Enabling large-scale screening of Barrett's esophagus using weakly supervised deep learning in histopathology

Kenza Bouzid[1*], Harshita Sharma[1*], Sarah Killcoyne[2], Daniel C. Castro[1], Anton Schwaighofer[1], Max Ilse[1], Valentina Salvatelli[1], Ozan Oktay[1], Sumanth Murthy[2], Lucas Bordeaux[2], Luiza Moore[3], Maria O'Donovan[2,3], Anja Thieme[1], Aditya Nori[1], Marcel Gehrung[2†], and Javier Alvarez-Valle[1†]

[1] *Microsoft Health Futures, Cambridge, UK*
[2] *Cyted Ltd, Cambridge, UK*
[3] *Department of Histopathology, Addenbrookes Hospital, Cambridge University NHS Foundation Trust, Cambridge, UK*

[*] *These authors contributed equally to this work.*
[†] *Corresponding authors: marcel.gehrung@cyted.ai, jaalvare@microsoft.com*

Timely detection of Barrett's esophagus, the pre-malignant condition of esophageal adenocarcinoma, can improve patient survival rates. The Cytosponge-TFF3 test, a non-endoscopic minimally invasive procedure, has been used for diagnosing intestinal metaplasia in Barrett's. However, it depends on pathologist's assessment of two slides stained with H&E and the immunohistochemical biomarker TFF3. This resource-intensive clinical workflow limits large-scale screening in the at-risk population. Deep learning can improve screening capacity by partly automating Barrett's detection, allowing pathologists to prioritize higher risk cases. We propose a deep learning approach for detecting Barrett's from routinely stained H&E slides using diagnostic labels, eliminating the need for expensive localized expert annotations. We train and independently validate our approach on two clinical trial datasets, totaling 1,866 patients. We achieve 91.4% and 87.3% AUROCs on discovery and external test datasets for the H&E model, comparable to the TFF3 model. Our proposed semi-automated clinical workflow can reduce pathologists' workload to 48% without sacrificing diagnostic performance.

Introduction

Early detection of cancer offers the best chance of long-term survival and good quality of life for patients. This is the main driver behind initiatives aimed at early detection in esophageal adenocarcinoma (EAC), which has a poor 5-year survival rate below 20%[1], primarily due to late diagnosis[2]. Barrett's esophagus (BE) is the pre-malignant tissue that presents an opportunity to detect and treat EAC early. However, it is estimated that less than 20% of patients with BE are diagnosed[3], resulting in the majority of EAC cases being diagnosed without the possibility of early treatment.

Currently the standard diagnostic test for BE is endoscopic biopsies with histopathology in patients who are at higher risk due to gastroesophageal reflux disease (GERD) symptoms. Considering the high prevalence of GERD (10–30%) in the adult population[4], screening at scale is challenging, as endoscopy is resource-intensive. Of these patients, an estimated 5–12% will be diagnosed with BE[5,6]. Increasing the detection and monitoring of BE is therefore a priority for EAC early diagnosis and treatment.

In recent years, minimally invasive capsule sponge devices such as the Cytosponge have been developed to enable large-scale screening. The capsule sponge samples cells throughout the length of the esophagus in a short procedure performed by a nurse in a clinic, and accurately identifies patients with BE or early cancer when coupled with specific biomarkers on a slide[7–10]. The biomarker trefoil factor 3 (TFF3) identifies goblet cells, the hallmark for intestinal metaplasia (IM) in BE[8,11], and the biomarker p53 detects malignant transformation of BE[12]. Lastly, hematoxylin and eosin (H&E) staining is used for cellular atypia as an indicator of pre-malignant changes.

Similar to endoscopic biopsies, these tests rely on manual inspection of histopathology specimens by pathologists for diagnosis. In current practice, a histopathologist inspects each of the three slides (TFF3, H&E, p53) for every patient. A BE diagnosis is made by inspecting both H&E for cellular morphology and TFF3 for goblet cells[7]. The complete visual inspection of a case may take the pathologist 8–10 minutes. As with any large-scale screening, pathologist spends most of their time examining negative cases, instead of prioritizing high-risk cases; this limits the scalability of the test for population-level screening.

Deep learning has demonstrated potential to improve screening coverage with successful application to digital histopathological images[13]. Such methods have illustrated promising diagnostic performance in cancer detection and classification[14,15]. In recent years, deep learning approaches have been shown to learn the spatial organization of cells in tumors[16], classify different tumor types[17], and learn histopathological characteristics related

to the underlying genomic mutations[18]. A common approach to image classification problems in histopathology has been supervised deep learning based on localized expert annotations in whole-slide images[9], where smaller regions in such large images (several gigabytes in size) need to be visually inspected and manually annotated by pathologists. Creating such annotations is time- and resource-intensive, limiting the scalability of the supervised deep learning methods for model development. Recent improvements in weakly supervised deep learning[19,20] based on the multiple instance learning (MIL) paradigm[21], enable model training from whole-slide images, using the diagnostic information rather than specialized local annotations. This makes it possible to leverage existing datasets composed of slides and pathologists' routine diagnostic reports, including labels for adjacent slides of the same patient. The approach can be scaled to larger sample sizes for training more effective and robust deep learning models. Consequently, the approach can be extended to new domains such as screening examinations for other cancer types in diagnostic histopathology.

In this paper, we propose a weakly supervised deep learning approach based on MIL to predict the presence of IM for the detection of BE directly from the routinely stained H&E whole-slide images (slides) of capsule sponge samples, without requiring TFF3 staining (Fig. 1a–b). We develop and test our approach on a discovery dataset from the DELTA implementation study and externally validate it on the BEST2 multi-center clinical trials dataset[7] (Fig. 1c). To ensure model interpretability, we conduct qualitative, quantitative and failure-modes analysis of the deep learning model outputs. We propose two semi-automated machine learning (ML)-assisted clinical workflows for Barrett's screening, which can considerably reduce the pathologists' manual workload to 48% without loss in diagnostic performance, and TFF3 staining to 37%, respectively.

Results

Weakly supervised deep learning models can accurately detect BE from H&E and TFF3 slides. The discovery dataset consists of 1,141 patient samples with paired pathology data containing adjacent H&E and TFF3 slides. It was randomly divided into development and test datasets using an 80:20 split (Fig. 1c). Two models were trained using four-fold cross-validation on the discovery development set following the BE-TransMIL model architecture (Fig. 1b, Methods), and evaluated on the discovery test set. The first model, H&E BE-TransMIL, addresses the main aim of detecting BE from H&E slides directly. The second model, TFF3 BE-TransMIL, was trained on TFF3 slides—this is intuitively an easier computer vision task, as goblet cells are distinctively stained as dark brown.

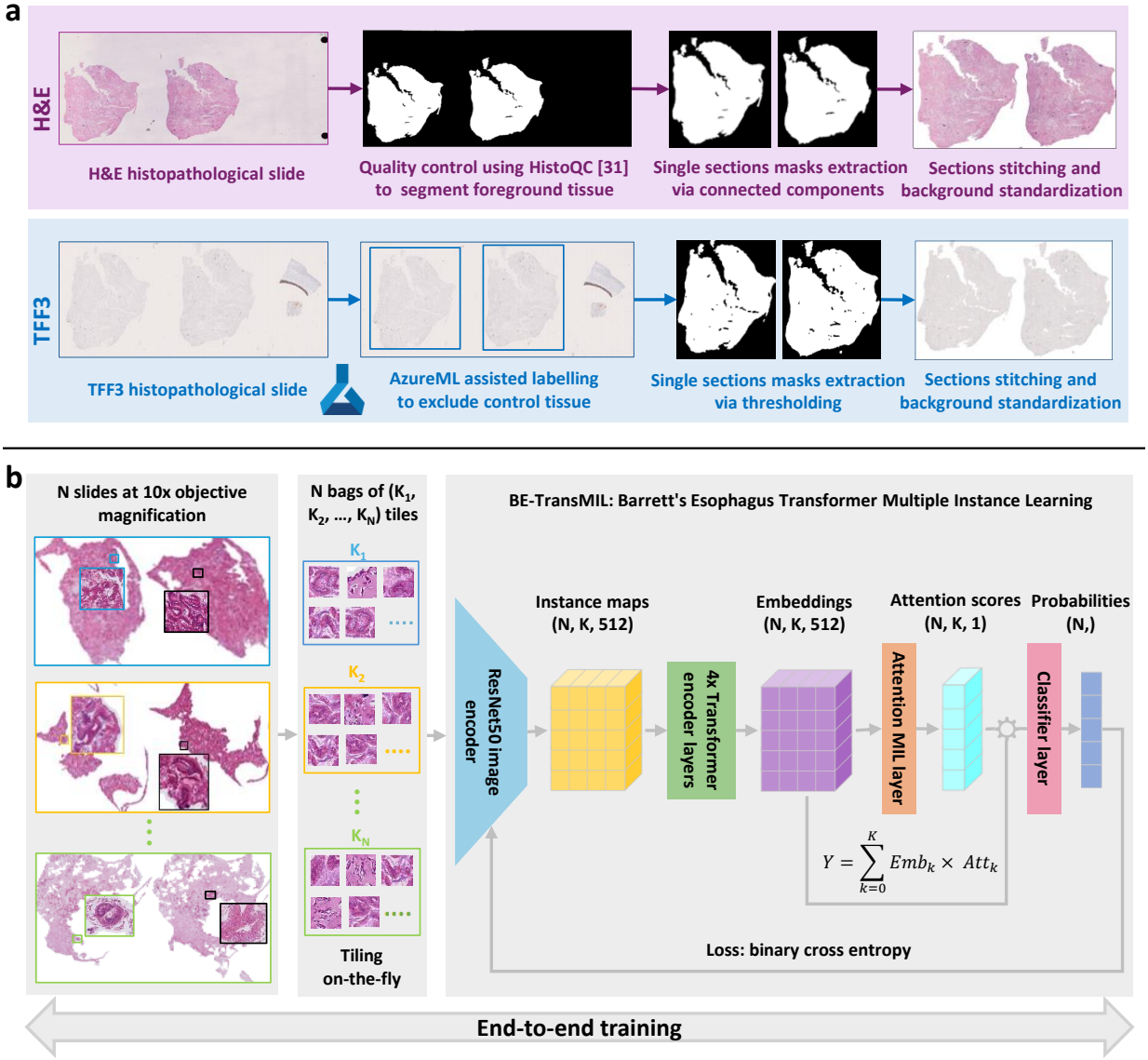
We benchmarked four different types of image encoders for each of the two models, namely, SwinT[22], DenseNet121[23], ResNet18[24], and ResNet50[24] for feature extraction (Methods). Cross-validation metrics on the discovery development dataset reveal that the model with the ResNet50 image encoder achieves the highest performance for detecting BE from H&E and TFF3 slides (Extended Data, Table 1 and Table 2). For the H&E model, ResNet50 achieves the highest area under receiver operating characteristic curve (AUROC) (mean $\pm$ standard deviation: 0.931 ± 0.021) and area under precision–recall curve (AUPR) (0.919 ± 0.031) among the four encoders. For the TFF3 model, ResNet50 achieves the highest or consistent AUROC (0.967 ± 0.003) with a consistent AUPR (0.951 ± 0.014), though DenseNet121 achieves the highest AUPR.

Summarizing the performance of the respective best models on the discovery test set at the selected operating points (see Methods for details), the H&E BE-TransMIL model achieves specificity: 0.922 (95% CI: 0.874–0.963), sensitivity: 0.727 (95% CI: 0.647–0.833) and AUROC: 0.914 (0.869–0.951)(Fig. 2a). The TFF3 BE-TransMIL model reaches specificity: 0.965 (95% CI: 0.919–0.986), sensitivity : 0.791 (95% CI: 0.707–0.882) and AUROC: 0.939 (95% CI: 0.901–0.971) (Fig. 3a). Note that the TFF3 BE-TransMIL model serves as the upper bound for the model trained with the adjacent H&E-stained slides.

Both H&E and TFF3 models focus on regions with goblet cells – the hallmark of IM for detecting BE. A key feature of BE-TransMIL models is a learnable attention mechanism, whereby the slide prediction is computed from the weighted feature representations of individual image tiles. Consequently, we can analyze distribution of attention weights and assess tiles contributing most (or least) to the model's prediction. To ensure interpretability of the trained deep learning models' outcomes, we perform detailed qualitative and quantitative analysis by investigating regions where the H&E and TFF3 models relatively focus on, including visual inspection of slide attention heatmaps and tile saliency maps, and TFF3 stain–attention correspondence analysis.

We analyze the slide attention heatmaps of the model and visually inspect high- and low-attention tiles, each of size 224×224 pixels (see Methods for details). For the H&E BE-TransMIL model, we analyze a true BE-positive slide (Fig. 2c) and observe that tiles that receive high attention (red regions in attention heatmap) contain goblet cells, indicative of BE. Moreover, these regions exhibit brown (positive) TFF3 staining in the adjacent TFF3 slide that further validates the presence of BE. Similarly, we confirm that the tiles with very low attention (blue regions in attention heatmap) do not contain goblet cells, and depict no brown staining in the adjacent TFF3 slide. For a true BE-negative slide (Fig. 2c), the attention heatmap shows uniform attention values without any high-attention regions, reflecting the absence of goblet cells. The adjacent TFF3 slide also does not indicate any positive brown staining. We repeat the analysis on the corresponding TFF3 slide with the attention heatmaps of the TFF3 BE-TransMIL model for the same BE-positive and negative slides (Fig. 3c and d). Again, we observe that the high-attention tiles indicate goblet cells clearly visible with dark brown TFF3 stain for BE-positive slide, and nearly uniform low-attention tiles without any brown TFF3 stain in the BE-negative slide.

In order to understand the relative importance given by trained model encoders at a more fine-grained level in the tiles of a slide, we generated saliency maps using gradient-weighted class activation mapping (Grad-CAM)[25] for individual tiles (Methods,



c

Dataset	Study	Patients (Slide pairs)	BE-positive	BE-negative
Discovery development set (train, validation)	DELTA	912	348 (38.2%)	564 (61.8%)
Discovery test set	DELTA	229	87 (38.0%)	142 (62.0%)
External test set (holdout)	BEST2	725	329 (45.4%)	396 (54.6%)
Total		**1,866**	**764 (40.9%)**	**1,102 (59.1%)**

Fig. 1 Method overview for automatic detection of Barrett's esophagus (BE) from H&E and TFF3 slides, including dataset preprocessing, model training, and data used in the study. a, H&E and TFF3 stained histopathological slides are scanned from adjacent sections and preprocessed using distinct pipelines for H&E and TFF3, as shown in purple and blue boxes, respectively. **b,** Preprocessed slides are split on-the-fly into non-overlapping tiles and used to train weakly supervised BE-TransMIL models end-to-end from H&E and TFF3 slides separately. A similar training procedure is performed for both stains. **c,** A summary of the datasets utilized in the study, with percentages of BE-positive and negative cases.

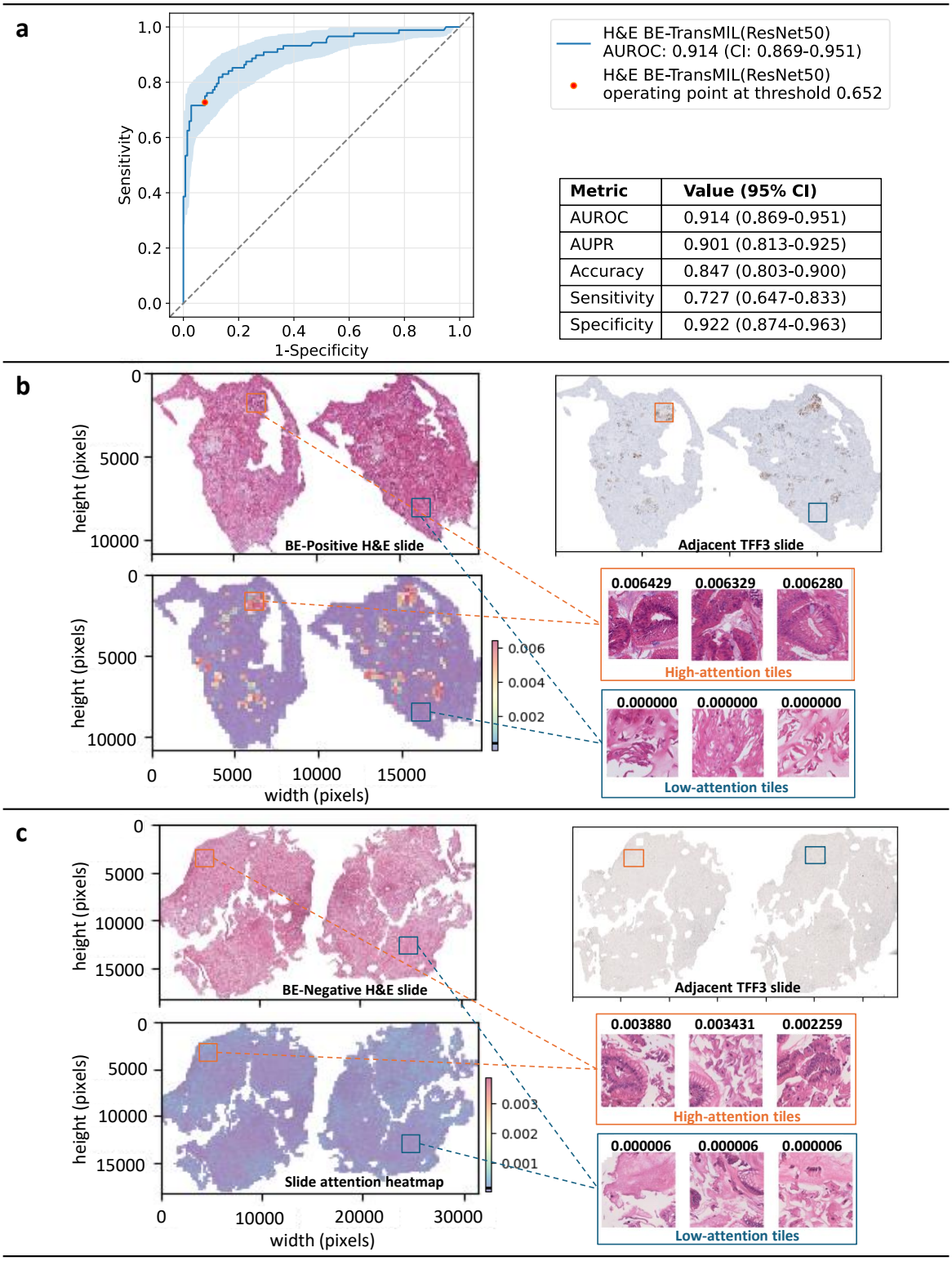
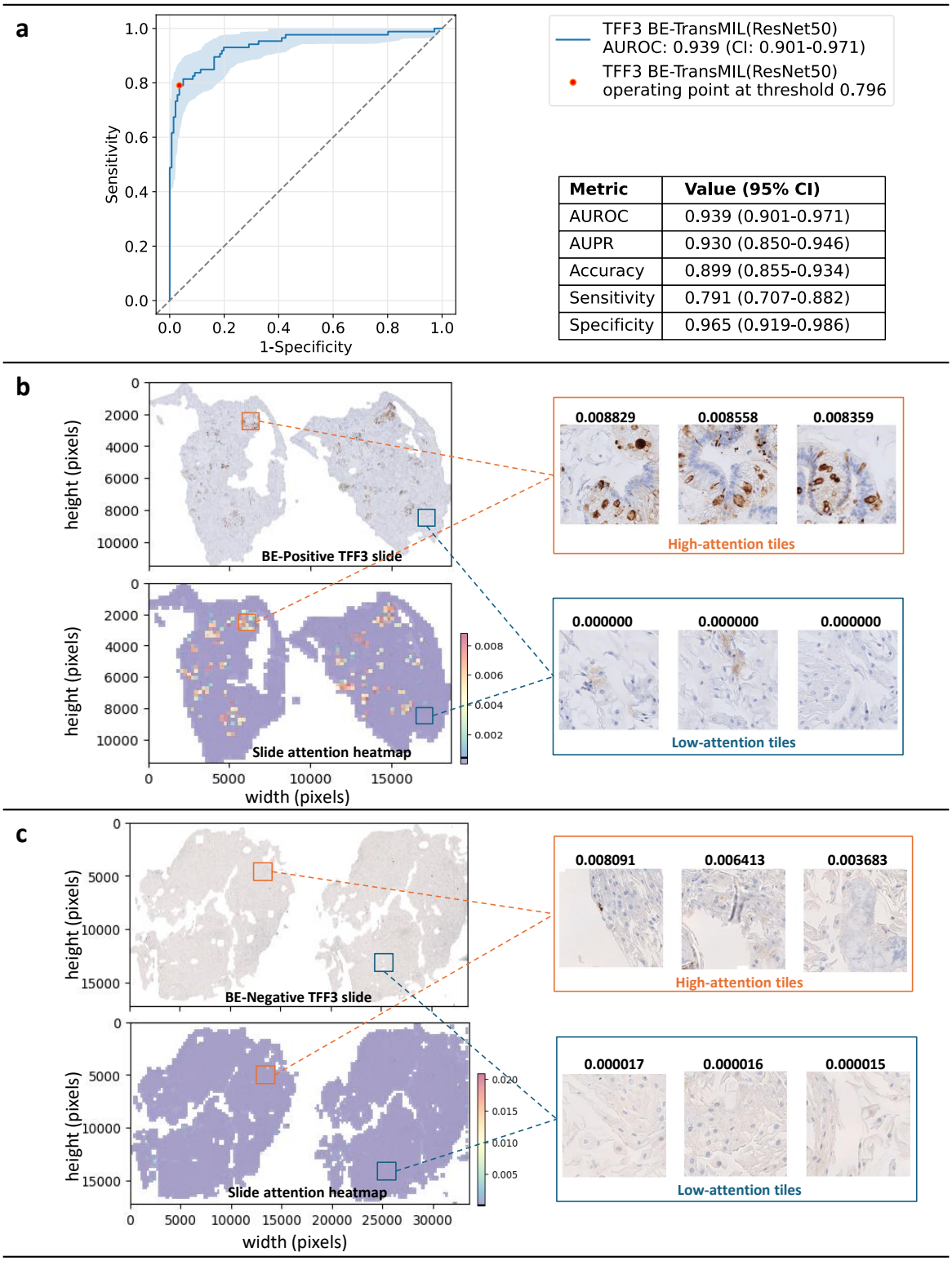


Fig. 2 Quantitative and qualitative analysis of the H&E BE-TransMIL (ResNet50) model on the discovery test dataset. a, ROC curve and metric values (95% confidence intervals (CI)) at the selected operating point. **b,** Example of a true positive H&E slide. Attention heatmap is heterogeneous, showing regions of high and low attentions that correspond to TFF3 staining in the adjacent TFF3 slide. Goblet cells visible in the tiles with high attention values; tiles with low attention values do not show any goblet cells. **c,** Example of a true negative H&E slide. Attention heatmap shows uniform attention without any high-attention regions; high- and low-attention tiles are without any goblet cells. Color bars along heatmaps show the range of attention values with marked mean value.



Metric	Value (95% CI)
AUROC	0.939 (0.901-0.971)
AUPR	0.930 (0.850-0.946)
Accuracy	0.899 (0.855-0.934)
Sensitivity	0.791 (0.707-0.882)
Specificity	0.965 (0.919-0.986)

Fig. 3 Quantitative and qualitative analysis of the TFF3 BE-TransMIL (ResNet50) model on the discovery test dataset. a, ROC curve and metric values (95% CI) at the selected operating point. **b,** Example of a true positive TFF3 slide. Brown TFF3-stained goblet cells are visible in the tiles with high attention values, whereas tiles with low attention values do not show any brown TFF3 stain or goblet cells. **c,** Example of a true negative TFF3 slide. Attention heatmap shows uniform attention without any markedly high-attention regions; tiles with highest and lowest attention values do not have any positively stained TFF3 regions, indicating absence of goblet cells.

Extended Data, Fig. 2). We observe that, for both H&E and TFF3 model encoders, there is visual agreement between the locations of goblet cells indicative of BE, and saliency map activations. Specifically, for the H&E model, we observe that higher importance is given by the models to the mucin-containing goblet cells (translucent, bluish appearance). The TFF3 model attributes more importance to regions with goblet cells showing positive TFF3 staining.

Going beyond qualitative visual inspection on few slides, we want to quantitatively establish across the discovery test dataset whether high-attention H&E tiles correspond to brown staining on the adjacent TFF3 slides, indicating the presence of goblet cells. A stain–attention correspondence analysis was performed, involving the spatial registration of TFF3 slide to the corresponding H&E slide (Extended Data, Fig. 3) extracting the 3,3'-diaminobenzidine (DAB) stain ratio in the TFF3 tiles, and computing the correspondence of the H&E attention heatmaps with the TFF3 stain ratios (Fig. 4a), see Methods for details. We use the same BE-positive and negative slides as in Fig. 2c–d and 3c–d to illustrate the slide-level stain–attention correspondence.

For the example slides of the H&E BE-TransMIL model (Fig. 4b), we find that in the attention plot of the BE-positive slide, high-attention tiles overlap with high stain ratio values. The higher attention values concentrate on the tiles with highest TFF3 expression, showing a high attention–stain agreement. The attentions are not uniform and become lower as stain ratio decreases. A Pearson's correlation coefficient $r > 0.5$ between stain and attentions substantiates a high correspondence. For the BE-negative slide, attention is largely uniform as the model detects no goblet cells, and attention values have a higher normalized entropy than BE-positive slide. All 50 true BE-positive H&E slides present a positive correlation between the stain ratio and attention weights, with mean $\pm$ standard deviation of 0.35 ± 0.23 and range 0.01–0.84. Additionally, the normalized attention entropies across all slides are higher for BE-negative slides, indicating more uniform and diffused attentions compared to BE-positive slides.

For the the example slides of the TFF3 BE-TransMIL model (Fig. 4c), we observe that high-attention tiles overlap with high stain ratio values and vice-versa with $r > 0.5$ for the BE-positive slide, similar to the observation for the H&E model. For the BE-negative slide, attention values are diffuse, with a higher normalized entropy than the BE-positive slide. The latter observation is supported by the box and strip plots of entropies over all slides. Again, all stain–attention correlation coefficients r for the 58 true-positive TFF3 slides are positive, with mean $\pm$ std. dev. of 0.28 ± 0.19 and range 0.02–0.67.

Failure-modes analysis reveals complex cases for BE detection from histopathological slide. To specifically understand where the models were unable to correctly detect BE, we evaluated the incorrectly predicted cases in the discovery test set (see Methods for details). We observe false prediction rates for the H&E model (27.3% false negative (FN), 7.8% false positive (FP)) and TFF3 model (20.9% FN, 3.5% FP) on the discovery test set (Fig. 2a, Fig. 3a). Our failure-modes analysis (see Extended Data, Table 4 for error quantities) reveals complex cases for which visual BE detection from histopathological slides is challenging.

We first focus on false negatives (FNs) due to their high incidence rate and clinical relevance. Out of all the FNs (Extended Data, Table 4), the majority (56% of the total FNs) are shared across both the H&E and TFF3 BE-TransMIL models. An expert pathologist reviewed these shared cases and reported that, in majority of the shared cases (48%), the goblet cell (i.e. hallmarks of IM) groups were not well-represented on the H&E slide, in 28%, there were small or few groups of goblet cells, and in 16% of these cases, the H&E slide was not well-preserved and mucin was not clearly visible. These observations suggested that such features may not be representative in the training dataset and were difficult to identify as positives. In addition, we analyzed the unique FNs of each model. H&E-only FNs (38% of the total FNs) may indicate 'pseudo-adjacent' tissue sections, wherein a TFF3 slide contains goblet cells but the paired H&E slide, obtained farther along the tissue block, may not. TFF3-only FNs (6% of the total FNs) contain background noise with low contrast between foreground tissue and background, and higher levels of stain blush (a faint or lower intensity staining, not necessarily associated with the location of goblet cells; also noticed in a few true negative cases), leading to nearly uniform attentions and a negative prediction. In summary, the majority of FNs were observed as non-trivial to visually detect BE by the pathologist, with none or sparse goblet cells in the H&E slide and unclear or equivocal staining in the adjacent TFF3 slide. These cases were labeled BE-positive by default to avoid missing any suspicious cases by the Cytosponge-TFF3 test; this observation informs our deployment strategy to design workflows to maximize specificity.

In the cases that were FP calls (Extended Data, Table 4), we observe that the shared FPs for the H&E and TFF3 BE-TransMIL models are much fewer (23% of the total FPs) than the shared FNs. Qualitative analysis of the shared and H&E-only FP slides reveals that very few (1–2) tiles show high attentions suggesting goblet cells. This appears to be related to the presence of pseudogoblet cells, which have a goblet cell-like appearance in the H&E slide and non-specific staining on the adjacent TFF3 slide[26]. Other artifacts were also mistaken by the models as goblet cells with high attentions due to darker intensities resembling positive TFF3 staining (e.g. bubbles' borders), and other non-specific staining.

The trained model for BE detection from H&E slides generalizes well to an external dataset. We used an external dataset of 725 cases from the multi-center BEST2 case–control clinical trial study[7] for external validation (Fig. 1c, Methods). We observe that the H&E BE-TransMIL model achieves 0.873 (95% CI: 0.844–0.899) AUROC (Fig. 5a). The model achieves a comparable performance on the external test dataset, i.e. similar to the discovery test dataset (Fig. 2a) with 0.914 (95% CI: 0.870–0.952) AUROC. In the visual inspection of the attention heatmaps (Fig. 5b–c), we find high-attention tiles containing goblet cells for the BE-positive slide and uniform attention for

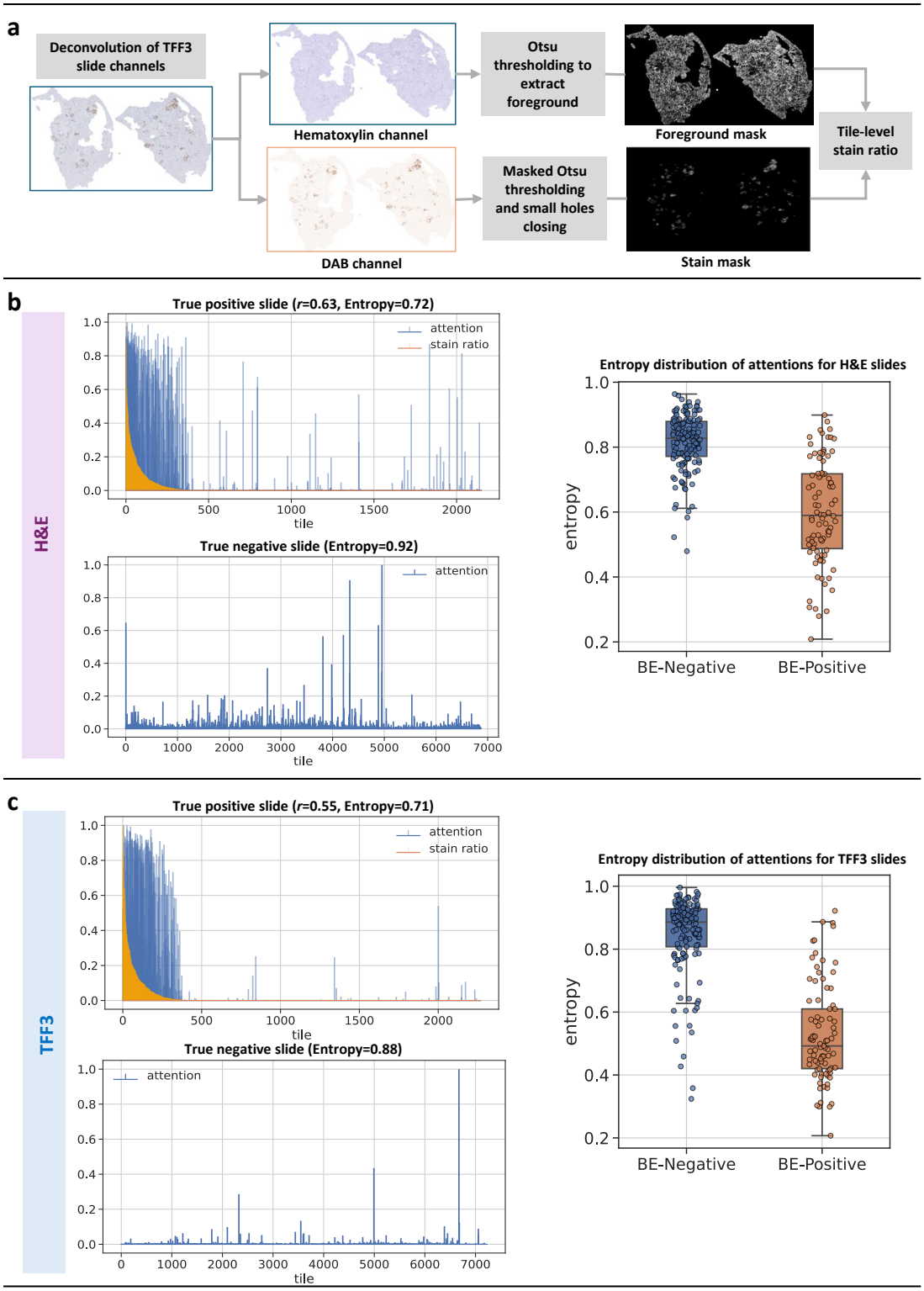


Fig. 4 Model attentions of BE-TransMIL models show high correspondence with TFF3 stain in true positive slides, and are uniform and diffused for true negative slides. a, Overview of TFF3 stain ratio computation. Left: slide-level attention plots for true positive slide (overlap with stain ratio, Pearson correlation r, and entropy) and true negative slide (with entropy). Right: Overlay of box and strip plots of normalized entropy of attention distributions of all slides in the test dataset using **b**, H&E BE-TransMIL (ResNet50) model and **c**, TFF3 BE-TransMIL (ResNet50) model.

BE-negative slide, as expected. Note that the H&E stain intensity of the external example slide is different from the H&E stain intensity of the example slide in the discovery dataset, owing to different staining protocols of the two datasets (Methods).

The average predictive performance of a pathologist on Cytosponge samples from the BEST2 study dataset with respect to endoscopy labels is discussed in[9], with a specificity of 0.927 and sensitivity of 0.817. We observe that on the external dataset from the same study, our weakly supervised deep learning model achieves a specificity of 0.881 and sensitivity of 0.720 at the selected operating point using only pathologists' diagnostic labels. Although, the model performance is not directly comparable to the pathologist's predictive performance, this observation on the BEST2 study informs the design of semi-automated ML-assisted workflows, as discussed in the next section.

Proposed ML-assisted workflows can substantially decrease manual review workloads. Integration of ML-assisted workflows in clinical practice could reduce pathologist workloads to assess histopathology slides by markedly lowering the cases requiring pathologist's manual review, and can improve cost-effectiveness by reducing the need for specialized stains. Comparing several workflows based on the above two criteria (see Methods), we propose two semi-automated ML-assisted workflows. In the first approach, we use either the H&E or TFF3 BE-TransMIL positive predictions to be followed by pathologist review. The second approach prioritizes the H&E model alone, such that TFF3 staining could be limited to cases with a positive finding in H&E. These are illustrated in Fig. 6a–b (see Extended Data, Table 5, Fig. 4 for detailed results). The workflows are designed to optimize specificity, to enable pathologists to review fewer negative cases and focus on the high risk cases.

The first workflow, "Pathologist reviews any positives" (Fig. 6a), requires both H&E and TFF3 models to analyze a sample, deferring to a pathologist if either model predicts positively. In other words, if both models agree that there are no signs of BE on either stain, the sample is assumed to be likely BE-negative and is not manually reviewed. This configuration can achieve 1.00 sensitivity and 1.00 specificity on the discovery dataset with respect to the pathologists' diagnosis alone, suggesting that the two models and pathologist are complementary (Extended Data, Fig. 4). In this scenario, only 48% (41–55%) of samples would need manual review, implying a 2.1× (1.8–2.5×) reduction in pathologist's workload. Among the samples that would reach pathologist review (i.e. likely positives), 14–20% are expected to be BE-positive, compared to the baseline prevalence of 5–12% in the fully manual clinical setting[5]. However, this workflow still relies on both conventional H&E and immunohistochemical TFF3 stains, similar to the current manual screening pathway.

The second workflow, "Pathologist reviews H&E model positives", stipulates that a sample is only manually reviewed if the H&E alone is positive (Fig. 6b), which would require TFF3 staining only for the 37% (31–45%) of samples that get reviewed by a pathologist. However, this workflow would result in a lower sensitivity of 0.91 (0.84–0.96) (Extended Data, Fig. 4) compared to the first workflow which uses both H&E and TFF3 models. In this scenario, pathologist workload could be reduced by 2.7× (2.2–3.4×). Among the samples reaching pathologist review, the observed prevalence of BE-positive would be 16–24%.

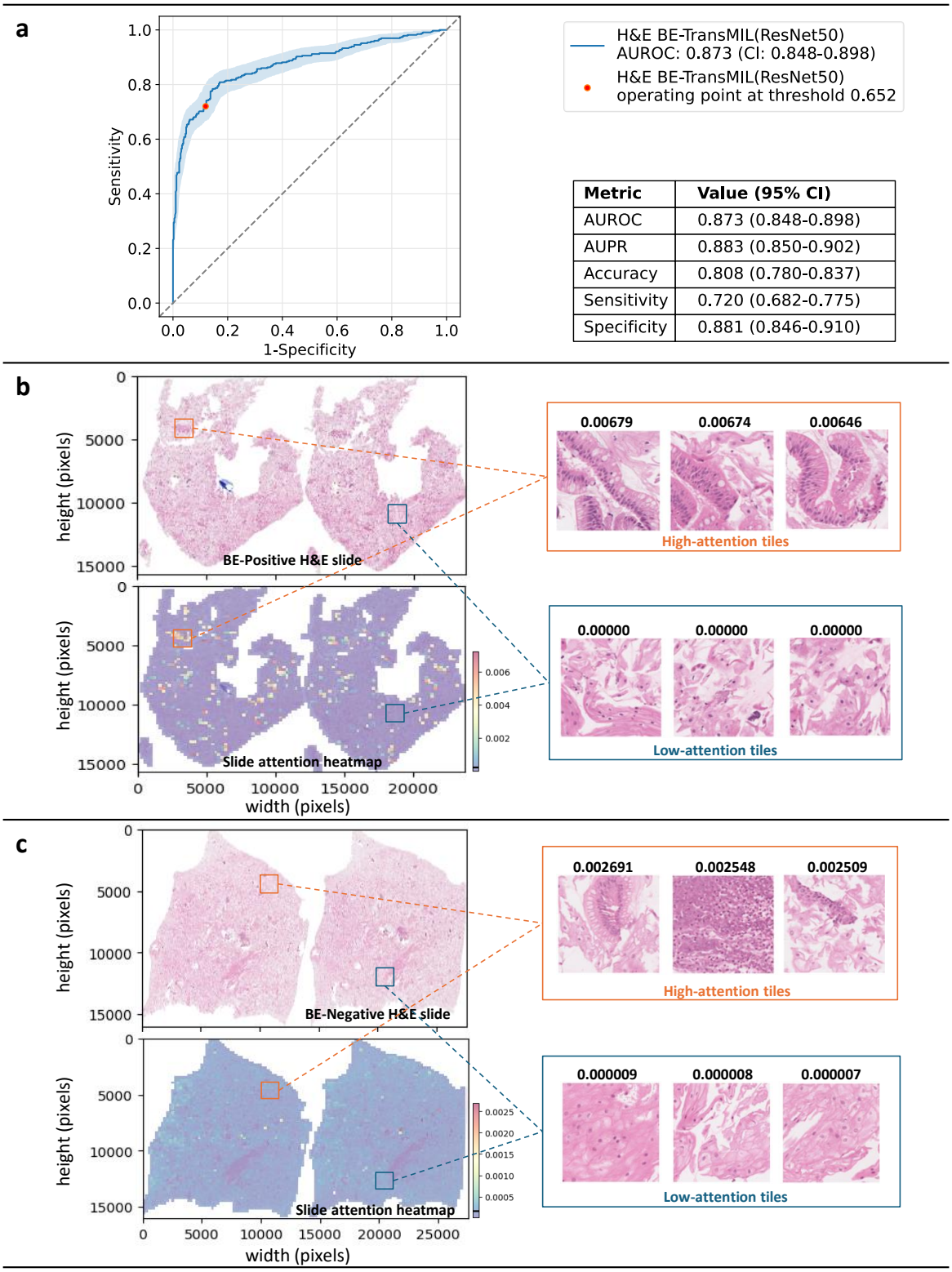
Discussion

Detection of BE from histopathology slides presently relies on a pathologist manually inspecting both the routine H&E and specialized TFF3 stained slides for each patient. The current resource-intensive clinical workflow represents a big hurdle for large-scale screening for BE using the Cytosponge-TFF3 test. Our work has demonstrated that weakly supervised deep learning models can detect BE using only the H&E slides. Most importantly, these models were able to identify the salient features used by pathologists, namely goblet cells. This approach shows that accurate models can be trained directly from the reported histopathology at the slide-level. This is an important difference from the previous models[9], as large-scale localized annotations in these slides require significant time and effort from expert pathologists.

Screening for BE currently requires pathologists to spend the bulk of their time reviewing cases that are negative. Exploring the potential of integrating deep learning models into clinical practice, we imagine that the first steps in a disease screening setting would be to optimize pathologists' manual workload to enable them prioritize high risk cases. We suggest an alternative using semi-automated workflows, including one that uses both H&E and TFF3 models. We estimate that the number of cases pathologists would need to review could be cut in half (48% current cases for manual review) without any loss of accuracy. This implies a 2.1× increase in screening coverage with the same number of pathologists by screening out the negatives and enabling pathologists to focus on positives.

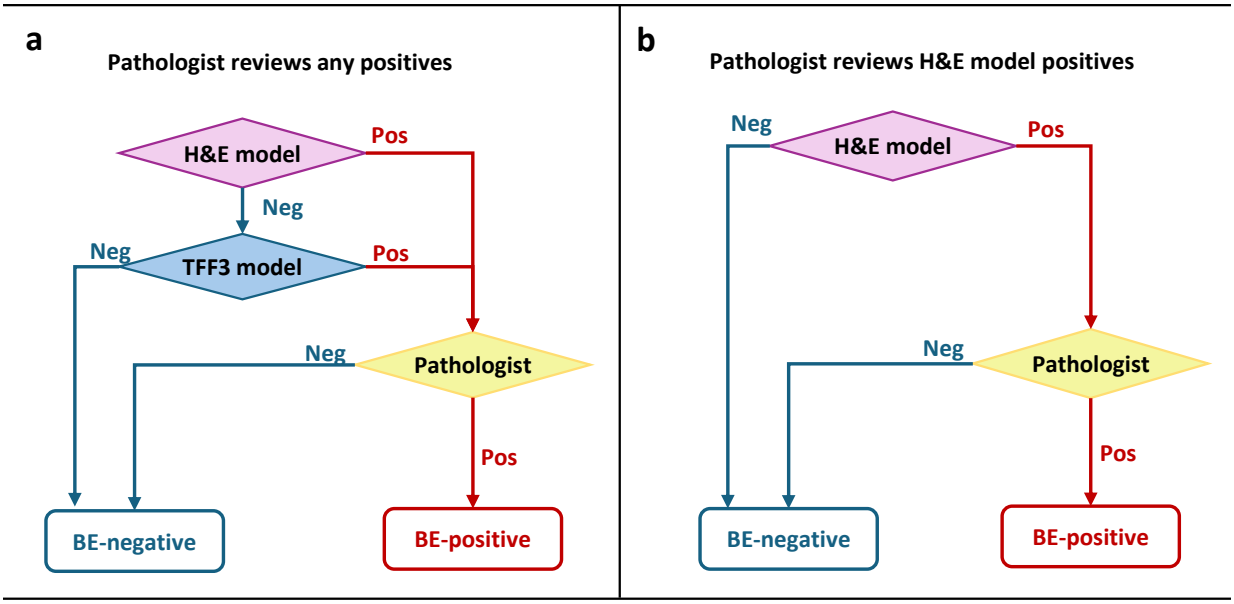
We observe that our models generalize well to an external dataset, demonstrating a comparable predictive performance as the discovery test set. These results are encouraging, considering the differences between the discovery and external datasets, including the slide preparation and staining protocols, patient populations, and reporting pathologists. The external evaluation demonstrates that our trained models could become a stepping stone to (semi-)automating the Cytosponge-TFF3 screening test, potentially allowing to scale it to larger populations.

One of the limitations of this study is that, while the discovery dataset is derived from various sites in the UK across different patient populations, the samples were sectioned and stained at a single site. This limitation was highlighted in the external validation dataset which showed greater variation in the stain, artifacts such as pen marks, and tissue preservation differences. We could not quantify whether or how much these differences relate to the failure cases. Future studies can account for this by mixing these now well-characterized datasets in training and test, as well as by continuing to include new data over time as sample processing protocols change. An additional source of error we observed during model development was manual errors in extraction of slide labels from pathologists' reports. This process could be improved using large language models (e.g.



Metric	Value (95% CI)
AUROC	0.873 (0.848-0.898)
AUPR	0.883 (0.850-0.902)
Accuracy	0.808 (0.780-0.837)
Sensitivity	0.720 (0.682-0.775)
Specificity	0.881 (0.846-0.910)

Fig. 5 Quantitative and qualitative analysis of the H&E BE-TransMIL (ResNet50) model on the external dataset. a, ROC curve and metric values (95% CI) at the selected operating point. **b,** Example of a true positive H&E slide. Similar to the H&E example slide of internal dataset (Fig. 2), the slide attention heatmap shows regions of high and low attentions, where goblet cells can be seen in the tiles with high attention values; tiles with low attention values do not show any goblet cells. **c,** Example of a true negative H&E slide. Attention heatmap shows nearly uniform attention; high- and low-attention tiles do not have any goblet cells.



c

Proposed ML-assisted workflow	Pathologist review	TFF3 staining	Observed prevalence	Sensitivity	Specificity
Pathologist reviews any positives	48% (41-55%)	100%	17% (14-20%)	1.00 (1.00-1.00)	1.00 (1.00-1.00)
Pathologist reviews H&E model positives	37% (31-45%)	37% (31-45%)	19% (16-24%)	0.91 (0.84-0.96)	1.00 (1.00-1.00)

Fig. 6 Proposed ML-assisted workflows. a, Workflow "Pathologist reviews any positives" **b,** Workflow "Pathologist reviews H&E model positives" **c,** Quantitative comparison of the proposed workflows in terms of the requirements for pathologist review as fraction of the current reviewed cases, TFF3 staining as fraction of the current cases, observed prevalence of BE, sensitivity and specificity (with 95% CIs in parentheses). "Pos" and "Neg" refer to BE-positive and negative, respectively.

GPT-4) to extract diagnostic information from unstructured text, ranging from the slide labels to clinical variables we have not accounted for in our models.

Although we offer two options for integrating deep learning models into the current clinical workflow, this continues to be an area of active research. Moreover, user interfaces have recently been introduced in ML-assisted histopathology workflows[27–29], where open questions include how specific visualizations can best assist pathologists' practice to accelerate their visual assessment of slides or aid their diagnostic decision-making. For instance, an overlay of model-generated attention heatmap on the whole-slide image with the ability to adjust opacity could help pathologists focus on the highlighted regions, leading to expert-time savings. Future work is required to quantify ML-assisted pathologist review times and compare with time to confirm or reject model results.

In summary, a weakly supervised deep learning approach using only routine H&E slides enables the training of a pathologically accurate model that offers the potential to reduce pathologist workloads through semi-automated workflows, allowing them to prioritize high risk cases, thereby facilitating large-scale screening of BE. Furthermore, the approach requires no extra efforts to create localized expert annotations. This also means that future models could be trained continually in real-time as new diagnostic data is generated, plausibly leading to further improvements in the performance of the trained models. Moreover, reliance on only diagnostic labels from pathologist assessments or reports facilitates the adoption of the approach to other screening applications in clinical histopathology.

Methods

Discovery and external evaluation datasets. The discovery dataset consists of 1,141 cases with both hematoxylin and eosin (H&E)- and trefoil factor 3 (TFF3)-stained whole-slide images from patients in the DELTA implementation study (integrated diagnostic solution for early detection of esophageal cancer study; funded by Innovate UK; ISRCTN91655550). Ethics approval was obtained from the East of England—Cambridge Central Research Ethics Committee (DELTA 20/EE/0141) and written informed consent was obtained from each patient.

The external test dataset consists of 725 cases with H&E slides from the BEST2 (ISRCTN12730505) clinical trials[11,30]. Ethics approval was obtained from the East of England—Cambridge Central Research Ethics Committee (BEST2 10/H0308/71) and the trials are registered in the UK Clinical Research Network Study Portfolio (9461). Written informed consent was obtained from each patient.

After retrieval, the Cytosponge was placed in SurePath Preservative Fluid (TriPath Imaging, Burlington, NC, USA) and kept at $4\,^{\circ}\mathrm{C}$. The sample was then processed to a formalin fixed block[7]. TFF3 staining was performed on slides 2 and 15 on serial sections according to established protocol (proprietary monoclonal antibody) using standard protocols on a BOND-MAX autostainer (Leica Biosystems, Newcastle upon Tyne, UK) as previously described[13]. Expert histopathologists scored the TFF3 slide in a binary fashion, where a single TFF3 positive goblet cell is sufficient to classify the slide as positive.

In the BEST2 trial all patients underwent an endoscopy within an hour of the Cytosponge procedure[7]. DELTA was a prospective trial with both known Barrett's esophagus (BE) patients and reflux screening patients. For this analysis, no follow-up endoscopic information was available. All class labels were based on the expert pathologists' reading of the Cytosponge slides.

Slides in both discovery and external datasets were scanned in digital pathology image formats (NDPI and SVS, respectively), with $5\times$, $10\times$, $20\times$, and $40\times$ as the available magnifications, with a resolution of $0.23\,\mu\mathrm{m/pixel}$ at the highest magnification. Quality control was performed to exclude the slides whose Cytosponge sample contained insufficient gastric tissue[9]. Additionally, visual quality control was performed to ensure correct H&E/TFF3 categorization of all images.

Data preprocessing. Preprocessing was performed to mitigate undesirable artifacts (e.g, bubbles, shadows, pen marks), standardize background effects, and to remove control tissue in TFF3 slides. Foreground masks for H&E slides were extracted via HistoQC[31] with configuration 'v2.1' (https://github.com/choosehappy/HistoQC), and all background pixels were set to a fixed plain-white value (255, 255, 255) in the RGB color model. Each H&E slide contains two tissue sections side-by-side, whose separate bounding boxes were determined based on morphological processing of the foreground masks. For the immunostained TFF3 slides, the low staining contrast led to unsatisfactory automatic foreground segmentation for some slides. Therefore, tissue section bounding boxes for TFF3 slides were obtained semi-automatically, using the Microsoft Azure Machine Learning (https://azure.microsoft.com/en-us/services/machine-learning/) data labeling tool. Foreground masks were then obtained for each section using the 80th percentile of the estimated hematoxylin concentration via stain deconvolution[32], as a threshold to select cell nuclei, followed by binary closing to fill in the gaps between the cells and finally binary opening to remove false positive pixels in the background. Both morphological operations were applied using a disk of 8 pixels radius at $1.25\times$ objective magnification (equivalent to $60\,\mu\mathrm{m}$). Lastly, the H&E and TFF3 tissue sections were cropped and stored at a single resolution in TIFF format ($10\times$ objective magnification at a fixed scale of approximately $0.92\,\mu\mathrm{m/pixel}$), resulting in a tenfold reduction in dataset size and improved training throughput. The specific preprocessing pipelines for H&E and TFF3 slides are demonstrated in Fig. 1a.

We selected a $10\times$ objective magnification for model training and inference, as it offers an adequate balance of contextual tissue architecture and cellular morphology, specifically for goblet cells, in the given field-of-view for a tile of 224×224 pixels (Extended Data, Fig.1). We also performed a sensitivity study of the H&E BE-TransMIL model using different objective magnifications at $5\times$, $10\times$, and $20\times$ (Extended Data, Table 3). Error bars were estimated via replication across random initializations of BE-TransMIL model parameters. We report the performance on a 10% random data split from the discovery development dataset, the rest 90% used for training. We found

that 10× achieves superior overall predictive performance compared to other magnifications, corroborating our visual assessment of tiles containing goblet cells at different magnifications (Extended Data, Fig.1). Additionally, we observed that 10× magnification offers a good trade-off between slide coverage during training phase and predictive performance, due to GPU memory limitations.

Model description. Due to high number of pixels in a slide (gigapixel sizes), it is not possible to process the entire slide at once with current hardware. The most common approach is to split the slide into tiles of equal size, such that batches of tiles can be easily handled by computer vision encoders. In the existing supervised learning approaches, each tile is given a label based on expert annotations of local regions on the slide ("dense annotations"); at prediction time, the results on individual tiles are aggregated as a proxy for the slide label[9,33]. Pathologists spend significant time and effort labeling specific cellular structures on a slide that are then used to train a model to classify new slides into one of the slide labels.

However, in a weakly supervised setting, instead of having access to a label for each tile, only slide-level labels are available. Classifying a set of tiles using a single binary slide label is a form of multiple instance learning (MIL), where we call the set of K instances (tiles) in a slide as a "bag", and assume that K could vary for each bag, that is, not all slides will have the same amount of tiles. We also assume that each bag (slide) has a binary label; in our case BE-positive or negative, and at least one instance (tile) should be positive for the entire bag (slide) to be positive[19,21]. In contrast to the supervised setting, the ground-truth label for each slide was obtained from pathologists' diagnostic labels for the cases, without the need to curate a training dataset with localized manual annotations.

Preprocessed slides were used to train weakly supervised models on H&E and TFF3 slides separately. For training and inference, the preprocessed slide were split on-the-fly into non-overlapping tiles of 224×224 pixels ($\approx 200\,\mu\text{m} \times 200\,\mu\text{m}$) with or without random offset (for training and inference, respectively) and background tiles were excluded, using the open-source MONAI library[34]. Tiling on-the-fly, as opposed to offline pre-tiling, offers greater flexibility in generating a wider variety of tiles that prevents the deep learning model from overfitting to the training set. At evaluation time, all the foreground tissue tiles in the slide were used to compute the model output (whole-slide inference). However, during training, only a subset of tiles (bag of size K) was used due to the limited GPU memory size. In order to ensure that relevant tissue regions were included in the bag during training, we applied a minimum intensity filter; this heuristic is based on the fact that dense cellular regions have an inherently darker appearance. Therefore, if K does not cover the entire slide, we ensured that the most relevant regions are selected. Additionally, we set an intensity threshold at 90% to exclude background regions previously set to plain-white in the preprocessing step for the H&E slides in the discovery dataset. Finally, we applied random geometric augmentations including 90° rotations and horizontal and vertical flipping to reduce overfitting effects during training.

The network architecture, depicted in Fig. 1b, is inspired by the model variant Transformer-MIL proposed in[19]. It is built upon attention-based MIL[21] paradigm, wherein a trainable module attributes an "attention" weight to each instance (tile) in the bag (slide). This has the benefit of being highly interpretable, as it facilitates inspection of whether the tiles with highest attention values are abnormal tissue sections that contain goblet cells in this context. The overall model architecture is composed of four main components. First, a feature extractor that encodes each image tile into lower-dimensional feature maps; it consists of one of the convolutional neural network (CNN) or vision transformer encoders. Second, a dependency module that captures spatial dependencies between individual tile maps in a bag into compact vector representations ("tile embeddings"); it is composed of four consecutive Transformer[35] encoder layers. Next, an attention MIL pooling module[21] using a multi-layer perceptron (MLP) with a single hidden layer of dimension 2048. Finally, a fully-connected classifier layer that receives a linear combination of all tile embeddings weighted by attention values to compute the final probability to predict a label for the slide.

We benchmarked different deep learning image encoder architectures including the 'tiny' version of Shifted Window Transformer (Swin-T)[22], DenseNet121[23], and two variants of ResNet[24], namely, ResNet18 and ResNet50. These encoders have achieved promising results on a variety of computer vision tasks, such as image classification, object detection, and segmentation. All encoders were initialized using weights from models pretrained on natural images. We used the highest possible bag size (K) with each encoder, constrained by the available GPU memory. To optimize the supported bag size K for each encoder, we implemented activation checkpointing[36], where we reduced the amount of memory required to store intermediate activations used to compute gradients during the backward pass, freeing up GPU memory for larger bag sizes processing. Comparative analysis of the encoders using four-fold cross-validation (Extended Data, Table 1, Table 2) depicts that ResNet50 (K=1200) encoder outperforms the other three encoders for detection of BE from H&E and TFF3 slides, which is then selected for further result analysis. Intuitively, ResNet50, owing to a deeper network architecture with more trainable parameters than ResNet18 (K=2300), encodes the image tiles more favorably and leads to superior performance even with a lower bag size. The other two encoders, Swin-T (K=1100) and DenseNet121 (K=700), are more computationally expensive to train than the ResNets (leading to lower bag sizes), and exhibit lower performance compared to ResNets.

The complete ("end-to-end") networks were finetuned using binary cross entropy loss using solely slide labels. Hyperparameter tuning was performed for high specificity, so that the models could confidently identify negative cases automatically (see Supplementary Material for tuned hyperparameters). For training the models, we used a batch size of 8 slides. Learning rate was fixed at 3×10^{-5} with a weight decay of 0.1, and models were trained for 50 epochs. Due to unbalanced datasets, class reweighting was applied using the Scikit-learn library[37].

At 10× objective magnification, H&E histopathological slides have a mean number of 3,779 tiles (range: 428 – 14,278 tiles), hence, it is not feasible to encode all tiles at once due to GPU memory constraints. Therefore, to perform whole-slide inference at evaluation time, we encoded the bag of tiles in chunks, concatenated the sub-feature maps into a large tensor, before feeding it to the transformer encoder that computes attention across the entire slide. Note that encoding in chunks is not feasible at training time due to parallel processing limitations that require exact number of forward passes to synchronize subprocesses, in addition to higher GPU memory requirements to store activations and gradients during the training phase. Training of all BE-TransMIL models was performed using compute nodes of 8 NVIDIA V100 GPUs in the Microsoft Azure cloud (https://azure.microsoft.com/). Inference was run on a single V100 GPU. 40 CPU cores were used to tile the whole-slide images (WSIs) on the fly.

Statistical methods. We split the discovery dataset into development and test as an 80:20 split (Fig. 1c). We performed four-fold cross-validation experiments on the discovery development set; this led to an effective train/validation/test split of 60:20:20 on the discovery dataset. Validation and test sets were randomly selected, stratified according to distributions of class labels and patient pathway (surveillance or screening).

To compare the performance of different weakly supervised models, we calculated area under receiver operating characteristic curve (AUROC) and area under precision–recall curve (AUPR), which are threshold-agnostic metrics, as well as accuracy, specificity, and sensitivity at 0.5 probability threshold. We report these metrics on the discovery validation set for each of the H&E and TFF3 models in Extended Data, Tables 1 and 2. We also performed replication experiments at different magnifications to observe the variation of metrics across random initializations of model parameters (Extended Data, Table 3).

After training the cross-validation models, we computed AUROC values for each fold on the validation dataset. For clinical relevance, the classification threshold was chosen at 0.85 sensitivity on the validation set for each model. The cross-validation fold with the highest AUROC was then used for inference and computing standard metrics (accuracy, AUROC, AUPR, specificity, and sensitivity) on the discovery and external test datasets at the selected probability threshold (Figs. 2b, 3b, and 5b). In addition, we plotted ROC curves with bootstrapping for confidence intervals (CI) (Figs. 2a, 3a, and 5a). CIs were defined as the 2.5^{th} and 97.5^{th} percentiles on distributions of 1000 samples (with replacement) of the test dataset size.

Qualitative analysis. For qualitative analysis, including visualization of results for interpretability of the BE-TransMIL models, we plotted attention heatmaps overlaid on the slides. For each tile, we color-coded the attention values based on the reversed spectral colormap (high to low values coded from red, yellow, green, to blue), and stitched the tile maps to get the slide attention heatmap. Also, we visually investigated the tiles with high and low attention values, for fine-grained inspection of the regions where the models gave highest or lowest importance while performing the prediction (Figs. 2c–d and 3c–d).

Gradient-weighted class activation mapping (Grad-CAM)[25] generates class localization maps by visualizing the gradients that flow into the last convolutional layer of the encoder of the weakly supervised model, which retains the class-specific spatial information from the input image before the fully- connected and pooling layers. The highlighted areas on these maps depict the specific locations within an image that are crucial for a model to identify a specific class. Grad-CAM requires no architectural modifications or retraining of the model, making it convenient to use. For the example BE-positive slide in the test set (Figs. 2 and 3), we generated Grad-CAM saliency maps for both H&E and TFF3 BE-TransMIL models. The target layer from the encoder architecture was the fourth ResNet block. We generated saliency maps of the 10 tiles of the example true-positive slide with highest attention (Extended Data, Fig. 2).

TFF3 expression quantification and stain–attention correspondence analysis. To quantify whether the high-attention tiles of the learned models correspond to tiles with high TFF3 expression, a detailed analysis was performed as follows. To obtain fine-grained TFF3 staining correspondence with the H&E tiles, the reference TFF3 tissue crops were spatially registered to the corresponding H&E crops[38] (Extended Data, Fig. 3). For registration, we first estimated the hematoxylin concentration from RGB pixel values via stain deconvolution[32], using the Macenko method[39] to estimate the stain matrix for H&E slides and employing the default hematoxylin–eosin–DAB (HED) stain matrix using the Scikit-image library[40] for TFF3 slides. The hematoxylin images were then registered with an affine transform (16×/4× downsampled) followed by a coarse cubic B-spline deformation (5×5 grid, 4×/2× downsampled), optimizing a mutual information criterion using SimpleITK[41]. The same fitted transform was then applied to the corresponding 3,3'-diaminobenzidine (DAB) image (DAB is the chromogen used in TFF3 staining). As registration is a computationally intensive process for gigapixel-sized whole-slide images, we registered the TFF3 slides at 5× objective magnification ($1.84\,\mu\text{m/pixel}$). The fitted transform parameters were then applied to register the corresponding slides at 10× objective magnification.

To analyze the correspondence of TFF3 expression and model attentions, we quantified the proportion of the DAB-stained pixels out of all tissue pixels in the TFF3 slides. We created a binary mask of the positive DAB stained regions based on the method described in[38]. Firstly, we separated the channels of the H&E into constituent hematoxylin and DAB stains, then we detected the foreground mask on the hematoxylin channel and the stain mask on the DAB channel using Otsu thresholding, and post-processed the stain mask to remove small holes, using parameters in[38] (only variance threshold was changed to take into account the difference in slide sizes). To compute the stain ratio for each tile in the TFF3 slides, the tiling operation was applied to the TFF3 slides using the same parameters as H&E slides. Tile coordinates were used to retrieve the foreground mask and DAB mask of each tile in the TFF3 slides. The tile-level stain

ratio was calculated by dividing the number of positive pixels in the tile DAB mask by the total number of positive pixels in the tile foreground mask.

We analyzed the correspondence between the TFF3 stain ratio and model attentions of each tile for BE-positive slides from the test set. We visually inspected the registration results between the corresponding TFF3 and H&E slides by plotting the differences of foreground masks of the registered slides (example in Extended Data, Fig. 3), and ensured that the registration quality was acceptable for most slide pairs for the correspondence analysis. To quantify the correspondence of TFF3 stain ratio and model attentions, we performed several types of analyses. Firstly, we computed normalized stain ratio and normalized model attentions (range 0–1) for each tile in the paired slides and found Pearson's correlation (r) between the two variables, higher values denoting higher correspondence between stain ratio and attention values. We computed normalized entropies of attention distributions for each slide to measure the dispersion of the learned attentions.

Visual inspection of the slides with lower correlation coefficients reveals noisy stain mask extraction due to low contrast and spuriously stained regions in the TFF3 slides, or sub-optimal registration in the case of H&E stain–attention correspondence analysis due to occasional mismatches in amount of tissue present in adjacent H&E- and TFF3-stained sections (e.g. missing pieces, ragged edges).

Failure-modes analysis. We computed the model agreement F_{agree} of H&E and TFF3 BE-TransMIL models on their false predictions (false positives (FPs) or false negatives (FNs)). This was computed as the Jaccard index (intersection over union) of the sets of false predictions made by TFF3 and H&E models independently. In addition, we inspected montages of false predictions made by both models. Specifically, the TFF3 slides were visually inspected to detect errors as the stain is specific for goblet cells, and we found sources of error that can confound the deep learning models occur, including background staining or low contrast between foreground and background. We observed that F_{agree} was much higher for the FNs than for FPs, hence, we prioritized these shared FNs to be manually reviewed by a trained pathologist.

Workflow analysis. We define a workflow as a semi-automatic decision process involving the H&E and TFF3 models as well as manual pathologist review of the corresponding H&E and TFF3 histopathology slides. We first outline two ML-assisted scenarios that still involve manual review by pathologists of all slides to minimize the risk of missed detections. Firstly, the provision of either the H&E or TFF3 BE-TransMIL model outputs (e.g. predictions, attention heatmaps) to pathologists to guide their review and speed-up assessment time of potential positives. Secondly, given the demonstrated prediction performance of the H&E model alone, the need to conduct the more expensive TFF3 staining if a positive finding is confirmed with the H&E model alone can be reduced; thereby lowering preparation costs. Upon demonstration of continued robustness and greater gains over manual pathology reviews, ML-assisted workflows to (semi-)automatically filter out certain cases (detected negatives, for example) could be particularly valuable.

We analyzed the performance (sensitivity, specificity) and the requirement for pathologist review and TFF3 stain for multiple combinations of H&E and TFF3 models as well as pathologist, leading to 14 different ML-assisted workflows (Extended Data, Fig. 4, Table 5). We analyzed the workflow performance on the discovery test dataset at an operating point corresponding to 0.95 sensitivity on the validation set. As majority of our proposed workflows are semi-automated involving a pathologist, this setting can help prevent overlooking suspicious positives. Note that the discovery dataset is heavily enriched for BE-positive cases (38.1%) (Fig. 1c), whereas the expected prevalence in a screening population is 5–12%[5]. Therefore, to simulate the real-world impact of integrating the presented systems into a clinical pathway, we applied importance re-weighting to the samples to achieve a more representative effective prevalence of 8%. Pathologists' workload reduction is computed as the reciprocal of the fraction of manual reviews.

The 14 ML-assisted workflows (Extended Data, Table 5) are named according to Boolean expressions and briefly explained as follows. *"Pathologist"*, *"H&E only"*, and *"TFF3 only"* are workflows involving the detection of BE solely by the pathologist, H&E BE-TransMIL model, and TFF3 BE-TransMIL model, respectively. *"H&E and TFF3"* refers to a workflow where a sample is BE-positive only if both TFF3 and H&E models predict it as BE-positive. *"H&E or TFF3"* workflow will detect a sample as BE-positive if either of the two ML models detect it as positive. The next four workflows are similar configurations as the previous two, combining pathologist and/or either of the two models. The workflow "*H&E and (TFF3 or Pathologist)"* will consider a sample BE-positive if it is labeled as positive by the H&E model and one of the TFF3 model or pathologist. "*H&E and TFF3 and Pathologist"* will consider a sample BE-positive only if it is labeled positive by the pathologist and both ML models. "*(H&E or TFF3) and Pathologist"* workflow will label a sample BE-positive if any of the two ML models and the pathologist label it as BE-positive, whereas "*(H&E and TFF3) or Pathologist"* workflow will consider a sample BE-positive if either both ML models or the pathologist call it BE-positive. Lastly, *"Consensus or Pathologist"* workflow will return the label of a sample as predicted by both ML models if they agree (consensus), otherwise it will consider the label of the pathologist.

Extended Data

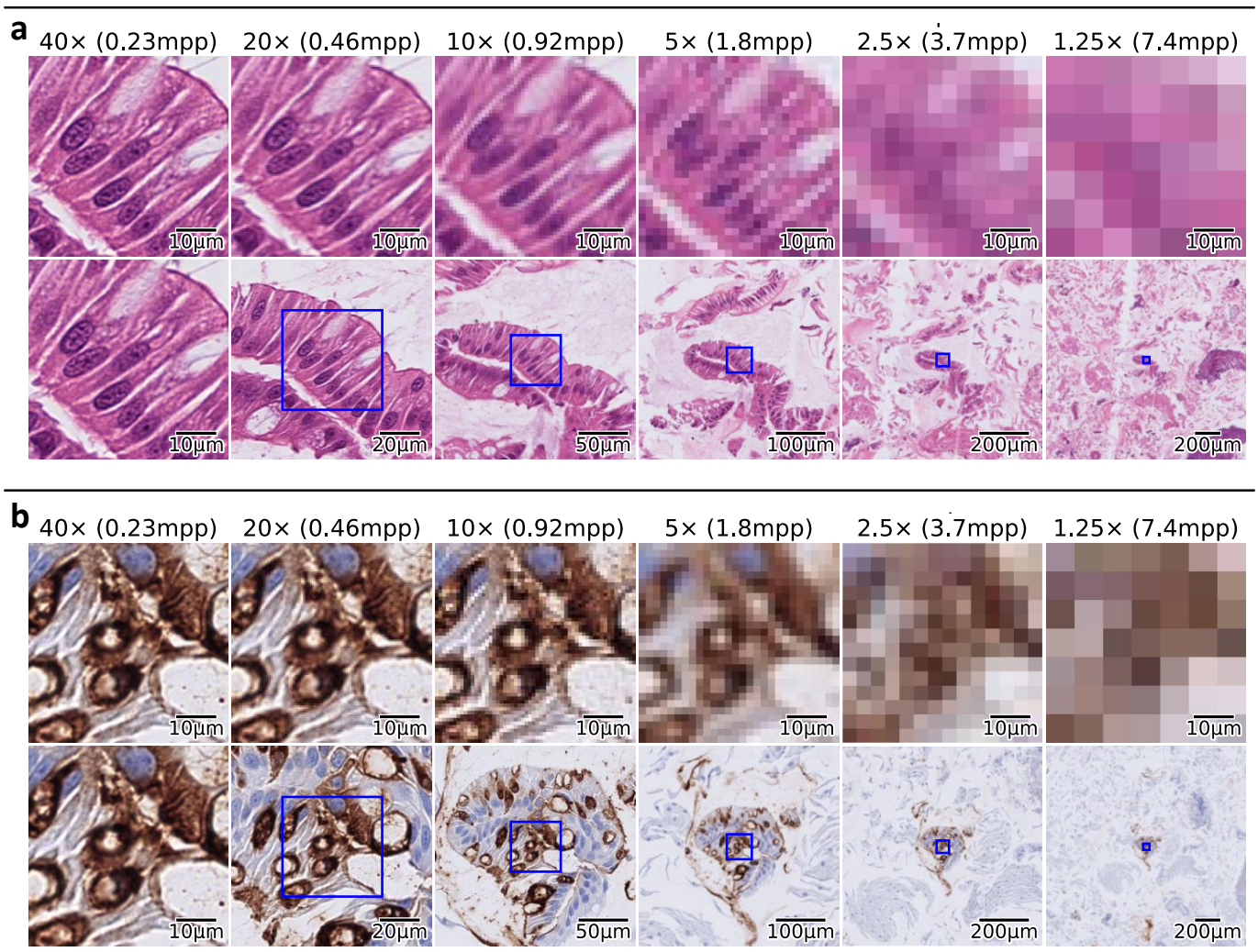


Fig. 1 224×224 tiles of a slide for different objective magnifications: top row shows the same field-of-view at different resolutions, bottom row shows different fields-of-view at different magnifications for a given tile size, with the blue box showing the corresponding region in the first row. **a**, H&E tiles. **b**, TFF3 tiles.

Encoder	AUROC	AUPR	Accuracy	Sensitivity	Specificity
SwinT	0.905 ± 0.018	0.892 ± 0.022	0.836 ± 0.017	0.801 ± 0.045	0.856 ± 0.047
DenseNet121	0.919 ± 0.026	0.907 ± 0.031	0.855 ± 0.025	0.824 ± 0.080	0.874 ± 0.025
ResNet18	0.922 ± 0.011	0.906 ± 0.012	0.849 ± 0.019	**0.833 ± 0.063**	0.858 ± 0.058
ResNet50	**0.931 ± 0.021**	**0.919 ± 0.031**	**0.876 ± 0.047**	0.813 ± 0.060	**0.915 ± 0.077**

Table 1 H&E BE-TransMIL 4-fold cross-validation performance (0.5 probability threshold) on the discovery validation data splits using different types of image encoders. ResNet50 encoder shows most favorable overall performance.

Encoder	AUROC	AUPR	Accuracy	Sensitivity	Specificity
SwinT	0.967 ± 0.006	0.955 ± 0.007	0.918 ± 0.016	**0.908 ± 0.014**	0.925 ± 0.032
DenseNet121	0.965 ± 0.006	**0.958 ± 0.008**	0.914 ± 0.014	0.867 ± 0.050	**0.943 ± 0.031**
ResNet18	0.963 ± 0.002	0.946 ± 0.014	0.907 ± 0.023	0.890 ± 0.030	0.918 ± 0.055
ResNet50	**0.967 ± 0.003**	0.951 ± 0.014	**0.922 ± 0.009**	0.893 ± 0.039	0.939 ± 0.016

Table 2 TFF3 BE-TransMIL 4-fold cross-validation performance (0.5 probability threshold) on the discovery validation data splits using different types of image encoders. ResNet50 encoder shows most favorable overall performance, better or consistent with DenseNet121 encoder.

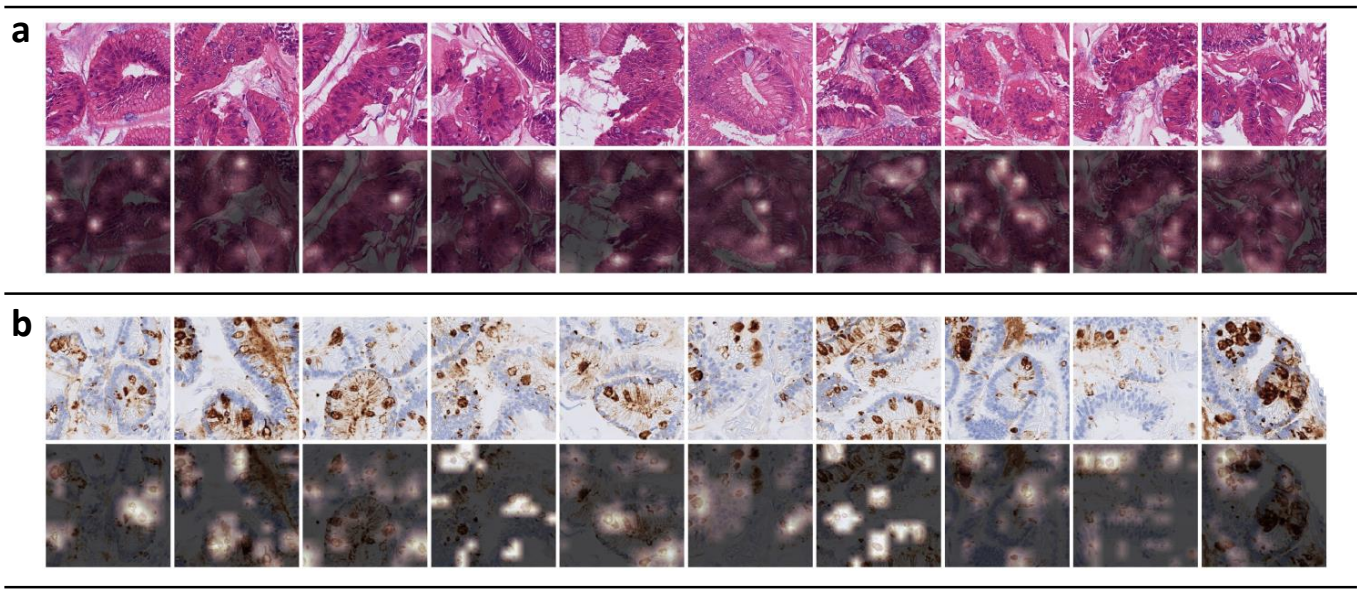


Fig. 2 Grad-CAM saliency maps of the top 10 tiles (with highest attention values) of a BE-positive slide (ResNet50, layer 4). **a**, H&E BE-TransMIL model. **b**, TFF3 BE-TransMIL model.

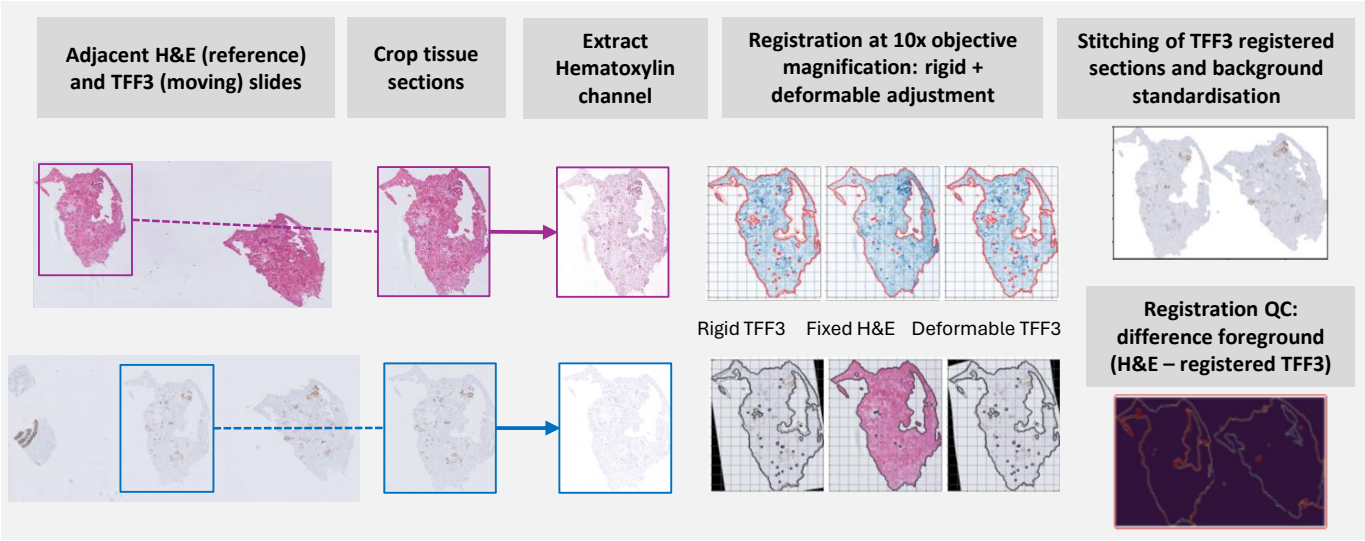


Fig. 3 Registration of adjacent TFF3 and H&E slides.

Mag.	AUROC	AUPR	Accuracy	Sensitivity	Specificity
5×	0.941 ± 0.003	0.918 ± 0.010	0.863 ± 0.0147	**0.857 ± 0.048**	0.867 ± 0.052
10×	**0.960 ± 0.006**	**0.954 ± 0.005**	**0.911 ± 0.014**	0.834 ± 0.033	0.958 ± 0.018
20×	0.953 ± 0.009	0.951 ± 0.007	**0.911 ± 0.011**	0.811 ± 0.059	**0.972 ± 0.032**

Table 3 H&E BE-TransMIL (ResNet50) replication experiments (mean and standard deviation over n=5 random initializations) at different objective magnifications ('Mag.'): performance (0.5 probability threshold) on a 10% data split from the discovery developmental set. 10× objective magnification shows most favorable overall performance. The variance across initializations is consistently small.

Failure type	Total errors	Shared errors	H&E-only errors	TFF3-only errors
False negatives	34 (100%)	19 (55.88%)	13 (38.23%)	2 (6.67%)
False positives	56 (100%)	13 (23.21%)	31 (55.35%)	12 (21.42%)

Table 4 Failure quantities computed for H&E and TFF3 BE-TransMIL models. Percentages in parentheses are with respect to the total number of errors for each failure type.

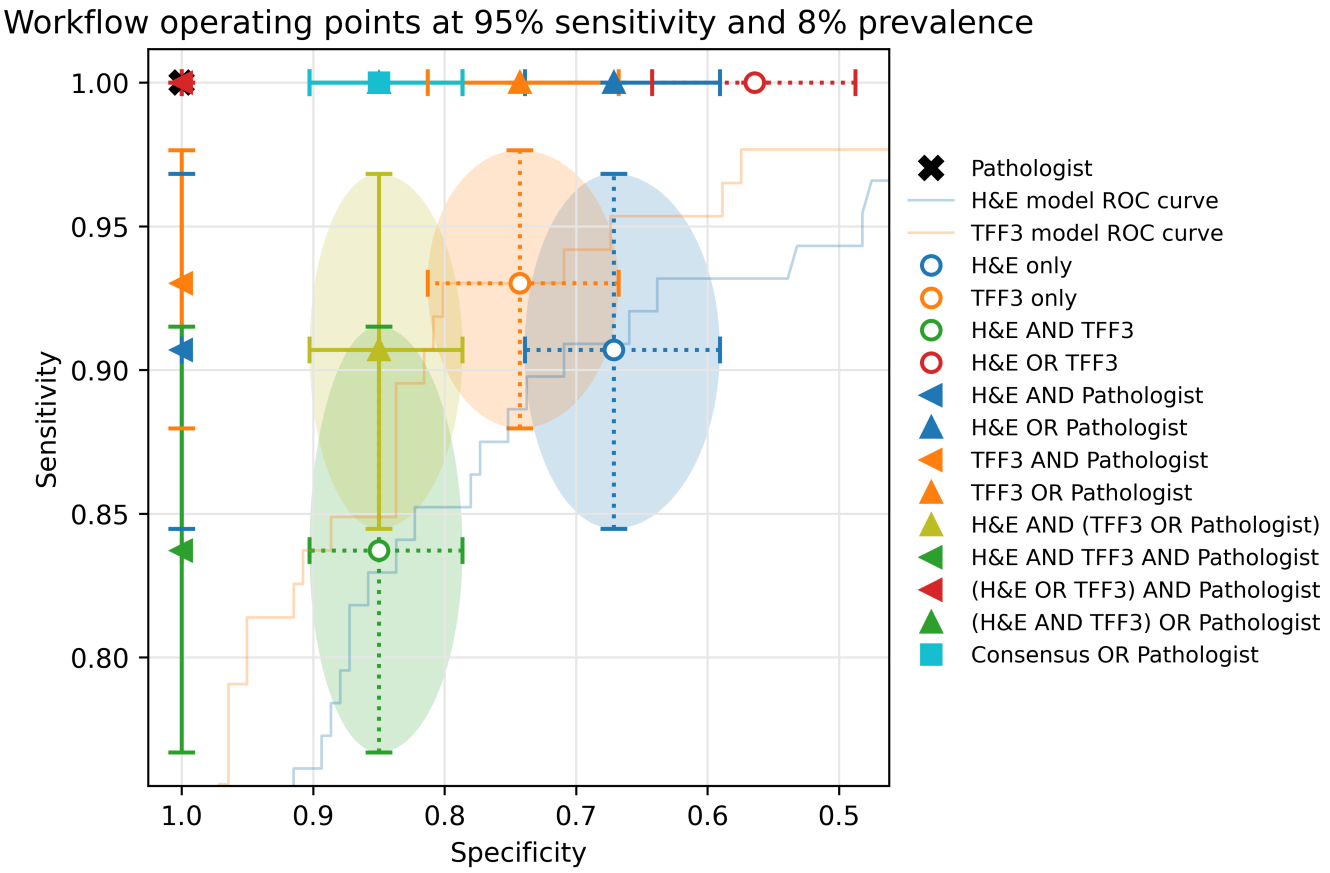


Fig. 4 Performance analysis of multiple ML-assisted workflows. The sensitivity and specificity of each workflow with respect to pathologist (cross at top-left corner) is presented alongside 95% confidence intervals. ROC curves of the H&E and TFF3 models are also presented.

Workflow	Pathologist review	TFF3 staining	Obs. prevalence
Pathologist	100%	100%	8%
H&E only	0%	0%	0%
TFF3 only	0%	100%	0%
H&E AND TFF3	0%	37% [31–45%]	0%
H&E OR TFF3	0%	63% [55–69%]	0%
H&E AND Pathologist	**37% [31–45%]**	**37% [31–45%]**	**19% [16–24%]**
H&E OR Pathologist	63% [55–69%]	63% [55–69%]	1% [1–2%]
TFF3 AND Pathologist	31% [25–38%]	100%	24% [20–31%]
TFF3 OR Pathologist	69% [62–75%]	100%	1% [0–2%]
H&E AND (TFF3 OR Path.)	17% [12–23%]	37% [31–45%]	3% [1–7%]
H&E AND TFF3 AND Path.	20% [16–26%]	37% [31–45%]	33% [26–43%]
(H&E OR TFF3) AND Path.	**48% [41–55%]**	**100%**	**17% [14–20%]**
(H&E AND TFF3) OR Path.	80% [74–84%]	100%	2% [1–2%]
Consensus OR Pathologist	28% [21–35%]	100%	5% [2–8%]

Table 5 Pathologists' workload as a fraction of current reviewed cases, TFF3 staining as a fraction of the current cases, and observed prevalence of BE for possible workflows using the BE-TransMIL models, along with their 95% CIs.

Code availability

The open-source repository https://github.com/microsoft/hi-ml/tree/main/hi-ml-cpath contains the code and library requirements for data preprocessing, core BE-TransMIL network architectures for training and evaluation of deep learning models, and statistical analysis presented in this study. More detailed code will be added to this repository before publication.

Data availability

Data cannot be shared by corresponding author due to license agreements of Cyted Ltd with partners. The study protocols for DELTA and BEST2 are publicly available. Data may be available upon request to the original institution. All data used was deidentified.

Acknowledgments

We would like to thank Rebecca C Fitzgerald for making this data available, the pathologists from Addenbrookes Hospital, Cambridge, UK for their work scoring the BEST2 slides, and the pathologists at Cyted Ltd for scoring the DELTA slides. We would like to extend our thanks to Sophie Ghazal for support that was instrumental in laying the foundation for the study. We also thank Hannah Richardson for guidance offered as part the compliance review of the datasets used in this study. We thank Melissa Bristow for helping in maintaining the open-source repository, and Fernando Peréz García for helping with slide visualization tools. This work was funded by Microsoft Research Ltd (Cambridge, UK).

Author information

These authors contributed equally: Kenza Bouzid, Harshita Sharma

Authors and Affiliations. Microsoft Health Futures, Cambridge, UK: Kenza Bouzid, Harshita Sharma, Daniel Coelho de Castro, Anton Schwaighofer, Max Ilse, Ozan Oktay, Valentina Salvatelli, Anja Thieme, Aditya Nori, Javier Alvarez-Valle
Cyted Ltd, Cambridge, UK: Sarah Killcoyne, Sumanth Murthy, Maria O'Donovan, Marcel Gehrung
Department of Histopathology, Addenbrookes Hospital, Cambridge University NHS Foundation Trust, Cambridge, UK: Luiza Moore, Maria O'Donovan

Contributions. HS, KB, DCC, VS, MI, OO, AS, JAV conceptualized and designed the methods presented in this manuscript. SK performed data collection and preparation. SK and LM provided domain understanding and problem feedback. HS, KB, AS, DCC, MI performed data preprocessing. KB and HS conducted experiments. KB, AS, VS and OO addressed training pipelines scalability for whole slide images. DCC, MI, KB, HS performed statistical analysis. SM and LB performed the external evaluation. MOD provided additional pathology feedback for the failure mode analysis. KB and AS addressed technical and infrastructure related challenges. HS, SK, KB, AS, DCC, MI drafted the manuscript, all authors revised and provided feedback. JAV and MG supervised the study. JAV and AN secured funding.

Corresponding authors. Correspondence to Javier Alvarez-Valle (jaalvare@microsoft.com) and Marcel Gehrung (marcel.gehrung@cyted.ai).

Ethics declaration

Ethics statements. For the DELTA study, ethics approval was obtained from the East of England—Cambridge Central Research Ethics Committee (DELTA 20/EE/0141) and written informed consent was obtained from each patient. For the BEST2 study[11,30], ethics approval was obtained from the East of England—Cambridge Central Research Ethics Committee (BEST2 10/H0308/71) and the trials are registered in the UK Clinical

Research Network Study Portfolio (9461). Written informed consent was obtained from each patient.

Competing interests. MOD is named on patents related to Cytosponge that have been licensed to COVIDien (now Medtronic). SK, SM, and MOD are employees of Cyted Ltd that provides the diagnostic testing for the Cytosponge product. MG is a co-founder and CEO of Cyted Ltd.

List of Acronyms

AUPR area under precision–recall curve

AUROC area under receiver operating characteristic curve

BE Barrett's esophagus

CI confidence intervals

CNN convolutional neural network

DAB 3,3'-diaminobenzidine

EAC esophageal adenocarcinoma

FN false negative

FP false positive

GERD gastroesophageal reflux disease

H&E hematoxylin and eosin

IM intestinal metaplasia

MIL multiple instance learning

ML machine learning

MLP multi-layer perceptron

Swin-T Shifted Window Transformer

TFF3 trefoil factor 3

WSI whole-slide image

References

1. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics. *CA: A Cancer Journal for Clinicians* **66**, 7–30 (2016).
2. Sung, H. *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* **71**, 209–249 (2021).
3. Cameron, A. J., Zinsmeister, A. R., Ballard, D. J. & Carney, J. A. Prevalence of columnar-lined (Barrett's) esophagus: comparison of population-based clinical and autopsy findings. *Gastroenterology* **99**, 918–922 (1990).
4. Richter, J. E. & Rubenstein, J. H. Presentation and epidemiology of gastroesophageal reflux disease. *Gastroenterology* **154**, 267–276 (2018).
5. Shaheen, N. J. *et al.* Diagnosis and management of Barrett's esophagus: an updated ACG guideline. *The American Journal of Gastroenterology* **117**, 559–587 (2022).
6. Modiano, N. & Gerson, L. B. Barrett's esophagus: Incidence, etiology, pathophysiology, prevention and treatment. *Ther. Clin. Risk Manag.* **3**, 1035–1145 (2007).
7. Ross-Innes, C. S. *et al.* Evaluation of a minimally invasive cell sampling device coupled with assessment of trefoil factor 3 expression for diagnosing Barrett's esophagus: A multi-center case–control study. *PLoS Medicine* **12**, e1001780 (2015). URL http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001780https://dx.plos.org/10.1371/journal.pmed.1001780.
8. Paterson, A. L., Gehrung, M., Fitzgerald, R. C. & O'Donovan, M. Role of TFF3 as an adjunct in the diagnosis of Barrett's esophagus using a minimally invasive esophageal sampling device—The Cytosponge™. *Diagnostic Cytopathology* **48**, 253–264 (2020). URL https://doi.org/10.1002/dc.24354.
9. Gehrung, M. *et al.* Triage-driven diagnosis of Barrett's esophagus for early detection of esophageal adenocarcinoma using deep learning. *Nature Medicine* **27**, 833–841 (2021). URL https://doi.org/10.1038/s41591-021-01287-9.
10. Landy, R. *et al.* Real world implementation of non-endoscopic triage testing for Barrett's oesophagus during COVID-19. *QJM: An International Journal of Medicine* (2023). URL https://dx.doi.org/10.1093/qjmed/hcad093.
11. Fitzgerald, R. C. *et al.* Cytosponge-trefoil factor 3 versus usual care to identify Barrett's oesophagus in a primary care setting: a multicentre, pragmatic, randomised controlled trial. *The Lancet* **396**, 333–344 (2020). URL www.thelancet.com.
12. Kaye, P. V. p53 immunohistochemistry as a biomarker of dysplasia and neoplastic progression in Barrett's oesophagus. *Diagnostic Histopathology* **21**, 89–98 (2015).
13. Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V. & Madabhushi, A. Artificial intelligence in digital pathology – new tools for diagnosis and precision oncology. *Nature Reviews Clinical Oncology* **16**, 703–715 (2019).

14. Srinidhi, C. L., Ciga, O. & Martel, A. L. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis* **67**, 101813 (2021).
15. Iizuka, O. *et al.* Deep learning models for histopathological classification of gastric and colonic epithelial tumours. *Scientific Reports* **10**, 1–11 (2020). URL http://dx.doi.org/10.1038/s41598-020-58467-9.
16. Saltz, J. *et al.* Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Reports* **23**, 181–193 (2018).
17. Kather, J. N. *et al.* Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature Cancer* **1**, 789–799 (2020).
18. Fu, Y. *et al.* Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature Cancer* **1**, 800–810 (2020).
19. Myronenko, A., Xu, Z., Yang, D., Roth, H. R. & Xu, D. Accounting for dependencies in deep learning based multiple instance learning for whole slide imaging. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, 329–338 (Springer, Cham, 2021).
20. Lu, M. Y. *et al.* Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering* **5**, 555–570 (2021).
21. Ilse, M., Tomczak, J. M. & Welling, M. Attention-based deep multiple instance learning (2018). 1802.04712.
22. Liu, Z. *et al.* Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992–10002 (2021).
23. Huang, G., Liu, Z. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708 (2017).
24. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, 770–778 (2016).
25. Selvaraju, R. R. *et al.* Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626 (2017).
26. Naini, B. V., Souza, R. F. & Odze, R. D. Barrett's esophagus: a comprehensive and contemporary review for pathologists. *The American journal of surgical pathology* **40**, e45 (2016).
27. Lindvall, M., Lundström, C. & Löwgren, J. Rapid assisted visual search: Supporting digital pathologists with imperfect AI. In *26th International Conference on Intelligent User Interfaces*, 504–513 (2021).
28. Gu, H. *et al.* Augmenting pathologists with NaviPath: Design and evaluation of a human-AI collaborative navigation system. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–19 (2023).
29. Gu, H. *et al.* Improving workflow integration with XPath: Design and evaluation of a human-AI diagnosis system in pathology. *ACM Transactions on Computer-Human Interaction* **30**, 1–37 (2023).
30. Ross-Innes, C. S. *et al.* Risk stratification of Barrett's oesophagus using a non-endoscopic sampling method coupled with a biomarker panel: a cohort study. *The Lancet Gastroenterology & Hepatology* **2**, 23–31 (2017). URL http://linkinghub.elsevier.com/retrieve/pii/S2468125316301182.
31. Janowczyk, A., Zuo, R., Gilmore, H., Feldman, M. & Madabhushi, A. HistoQC: An open-source quality control tool for digital pathology slides. *JCO Clinical Cancer Informatics* 1–7 (2019).
32. Ruifrok, A. C. & Johnston, D. A. Quantification of histochemical staining by color deconvolution. *Analytical and Quantitative Cytology and Histology* **23**, 291–299 (2001).
33. Tomita, N. *et al.* Attention-based deep neural networks for detection of cancerous and precancerous esophagus tissue on histopathological slides. *JAMA Network Open* **2**, e1914645–e1914645 (2019). URL https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2753982.
34. Cardoso, M. J. *et al.* MONAI: An open-source framework for deep learning in healthcare (2022). 2211.02701.
35. Vaswani, A. *et al.* Attention is all you need (2017). 1706.03762.
36. Siskind, J. M. & Pearlmutter, B. A. Divide-and-conquer checkpointing for arbitrary programs with no user annotation. *CoRR* **abs/1708.06799** (2017). URL http://arxiv.org/abs/1708.06799. 1708.06799.
37. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
38. Su, A. *et al.* A deep learning model for molecular label transfer that enables cancer cell identification from histopathology images. *npj Precision Oncology* **6**, 14 (2022).
39. Macenko, M. *et al.* A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 1107–1110 (2009).
40. van der Walt, S. *et al.* scikit-image: image processing in Python. *PeerJ* **2**, e453 (2014).
41. Lowekamp, B., Chen, D., Ibáñez, L. & Blezek, D. The design of SimpleITK. *Frontiers in Neuroinformatics* **7** (2013). URL https://www.frontiersin.org/articles/10.3389/fninf.2013.00045.