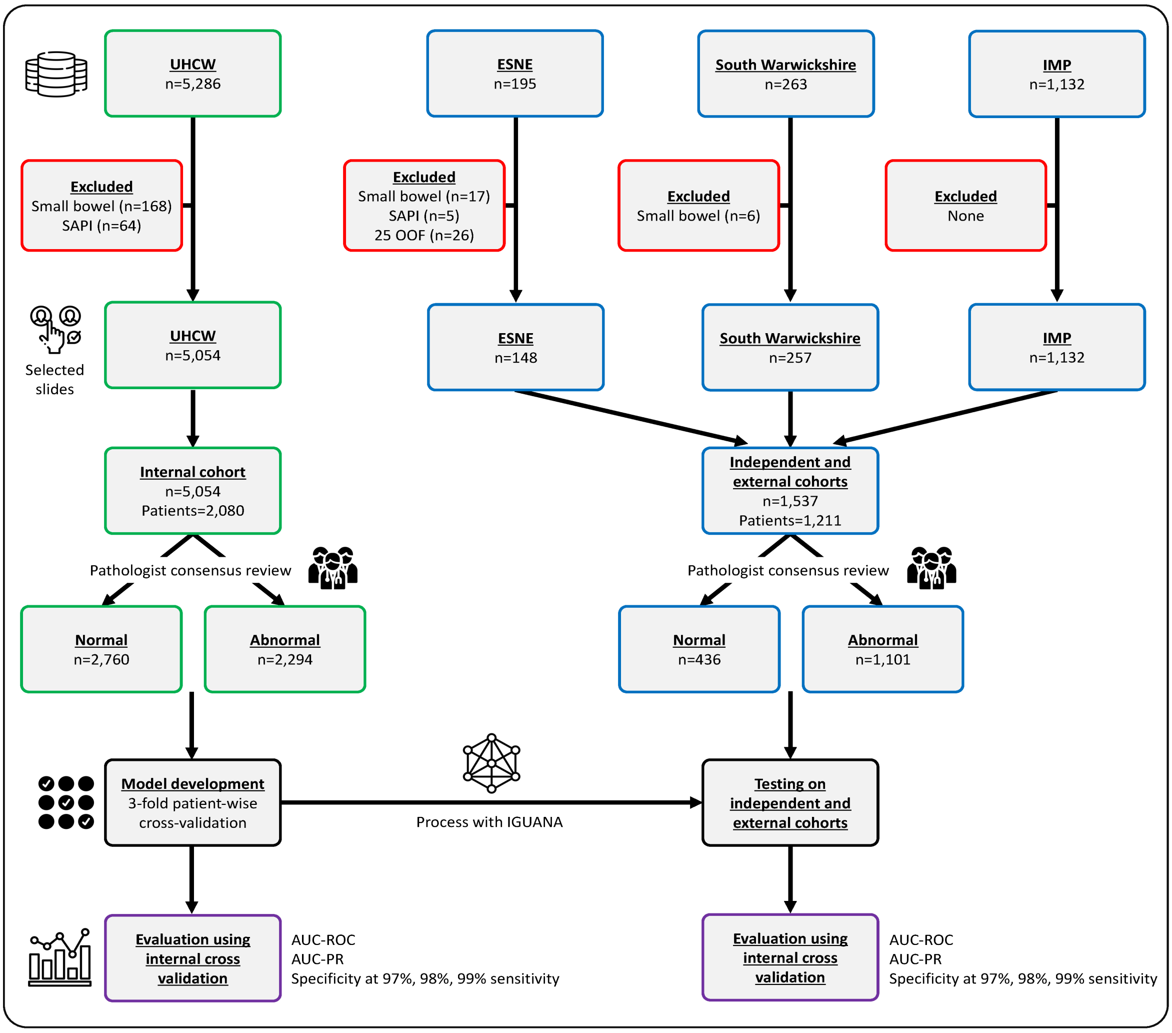
Supplementary Figure 1: Pathologist colon screening diagnostic algorithm. **a**, Decision process for diagnosing colon biopsies as normal. If any abnormal feature is found during this process, then the entire tissue sample is reported as abnormal. **b**, feature breakdown within each category. Red regions show features not yet explicitly modelled in our approach.

S2 Overview of NHS large bowel requests

Supplementary Table 1: Internal audit of seven UK NHS trusts for large bowel biopsies in 2019.

	Histopathology Requests	Large Bowel Biopsies (%)	Normal Large Bowel Biopsies (%)
Coventry	41,771	4,877 (11.7)	1,680 (34.4)
Wolverhampton	52,008	9,708 (18.7)	4,140 (42.6)
Oxford	56,575	7,766 (13.7)	3,938 (50.7)
Nottingham	59,851	10,562 (17.6)	3,428 (32.4)
Newcastle	59,843	5,348 (8.9)	2,015 (37.7)
Durham	34,958	6,240 (17.8)	2,353 (37.7)
Glasgow	108,000	29,000 (13.9)	7,830 (27.0)
Total	**413,006**	**73,501 (17.8)**	**25,384 (34.5)**

S3 Experiment design


UHCW n=5,286
ESNE n=195
South Warwickshire n=263
IMP n=1,132
Excluded Small bowel (n=168) SAPI (n=64)
Excluded Small bowel (n=17) SAPI (n=5) 25 OOF (n=26)
Excluded Small bowel (n=6)
Excluded None
Selected slides
UHCW n=5,054
ESNE n=148
South Warwickshire n=257
IMP n=1,132
Internal cohort n=5,054 Patients=2,080
Independent and external cohorts n=1,537 Patients=1,211
Pathologist consensus review
Pathologist consensus review
Normal n=2,760
Abnormal n=2,294
Normal n=436
Abnormal n=1,101
Model development 3-fold patient-wise cross-validation
Process with IGUANA
Testing on independent and external cohorts
Evaluation using internal cross validation
AUC-ROC
AUC-PR
Specificity at 97%, 98%, 99% sensitivity
Evaluation using internal cross validation
AUC-ROC
AUC-PR
Specificity at 97%, 98%, 99% sensitivity

Supplementary Figure 2: Data description diagram showing the experiment design and the inclusion and exclusion criteria used in the study. Here n denotes the number of whole-slide images, OOF refers to Out-Of-Focus slides and SAPI refers to biopsies that were Slightly Abnormal but Pathologically Insignificant.

S4 Methods

S4.1 WSI datasets

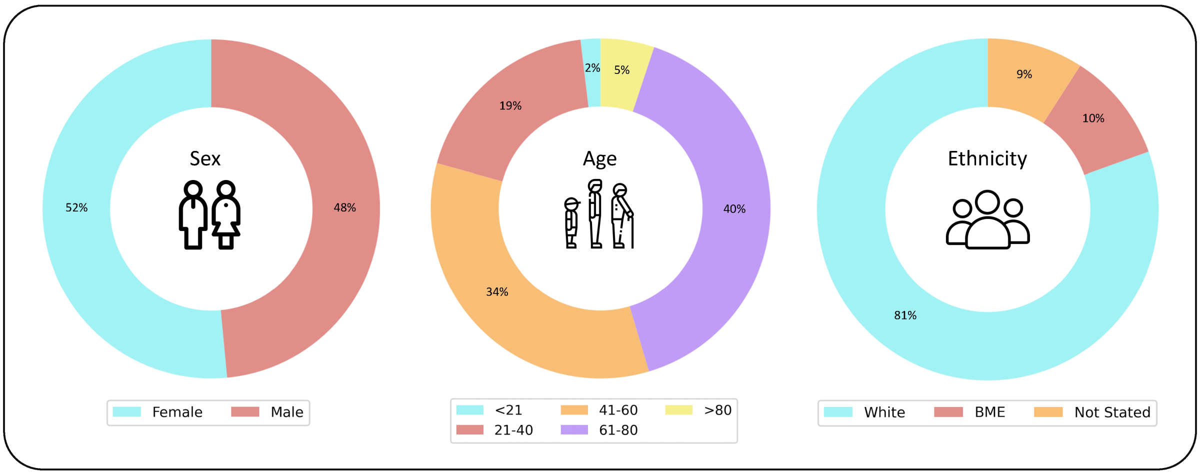
We collected data from four patient cohorts containing routine Haematoxylin and Eosin (H&E) stained WSIs of endoscopic colon biopsies from the following centres: 1) University Hospitals Coventry and Warwickshire (UHCW) NHS Trust, United Kingdom; 2) South Warwickshire NHS Foundation Trust, United Kingdom; 3) East Suffolk and North Essex (ESNE) NHS Foundation Trust, United Kingdom and 4) IMP Diagnostics Laboratory, Portugal[1]. UHCW, South Warwickshire and ESNE WSIs were sampled consecutively, using retrospective data originally scanned between the years 2017-2020. Glass slides from UHCW were digitised with a GE Omnyx slide scanner at a pixel resolution of 0.275 microns per pixel (MPP). Slides from ESNE and South Warwickshire Hospitals were digitised with 3DHISTECH scanners at pixel resolutions of 0.122 MPP and 0.139 MPP, respectively. IMP Diagnostics slides were digitised with a Leica GT450 scanner at a pixel resolution of 0.263 MPP. In total, we collected 6,591 WSIs from 3,291 patients, with 5,054 from UHCW, 148 from ESNE, 257 from South Warwickshire and 1,132 from IMP Diagnostics. To rigorously evaluate our approach for colon biopsy screening, we performed 3-fold internal cross-validation on the UHCW dataset and held out the remaining three datasets for independent external validation. When creating the folds for internal cross-validation, the data was split with stratification at patient-level to ensure that our method was evaluated on completely unseen cases.

Collaborating pathologists categorised WSIs from UHCW, ESNE and South Warwickshire at slide-level into a ground-truth diagnosis label of either normal, non-neoplastic or neoplastic with consensus review of discordant cases. For this study, non-neoplastic and neoplastic classes were combined into a single *abnormal* category to reflect the clinical screening procedure. The slide review was completed by a team spanning five consultant and four trainee pathologists. Normal slides were reviewed by pathologists in training, but all were originally signed out of the laboratory by consultant pathologists holding FRCPath and working at UHCW. All abnormal slides, including SAPI (Slightly Abnormal but Pathologically Insignificant), were reviewed by consultant pathologists holding fellowship of the Royal College of Pathologists and with a sub-specialty interest in GI pathology. SAPI slides, containing only a subtle level of inflammation, were excluded from experiments as they did not clearly belong to either category. Yet, it is important to note that mild inflammatory slides were still included in all experiments. In these cases, distinction between normal and mild inflammation was done by consensus review of slides by consultant pathologists working at UHCW with sub-specialty expertise in GI biopsy reporting.

A wide range of histological conditions were present across the datasets to reflect the clinical screening procedure. Therefore, the cohorts are representative of real-world settings for the clinical question at hand. A full summary of the specific diagnoses is shown in Supplementary Table 2. WSIs from IMP diagnostics were originally categorised as either non-neoplastic, low-grade dysplasia or high-grade dysplasia[1], where the non-neoplastic category contained a mixture of normal, inflammatory and hyperplastic slides. Therefore, our team of pathologists additionally reviewed non-neoplastic slides from IMP to separate normal from abnormal tissue samples. In our final curated datasets, 42% of slides from UHCW, 61% from ESNE, 40% from South Warwickshire and 84% from IMP Diagnostics were labelled as abnormal. We provide a data description diagram showing the experiment design and the inclusion and exclusion criteria used in Supplementary Figure 2. In addition, we give a demographic summary of patients within the development set in Supplementary Figure 3 and provide a more in-depth breakdown in Supplementary Tables 3 and 4. We also give overview of all datasets used in this study in Supplementary Figure 4.

Supplementary Table 2: Summary of all conditions present in the datasets used in the paper.

	UHCW	IMP Diagnostics	ESNE	South Warwickshire
Normal	✓	✓	✓	✓
Ulcerative Colitis	✓	✗	✓	✓
Collagenous colitis	✓	✗	✓	✓
Crohn’s disease	✓	✗	✗	✓
Mild inflammation	✓	✓	✓	✓
Moderate inflammation	✓	✓	✓	✓
Chronic inflammation	✓	✓	✓	✓
CMV	✓	✗	✗	✗
Dysplasia	✓	✓	✓	✓
Hyperplastic polyp	✓	✓	✓	✓
Tubular adenoma	✓	✓	✓	✓
Villous adenoma	✓	✓	✓	✓
Lymphoma	✓	✗	✗	✗
Lipoma	✓	✗	✗	✗
Adenocarcinoma	✓	✓	✓	✓



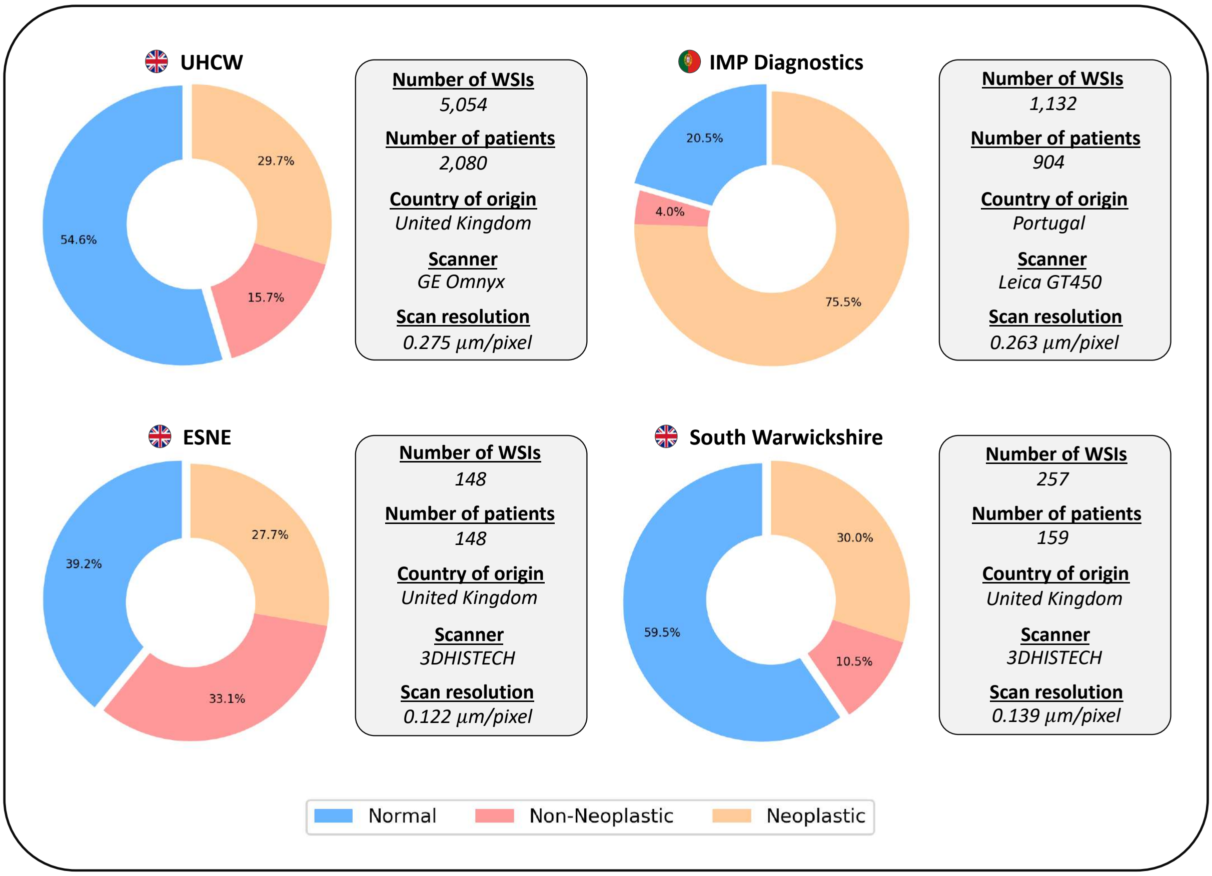
Supplementary Figure 3: Demographic information of the UHCW dataset, that was used for algorithm development. BME denotes black and minority ethnic groups.

Supplementary Table 3: Breakdown of the age of patients present in the UHCW dataset, used for model development.

Age group	Proportion (%)
< 6	0.0
6 to 10	0.1
11 to 15	0.1
16 to 20	1.6
21 to 25	3.1
26 to 30	4.9
31 to 35	5.0
36 to 40	5.9
41 to 45	6.0
46 to 50	7.2
51 to 55	10.6
56 to 60	10.1
61 to 65	10.8
66 to 70	10.5
71 to 75	12.0
76 to 80	6.9
81 to 85	3.9
86 to 90	1.1
> 90	0.2

Supplementary Table 4: Breakdown of the ethnicity of patients present in the UHCW dataset, used for model development.

Ethnicity Summary	Proportion (%)
Asian Bangladesh / British Bangladeshi	0.4
Asian Indian / British Indian	5.6
Asian Other	0.9
Asian Pakistani / British Pakistani	0.6
Black African	0.5
Black Caribbean	0.3
Black Other	0.4
Chinese	0.2
Mixed Other	0.1
Mixed White / Black Caribbean	0.1
Mixed White/ Black African	0.2
Other	1.4
White British	76.9
White Irish	1.4
White Other	2.2
Not stated	9.1



Supplementary Figure 4: Summary of the data used in this study. Despite using WSIs labelled as either normal or abnormal in our experiments, we also show the breakdown of abnormal slides into non-neoplastic and neoplastic categories.

S4.2 Identification of histological objects

The first step of IGUANA requires the segmentation of various histological objects within the WSI, which enables subsequent graph construction and feature extraction. For this, we utilise our recently published Cerberus[2] model, which performs simultaneous segmentation and classification of nuclei, glands, lumen and different tissue regions. Here, the tissue type classification output is used to estimate the lamina propria. During training, we use a multi-task learning strategy, which allows the utilisation of multiple independent datasets and enables simultaneous prediction with a single network. Therefore, our localisation step is computationally efficient and does not require multiple passes through various networks. As well as delineating object boundaries, Cerberus determines the category of each nucleus and differentiates the surface epithelium from other glands. Cerberus is trained on a large amount of data from 12 different centres, including more than 535 thousand nuclei, 51 thousand glands and 56 thousand lumen annotations. In our previous work, we have shown that this crucial step of initial localisation generalises well to unseen examples[2].

S4.3 Extraction of clinically interpretable features

After performing segmentation of various histological objects using Cerberu*s*, we extract a set of clinically meaningful features, which were carefully chosen in collaboration with pathologists so that they reflect what features are considered during the screening procedure. Our model's ability to localise glands, lumen and nuclei within the tissue allows us to extract interesting gland, intra-gland and inter-gland features that are potentially capable of identifying various histological conditions. Specifically, the inter-gland features are defined in the non-glandular surrounding area, also known as the lamina propria. To obtain this region, we utilise the patch-based tissue type classification output from Cerberus and consider both normal gland and tumour patch predictions. We then subtract the gland segmentation output from the prediction map and carry out a series of refinement steps to obtain the final estimation of the lamina propria. We ensure that each feature that we consider has a key biological significance. For example, we consider the size and morphology of glands, which can be indicative of cancer. For quantifying the morphology, we utilise the best alignment metric (BAM)[3], which provides a measure of how elliptical an object is, to help capture abnormal glands with irregular shapes. We also take into account the number of lumen along with their corresponding morphology, which can be suggestive of conditions such as cribriform architecture and serrated polyps, respectively. Furthermore, the organisation of epithelial cells and the amount of different inflammatory cells within the gland are diagnostically informative. For example, normal glands will have epithelial cells organised at the boundary and neutrophils within the gland are indicative of crypt abscesses. For measuring the epithelial organisation, we compute the mean and standard deviation of distances of epithelial nuclei centroids to their nearest gland boundary. We also utilise the mean and standard deviation of inter-epithelial nuclear distances within the gland. Certain inflammatory conditions, such as lymphocytic colitis, will have an increased number of inflammatory cells within the lamina propria. Therefore, we extract inter-glandular features indicative of the local density of inflammatory cells and report the associated cellular composition. Overall, we compute a set of 25 features, which are standardised before utilisation within our graph-based machine learning model.

S4.4 Graph neural networks for computational pathology

Recently, graph neural networks (GNNs) have become popular in Computational Pathology (CPath)[4] due to their ability to model a large WSI as an interconnection of nodes representing histologically important constructs characterized by node-level features[5-7]. An added advantage of using GNNs for predictive modelling in CPath is their ability to generate an explanation of their output in terms of the node level features[8-10].

Existing graph neural networks (GNNs) for computational pathology usually consider fixed-size image patches at each node[11-13] and so fail to incorporate features derived from macrostructures, which can span multiple image patches. However, GNNs that use nodes at centres of image patches in WSIs[14] may have poor interpretability. Instead, nodes can be centred at known histological entities, such as nuclei and glands, allowing pathologists to directly reason with a model's predicted outcome[8]. Although some methods position nodes at known entities, Deep Learning-based features are commonly used[15 16], again leading to reduced interpretability. Rather than using features derived from image regions, graphs built on top of meaningful entities enable the extraction of morphological features. For example, previous methods have constructed graphs on top of nuclei (also known as cell graphs), allowing utilisation of interpretable cellular features. However, nuclei are the most basic building blocks in the tissue and therefore associated features may have limited expressive power, and may fail to model important multi-cellular structures, such as glands. Cell graphs can also be very large, where a single tissue sample can contain tens of thousands of nuclei, leading to the generation of intractable graph models.

S4.5 Gland-graph neural network for accurate large bowel screening

To overcome recent limitations in the literature, our proposed method utilises the concept of gland-graphs for WSI classification, where the nodes are positioned at glands within the tissue, with associated human-interpretable features. The features that we utilise are clinically-meaningful and in line with pathologist diagnostic pathways, leading to excellent performance and providing a highly-interpretable output.

Thus, once the different histological objects have been identified, each WSI is represented as a gland-graph. Here, glands are represented as nodes on a 2D plane that are connected by edges if they are within a certain distance of each other. Each gland is then associated with a set of 25 features that were previously described. Therefore, the overall graph provides a mechanism for representing local features across the entire tissue sample. As opposed to surgical resections, which usually contains a large bulk of tissue, biopsies can contain many separate tissue segments on the slide. This arrangement has no biological significance and, therefore, it would be unreasonable for glands to be connected between neighbouring tissue regions. Thus, we also ensure that an edge between any two given glands only exists if they are both located within the same tissue segment.

Mathematically, a graph is defined as $G \equiv (V, E)$, where V is a set of N vertices (or nodes) and E is a set of edges, where $e_{i,j} \in E$ denotes an edge between nodes i and $j \in V$. In our case, V describes the set of all glands in a WSI. Each node typically has an associated k-dimensional feature vector $\boldsymbol{x}_i$ for $i \in V$. In existing methods, an edge $e_{i,j}$ is constructed if the Euclidean distance between the centroids of nodes i and j is less than a certain threshold[7 17]. The distance between neighbouring node centroids is suitable for convex node entities, such as nuclei, because centroids will usually be located within the object. However, glands can often be non-convex, especially when they become cancerous. Therefore, we instead define an edge $e_{i,j}$ in our gland-graph if the minimum distance between points on the boundary contours of two glands i and j is less than a certain distance α.

Upon formation of our gland-graph representation $G \equiv (V, E)$ of a WSI in terms of its nodes V and their non-directional edges E, we pass the input through a GNN, which sequentially aggregates features within the slide to predict the diagnosis. Each node vector $\boldsymbol{x}_i$ represents a gland in terms of the previously described features and the GNN aggregates information across nodes using the edges in its computation. Note that the number of nodes and edges in each graph can be different depending upon the tissue structure. The GNN first applies a linear operation on $\boldsymbol{x}_i \in \mathbb{R}^{25}$ to produce another node-level feature representation h_i^0 for input into two Principal Neighbourhood Aggregation (PNA) graph convolution[18] layers. Each PNA layer ($l = 1{,}2$) updates each node representation by aggregating information from its neighbours $j \in \aleph_i$ according to the following rule: $h_i^l = \gamma_l\left(h_i^{l-1}, \bigoplus_{j \in \aleph_i} \rho_l\left(h_i^{l-1}, h_j^{l-1}\right)\right)$, where γ_l and ρ_l are multi-layer perceptrons (MLPs) each with their own trainable weights. PNA uses a combination of aggregation strategies (denoted by $\bigoplus$) based on scaling of mean, standard deviation, minimum and maximum aggregation operators over node features. It has been shown recently that using this aggregation approach is superior to methods that use a single aggregation step, such as computing the sum as it enables the resulting GNN to be a better discriminator of local graph structures[18]. Outputs of the two PNA layers are concatenated and fused with a linear operation to arrive at the final node-level feature embedding $\boldsymbol{r}_i$. The final output $\boldsymbol{f}(G) \in \mathbb{R}^C$ is obtained by performing global attention pooling:

$$\boldsymbol{f}(G) = \sum_{i=1}^{N} \frac{\exp\left(\psi(\boldsymbol{r}_i)\right)}{\sum_{j=1}^{N} \exp\left(\psi(\boldsymbol{r}_j)\right)} \odot \omega(\boldsymbol{r}_i)$$

where ψ and ω are MLPs, C denotes the number of classes predicted by the network, $\odot$ denotes element-wise multiplication and hence the global pooling operator learns to assign a varying weight to different gland representations, signifying their relative importance in the final prediction. Finally, a softmax function is applied and all the trainable weights in the GNN are optimised in an end-to-end fashion by minimising the binary cross entropy loss between the output and the ground-truth labels of training slides.

S4.6 Gland-graph neural network for interpretable diagnosis

An important aspect of IGUANA is its ability to provide an interpretable and explainable output, which can be used to facilitate the diagnostic process and potentially for biomarker discovery. For this, we utilise GNNExplainer[9], which generates a subset of nodes and features that play a crucial role in the GNN's prediction. To obtain a WSI-level explanation, the local features are averaged within the top ten most predictive nodes. This enables analysis over larger cohorts to identify existing sub-populations. To further increase model interpretability, we can also visualise the intermediate nuclear, lumen and gland localisation results overlaid on top of the original WSI.

To facilitate this, we utilise the graph pruning method GNNExplainer[9], which generates a subset of nodes and features that play a crucial role in the GNN's prediction. The intuition here is that unimportant nodes and features should have a negligible impact on the performance and can therefore be removed. Specifically, GNNExplainer is formulated as an optimisation task that maximises the mutual information between a GNN's prediction and the distribution of possible subgraph structures. Practically, this is achieved by learning a real-valued mask, which gives less weight to unimportant graph components. For our approach, we learn a node explanation mask $M_n \in \mathbb{R}^N$ and a feature explanation mask $M_f \in \mathbb{R}^{N\times 25}$, where N denotes the number of nodes in each WSI and 25 is the pre-defined number of features. Rather than applying a threshold to the learned masks to give a compact subgraph, we visualise the raw mask output, which gives an interpretable and explainable output that can be discussed with clinicians. Specifically, the learned mask M_n provides the node explanation which can be overlaid on top of the glands in each WSI as a heatmap. Similarly, M_f can be used to identify the top features for each node/gland and the corresponding importance values. To obtain a WSI-level explanation, the local features are averaged within the top ten most predictive nodes. To further increase model interpretability, we can also visualise the intermediate nuclear, lumen and gland localisation results overlaid on top of the original WSI.

To assess which node explanation method was best, we utilised the metric proposed by Jaume *et al.*[8]. The intuition behind their proposed metric is that a superior node explanation technique should be able to locate top nodes that can better differentiate between classes (in our case normal *vs* abnormal). We compared node explanations given by GNNExplainer, integrated gradients[31] and the attention scores given by our network and found that GNNExplainer gave the best results in terms of class separability. Based on this result, we also used GNNExplainer for obtaining feature explanations. It is important to note that GNNExplainer is a *post-hoc* method, which is why we used attention pooling during optimisation of our predictive model.

S4.7 Software, optimisation and reproducibility

We implemented our framework with the open-source software library PyTorch version 1.10[19], PyTorch Geometric version 2.1.1[20] and Python version 3.6 on a workstation equipped with one NVIDIA Tesla V100 GPU. We utilised scikit-learn version 1.0.2[21] to perform the comparative experiments using random forest and fastcluster version 1.2.6[22] to perform biclustering analysis. We trained our graph neural network for 50 epochs using a batch size of 64 and an initial learning rate of 0.005, which was reduced by a factor of 0.2 after 25 epochs. It should be noted, that despite using a GPU with 32GB RAM, our GNN framework incurred a low memory utilisation and therefore different specification GPUs may also be used. The interactive demo was developed using the tile server from TIAToolbox[23] and Bokeh 2.4.3. No changes were made to the AI system or hardware over the course of this study.

Model code, along with a full list of software requirements, is located at https://github.com/TissueImageAnalytics/iguana. Code is shared under a copyleft license, whereas model weights are for research purposes only and are therefore shared under a non-commercial Creative Commons license.

S5 Extended results

S5.1 Detailed comparative results

To assess the performance of our approach, we compare IGUANA with IDaRS[24], CLAM[25] and a random forest classifier using the interpretable glandular features that we extracted (denoted by Gland-RF). Both IDaRS and CLAM are recent top-performing deep learning models that use H&E image patches as input in a multiple-instance learning (MIL) framework. The Gland-RF model computes the mean and standard deviation of all local features within the slide to obtain a fixed-size global feature vector before input to the model. For all approaches, we select the best model in terms of its best AUC-ROC on the validation set. When fitting the Gland-RF for each fold, we perform a grid search over the hyperparameters and select the best models in terms of their performance on the validation set. Despite labels being available, test sets were only processed upon conclusion of cross-validation experiments to prevent test data hacking.

Below, we provide detailed numerical results obtained during the study. For IDaRS and CLAM a prediction is made per input patch and the results are then aggregated to give a single slide-level score. Therefore, we compare the results of IDaRS with different patch aggregation strategies, denoted by *Avg*, *Max* and *AT*. *Avg* and *Max* compute either the average or the maximum score for all patches in the slide, whereas *AT* computes the average score within the top-scoring patches (patches with scores above the slide-level median). Average aggregation is the technique used in the original IDaRS publication.

Supplementary Table 5: Area under the receiver operating characteristic curve (AUC-ROC) for each of the comparative methods. Avg, Max and AT refer to the aggregation strategy used in IDaRS.

	UHCW	**IMP Diagnostics**	**ESNE**	**South Warwickshire**
Gland-RF	0.9581 ± 0.0011	0.9462 ± 0.0064	0.9391 ± 0.0066	0.9575 ± 0.0013
CLAM	0.9658 ± 0.0098	0.9701 ± 0.0019	0.9379 ± 0.0080	0.9492 ± 0.0081
IDaRS (Avg)	0.9738 ± 0.0035	0.9337 ± 0.0147	0.8640 ± 0.0293	0.9267 ± 0.0326
IDaRS (Max)	0.9721 ± 0.0019	0.9279 ± 0.0288	0.8742 ± 0.0414	0.9085 ± 0.0252
IDaRS (AT)	0.9757 ± 0.0030	0.9600 ± 0.0073	0.8791 ± 0.0276	0.8791 ± 0.0314
IGUANA	0.9783 ± 0.0036	0.9789 ± 0.0023	0.9567 ± 0.0155	0.9649 ± 0.0025

Supplementary Table 6: Area under the precision recall curve (AUC-PR) for each of the comparative methods. Avg, Max and AT refer to the aggregation strategy used in IDaRS.

	UHCW	**IMP Diagnostics**	**ESNE**	**South Warwickshire**
Gland-RF	0.9598 ± 0.0045	0.9870 ± 0.0014	0.9683 ± 0.0028	0.9426 ± 0.0021
CLAM	0.9678 ± 0.0099	0.9922 ± 0.0006	0.9589 ± 0.0131	0.9357 ± 0.0131
IDaRS (Avg)	0.9747 ± 0.0013	0.9806 ± 0.0049	0.9168 ± 0.0179	0.9187 ± 0.0281
IDaRS (Max)	0.9728 ± 0.0019	0.9802 ± 0.0071	0.9205 ± 0.0315	0.9067 ± 0.0195
IDaRS (AT)	0.9769 ± 0.0011	0.9894 ± 0.0022	0.9288 ± 0.0161	0.9240 ± 0.0266
IGUANA	0.9798 ± 0.0031	0.9949 ± 0.0006	0.9731 ± 0.0105	0.9466 ± 0.0034

Supplementary Table 7: Specificity at 97% sensitivity. The specificity indicates the percentage reduction in normal slides that require review. Avg, Max and AT refer to the aggregation strategy used in IDaRS.

	UHCW	**IMP Diagnostics**	**ESNE**	**South Warwickshire**
Gland-RF	0.5892 ± 0.0529	0.4356 ± 0.0980	0.2874 ± 0.0800	0.6471 ± 0.0515
CLAM	0.6407 ± 0.1192	0.7852 ± 0.0247	0.5402 ± 0.0533	0.6078 ± 0.0753
IDaRS (Avg)	0.7628 ± 0.0751	0.5407 ± 0.0556	0.3161 ± 0.0722	0.4583 ± 0.2041
IDaRS (Max)	0.7442 ± 0.0358	0.6119 ± 0.0517	0.3621 ± 0.0614	0.2917 ± 0.1482
IDaRS (AT)	0.7717 ± 0.0581	0.6726 ± 0.0337	0.3621 ± 0.1117	0.5088 ± 0.2258
IGUANA	0.7865 ± 0.0350	0.8341 ± 0.0091	0.6552 ± 0.0614	0.7647 ± 0.0489

Supplementary Table 8: Specificity at 98% sensitivity. The specificity indicates the percentage reduction in normal slides that require review. Avg, Max and AT refer to the aggregation strategy used in IDaRS.

	UHCW	**IMP Diagnostics**	**ESNE**	**South Warwickshire**
Gland-RF	0.4846 ± 0.0197	0.3422 ± 0.0930	0.2184 ± 0.1353	0.6362 ± 0.0607
CLAM	0.4989 ± 0.1774	0.6711 ± 0.0725	0.4828 ± 0.0244	0.5948 ± 0.0615
IDaRS (Avg)	0.6725 ± 0.0831	0.4726 ± 0.0588	0.2759 ± 0.0282	0.4035 ± 0.2224
IDaRS (Max)	0.6506 ± 0.0221	0.5481 ± 0.0437	0.3161 ± 0.0775	0.2697 ± 0.1351
IDaRS (AT)	0.6758 ± 0.0749	0.5822 ± 0.0604	0.2759 ± 0.0141	0.4276 ± 0.1723
IGUANA	0.6720 ± 0.0921	0.7378 ± 0.0363	0.5862 ± 0.0488	0.6797 ± 0.0282

Supplementary Table 9: Specificity at 99% sensitivity. The specificity indicates the percentage reduction in normal slides that require review. Avg, Max and AT refer to the aggregation strategy used in IDaRS.

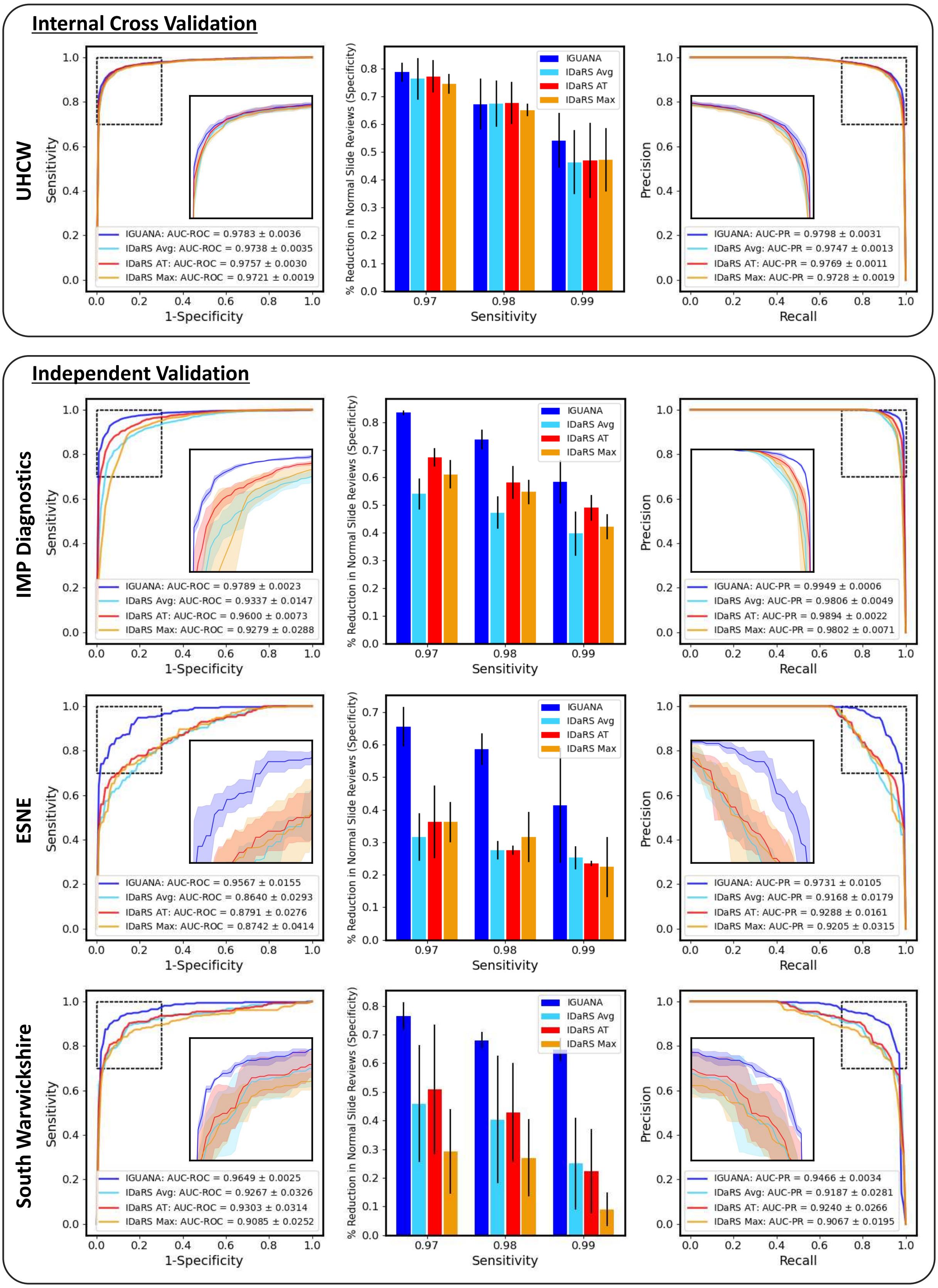
	UHCW	**IMP Diagnostics**	**ESNE**	**South Warwickshire**
Gland-RF	0.3157 ± 0.00478	0.1822 ± 0.0700	0.0517 ± 0.0141	0.5708 ± 0.0263
CLAM	0.3262 ± 0.1771	0.5215 ± 0.0927	0.3218 ± 0.1214	0.3987 ± 0.2067
IDaRS (Avg)	0.4629 ± 0.1144	0.3970 ± 0.0796	0.2529 ± 0.0354	0.2500 ± 0.1604
IDaRS (Max)	0.4706 ± 0.1137	0.4222 ± 0.0453	0.2241 ± 0.0923	0.0899 ± 0.0594
IDaRS (AT)	0.4686 ± 0.1350	0.4904 ± 0.0467	0.2356 ± 0.0081	0.2237 ± 0.1472
IGUANA	0.5409 ± 0.0988	0.5852 ± 0.0789	0.4138 ± 0.1764	0.6471 ± 0.0403

S5.2 Performance across patient subgroups

We perform an analysis of the test performance on the internal UHCW dataset to analyse potential differences in model performance across sex, age, ethnicity and anatomical site of the biopsy. For each subgroup-level analysis, we run 100 bootstrap runs to compute average AUC-ROC and its standard deviation across sub-categories (Supplementary Table 10). We observe that our method is not biased towards any particular subgroup with only minor differences. The effect sizes are quite small for differences in performance across sex and ethnicity, but for age the effect is more pronounced. The effect sizes are also generally small between anatomical sites, but there is a slightly more noticeable effect for the transverse colon. These differences can be due to multiple factors, such as the data used for model training and potential variability in disease patterns within subgroups.

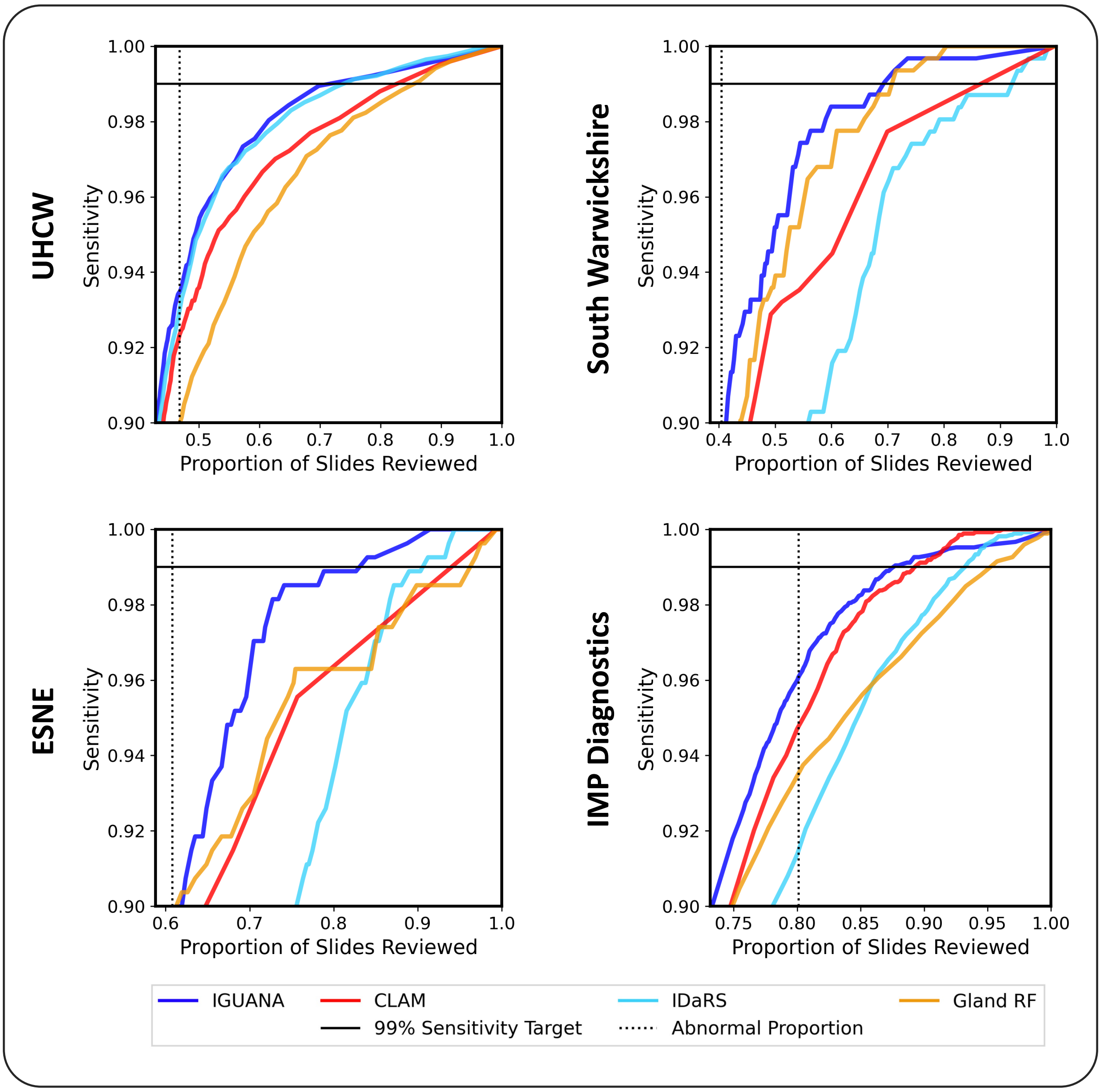
Supplementary Table 10: Performance across different subgroups present in the UHCW dataset.

Group	**Subgroup**	**AUC-ROC**
Sex	Male	0.9774 ± 0.0031
	Female	0.9765 ± 0.0033
Age	< 50	0.9644 ± 0.0047
	> 50	0.9819 ± 0.0024
Ethnicity	White	0.9771 ± 0.0026
	Black & minority ethnic backgrounds	0.9790 ± 0.0082
Anatomical site	Ascending	0.9687 ± 0.0071
	Descending	0.9740 ± 0.0059
	Transverse	0.9822 ± 0.0053
	Sigmoid	0.9772 ± 0.0045
	Rectum	0.9723 ± 0.0060
	Caecum	0.9728 ± 0.0078



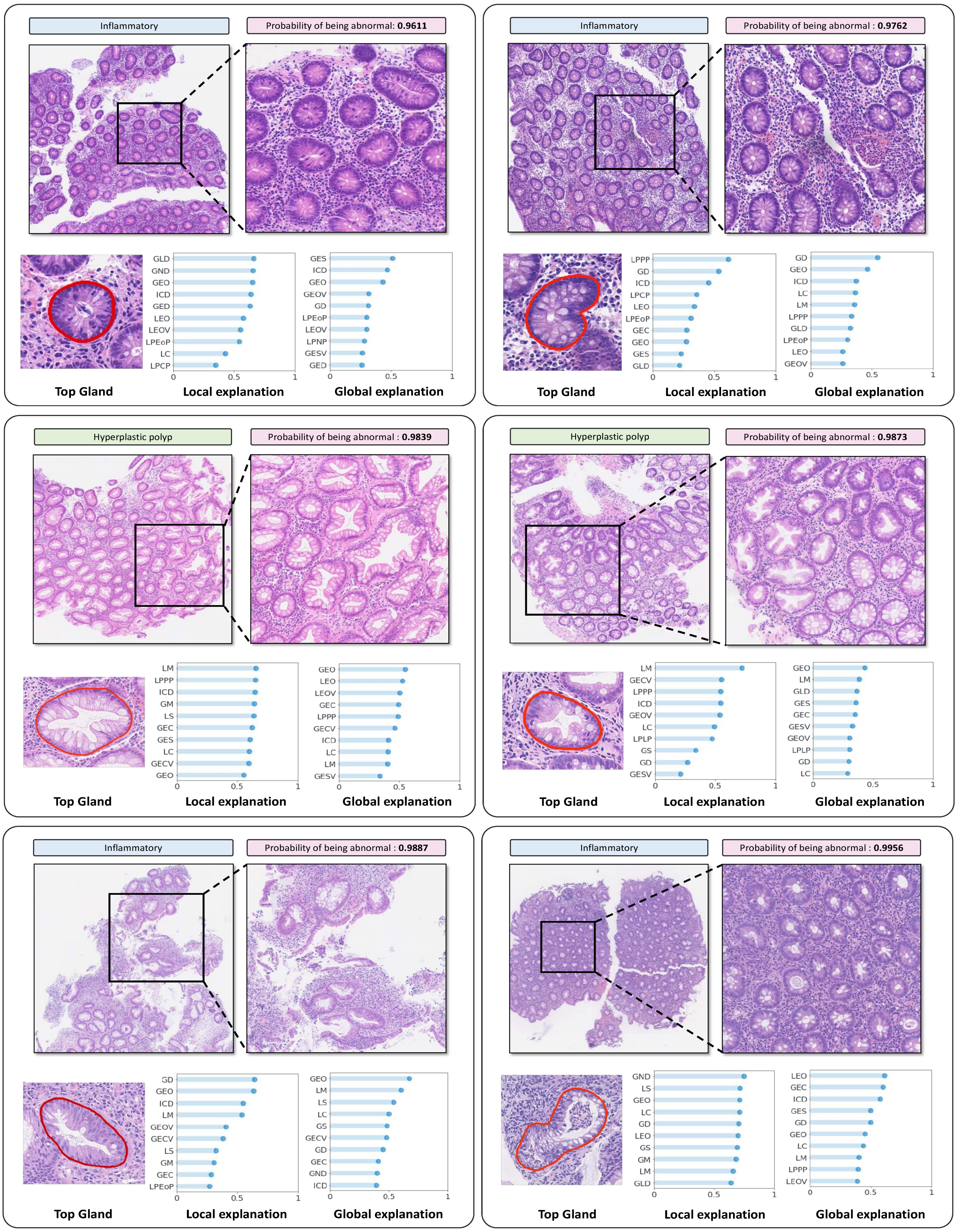
Supplementary Figure 5: Results of IGUANA across the four cohorts used in our experiments compared to IDaRS with different aggregation strategies. Here, AT corresponds to the average of top tiles, where top tiles are those with a score above the median. We display the ROC and PR curves along with the respective AUC scores for each method. We also display the specificities obtained at sensitivity cut-offs of 0.97, 0.98 and 0.99. The shaded areas in the curves and the error bars in the bar plots show 1 standard deviation from the results.

S5.3 Reduction of pathologist workload



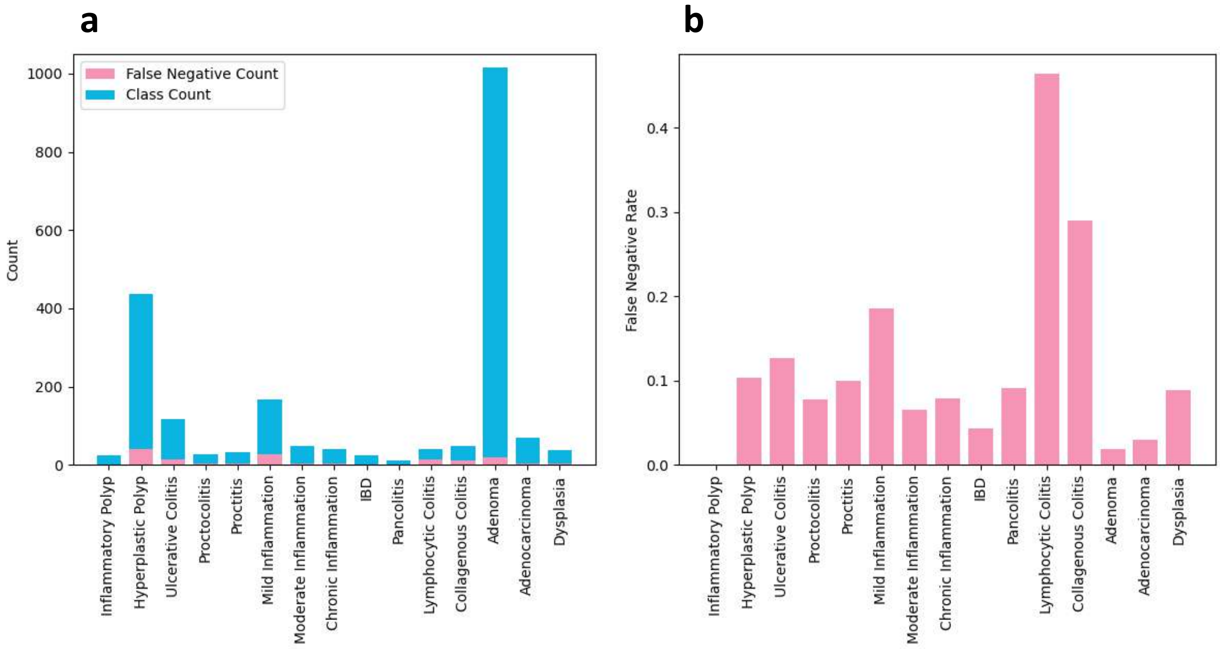
Supplementary Figure 6: Impact of the automatic colon biopsy screening tool on clinical practice. For each dataset we show the proportion of slides that need to be reviewed to ensure a specific sensitivity. Our target sensitivity is 0.99. We also show the proportion of abnormal slides with a vertical dashed line to indicate the minimum number of slides that need to be reviewed to ensure high sensitivity.

S5.4 Analysis of misclassifications



Supplementary Figure 7: Analysis of false positives in the UHCW dataset. In each panel, we display zoomed in images, the gland that contributes most to the prediction, the local feature explanation (corresponding to the top gland) and the global feature explanation. GS: Gland size, GM: Gland morphology, GD: Gland density, LN: Lumen number, LC: Lumen composition, LM: Lumen morphology, LS: Lumen size, GED: Gland epithelial density, GLD: Gland lymphocyte density, GND: Gland neutrophil density, GEoD: Gland eosinophil density, GEC: Gland epithelial clustering, GECV: Gland epithelial clustering variation, GEO: Gland epithelial organisation, GEOV: Gland epithelial organisation variation, LEO: Lumen epithelial organisation, LEOV: Lumen epithelial organisation variation, GES: Gland epithelial size, GESV: Gland epithelial size variation, LPLP: Lamina propria lymphocyte proportion, LPPP: Lamina propria plasma proportion, LPNP:

Lamina propria neutrophil proportion, LPEoP: Lamina propria eosinophil proportion, LPCP: Lamina propria connective proportion, ICD: Inflammatory cell density.

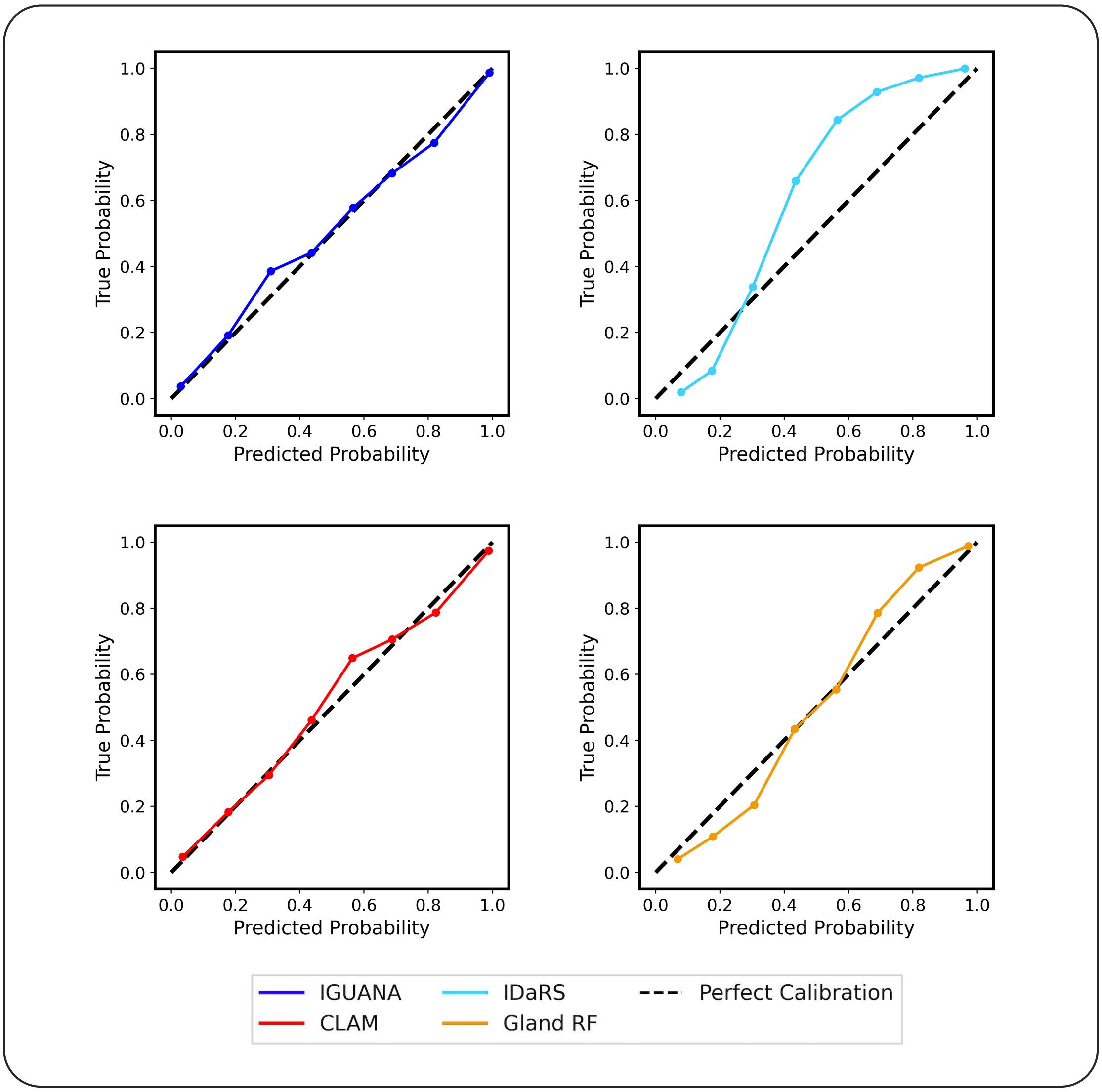


Supplementary Figure 8: False negative analysis. On the left we show the class counts along with the corresponding number of false negatives. On the right, we show the false positive rate per class. For this figure, we don't consider sub-conditions with minimal examples.

S5.5 Model calibration

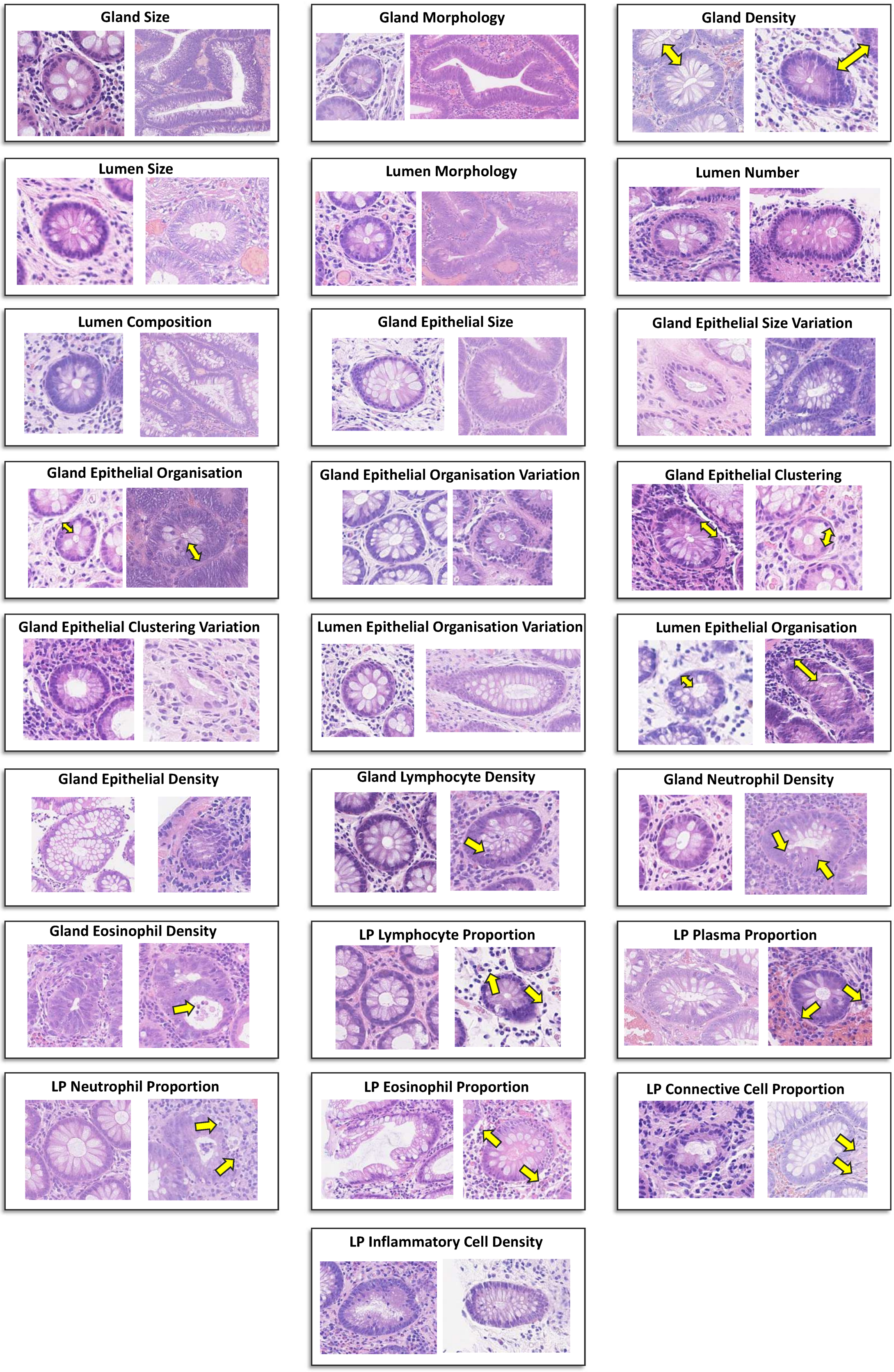
To assess whether the model output can be interpreted as a probability and hence be used as a measure of confidence to guide clinicians, we investigated model calibration. For this, we compare the model output with the true probability, which is calculated by measuring the proportion of correctly classified examples within a certain probability range.

In Supplementary Figure 9, we display model calibration plots for our approach compared to IDaRS, CLAM and Gland-RF on the UHCW dataset. For a perfectly calibrated model, we would expect the plot to lie across the dashed diagonal line, indicating that the model output is equal to the true probability. We observe that both IGUANA and CLAM are fairly well calibrated and so the model output can be used as a proxy to the model confidence. However, after inspection IDaRS is poorly calibrated and therefore outputs should be interpreted with caution. This is due to the aggregation strategy used, where patch-level scores are averaged to give the output.



Supplementary Figure 9: Calibration plot of our approach compared to IDaRS, CLAM and a gland-based random forest. Curves closer to the black dashed line indicate better calibration, which means that model outputs are closely associated with the probability of a prediction being correct. Hence calibrated model outputs can be used as a measure of confidence to better inform decisions.

S6 Interpretable features



Supplementary Figure 10: Examples of cropped image regions from the UHCW dataset containing features taken from the 5th (first image in each panel) and 95th (second image in each panel) percentiles. Yellow arrows show areas within the image relevant to the associated feature.

Supplementary Table 11: Description of all 25 features used in our experiments, along with various conditions that they may be able to detect. LP denotes lamina propria.

Feature Name	Feature Description	Histological Description	Main Conditions Modelled
Gland size	Size of gland (number of pixels at 0.5 microns/pixel)	Gland enlargement	Neoplasia, dysplasia, adenomatous polyps
Gland morphology	How far gland is from being elliptical – BAM distance[38]	Gland distortion, gland branching	Neoplasia, dysplasia. adenomatous polyps
Gland density	Distance to nearest gland	Gland dropout	Inflammation
Lumen size	Size of lumen (number of pixels at 0.5 microns/pixel)	Lumen dilation	Neoplasia, dysplasia, hyperplastic polyps
Lumen morphology	How far lumen is from being elliptical – BAM distance[38]	Lumen serrations	Hyperplastic polyps
Lumen number	Lumen count within gland	Cribriform architecture	Neoplasia
Lumen composition	Ratio of lumen to gland area	Gland dilation	Hyperplasia
Gland epithelial size	Average size of epithelial nuclei within a gland	Epithelial cell atypia	Neoplasia, dysplasia, adenomatous polyps
Gland epithelial size variation	Standard deviation of epithelial nuclei size within a gland	Epithelial cell atypia	Neoplasia, dysplasia, adenomatous polyps
Gland epithelial organisation	Average distance of intra-gland epithelial nuclei to nearest gland boundary	Stratification of epithelial cells	Neoplasia, dysplasia, adenomatous polyps
Gland epithelial organisation variation	Standard deviation of intra-gland epithelial nuclei distances to nearest gland boundary	Uneven stratification of epithelial cells	Neoplasia, dysplasia, adenomatous polyps
Gland epithelial clustering	Average distance between intra-gland epithelial nuclei	Epithelial cells tightly packed	Neoplasia, dysplasia, adenomatous polyps
Gland epithelial clustering variation	Standard deviation of intra-gland epithelial nuclei distances to nearest gland boundary	Epithelial cells unevenly spaced	Neoplasia, dysplasia, adenomatous polyps
Lumen epithelial organisation	Average distance of intra-gland epithelial nuclei to nearest lumen boundary	Gland dilation, cribriform architecture	Neoplasia, dysplasia, hyperplastic polyps
Lumen epithelial organisation variation	Standard deviation of intra-gland epithelial nuclei distances to nearest lumen boundary	Lumen serrations, cribriform architecture	Neoplasia, dysplasia, hyperplastic polyps
Gland epithelial density	Number of intra-gland epithelial nuclei, normalised by the gland size	Solid sheets of epithelial cells	Neoplasia, dysplasia, adenomatous polyps
Gland lymphocyte density	Number of intra-gland lymphocytes, normalised by the gland size	Gland lymphocyte infiltration	Inflammation
Gland neutrophil density	Number of intra-gland neutrophils, normalised by the gland size	Gland neutrophil infiltration (crypt abscess)	Inflammation
Gland eosinophil density	Number of intra-gland eosinophils, normalised by the gland size	Gland eosinophil infiltration	Inflammation
LP lymphocyte proportion	Proportion of lymphocytes within nearest 220 nuclei to gland	Lymphocytic colitis	Inflammation
LP plasma cell proportion	Proportion of plasma cells within nearest 220 nuclei to gland	Colitis	Inflammation
LP neutrophil proportion	Proportion of neutrophils within nearest 220 nuclei to gland	Acute inflammation	Inflammation
LP eosinophil proportion	Proportion of eosinophils within nearest 220 nuclei to gland	Eosinophilic colitis	Inflammation
LP connective tissue cell proportion	Proportion of connective tissue cells within nearest 220 nuclei to gland	Desmoplasia	Inflammation
LP inflammatory cell density	Mean distance of nearest 250 inflammatory nuclei to gland	General inflammation	Inflammation, hyperplastic polyps

S7 Extended discussion

S7.1 Application of the model to surgical resections and other tissues

Despite the screening of endoscopic large bowel biopsies being a focus of this study, the proposed approach could be applied to resection samples with minimal modification. Our method may be especially powerful in this case because each tissue segment within the slide is typically larger, allowing greater spatial context to be explored. Two potential areas of interest using resection samples include the prediction of genetic alterations[26] and survival analysis[27 28]. In these cases, histological biomarkers are less well known, as compared to those used for routine screening tasks. Therefore, our approach might be used to aid biomarker discovery and help toward further understanding of which morphological patterns are associated with certain genetic alterations and clinical outcomes.

In principle, our graph model can target other histological entities as nodes, such as blood vessels, but we chose to focus on epithelial structures in case of large bowel tissue as most large bowel abnormalities are associated with epithelial structures. Our algorithm will naturally translate to other tissues with tubular structures, such as endometrial and breast tissue.

S7.2 Using the application is clinical practice

As well as performing a thorough validation, we will need to ensure that the user interface seamlessly integrates with existing pathologist workflows. Our current interactive solution is a proof of concept, which will inevitably undergo numerous rounds of refinement before it is deployed. In particular, the optimal way to present the top features and overlay to the pathologists is yet to be determined. For this, we envisage a future pathologist user study to collect extensive feedback on the tool.

Supplementary References

1. Oliveira SP, Neto PC, Fraga J, et al. CAD systems for colorectal cancer from WSI are still not ready for clinical acceptance. *Scientific Reports* 2021;11(1):1-15.
2. Graham S, Vu QD, Jahanifar M, et al. One model is all you need: multi-task learning enables simultaneous histology image segmentation and classification. *Medical Image Analysis* 2023;83:102685.
3. Awan R, Sirinukunwattana K, Epstein D, et al. Glandular morphometrics for objective grading of colorectal adenocarcinoma histology images. *Scientific reports* 2017;7(1):1-12.
4. Ahmedt-Aristizabal D, Armin MA, Denman S, et al. A survey on graph-based deep learning for computational histopathology. *Computerized Medical Imaging and Graphics* 2021:102027.
5. Lu W, Graham S, Bilal M, et al. Capturing cellular topology in multi-gigapixel pathology images. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops; 2020.
6. Pati P, Jaume G, Foncubierta-Rodríguez A, et al. Hierarchical graph representations in digital pathology. *Medical image analysis* 2022;75:102264.
7. Zhou Y, Graham S, Koohbanani NA, et al. Cgc-net: Cell graph convolutional network for grading of colorectal cancer histology images. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops; 2019.
8. Jaume G, Pati P, Bozorgtabar B, et al. Quantifying explainers of graph neural networks in computational pathology. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021.
9. Ying Z, Bourgeois D, You J, et al. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems* 2019;32.
10. Luo D, Cheng W, Xu D, et al. Parameterized explainer for graph neural network. *Advances in neural information processing systems* 2020;33:19620-31.
11. Chen RJ, Lu MY, Shaban M, et al. Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. International Conference on Medical Image Computing and Computer-Assisted Intervention; 2021. Springer.

12. Chen RJ, Lu MY, Wang J, et al. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging* 2020.
13. Wang J, Chen RJ, Lu MY, et al. Weakly supervised prostate tma classification via graph convolutional networks. 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI); 2020. IEEE.
14. Li R, Yao J, Zhu X, et al. Graph CNN for survival analysis on whole slide pathological images. International Conference on Medical Image Computing and Computer-Assisted Intervention; 2018. Springer.
15. Anand D, Gadiya S, Sethi A. Histographs: graphs in histopathology. Medical Imaging 2020: Digital Pathology; 2020. SPIE.
16. Xie C, Vanderbilt C, Feng C, et al. Computational biomarker predicts lung ICI response via deep learning-driven hierarchical spatial modelling from H&E. 2022.
17. Wang Y, Wang YG, Hu C, et al. Cell graph neural networks enable the precise prediction of patient survival in gastric cancer. *NPJ precision oncology* 2022;6(1):1-12.
18. Corso G, Cavalleri L, Beaini D, et al. Principal neighbourhood aggregation for graph nets. *Advances in Neural Information Processing Systems* 2020;33:13260-71.
19. Paszke A, Gross S, Massa F, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 2019;32.
20. Fey M, Lenssen JE. Fast graph representation learning with PyTorch Geometric. *arXiv preprint arXiv:190302428* 2019.
21. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 2011;12:2825-30.
22. Müllner D. fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. *Journal of Statistical Software* 2013;53:1-18.
23. Pocock J, Graham S, Vu QD, et al. TIAToolbox as an end-to-end library for advanced tissue image analytics. *Communications medicine* 2022;2(1):1-14.
24. Bilal M, Raza SEA, Azam A, et al. Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study. *The Lancet Digital Health* 2021;3(12):e763-e72.
25. Lu MY, Williamson DF, Chen TY, et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering* 2021;5(6):555-70.
26. Kather JN, Heij LR, Grabsch HI, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature cancer* 2020;1(8):789-99.
27. Kather JN, Krisam J, Charoentong P, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine* 2019;16(1):e1002730.
28. Zhu W, Xie L, Han J, et al. The application of deep learning in cancer prognosis prediction. *Cancers* 2020;12(3):603.