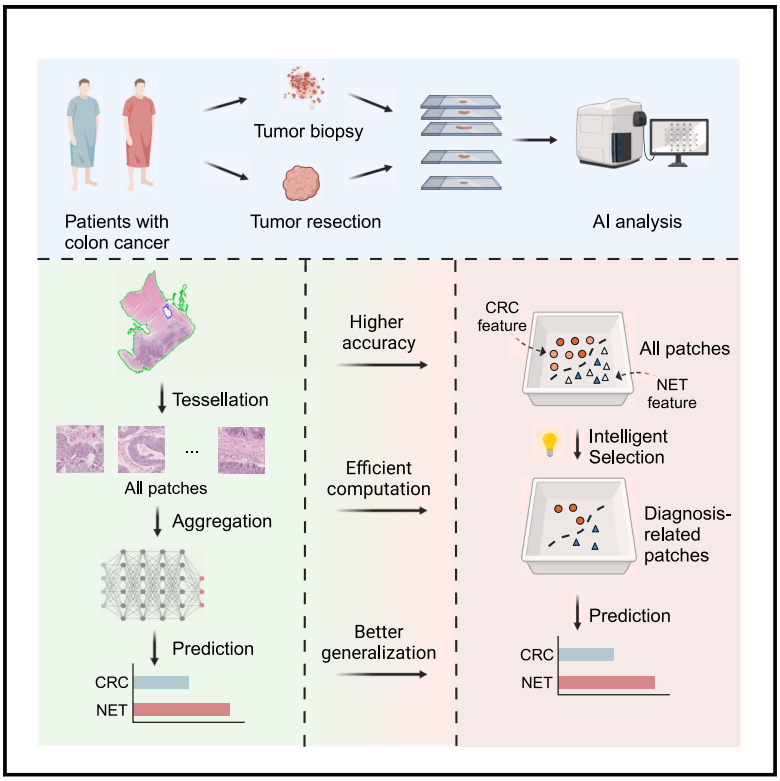
Cell Reports
Medicine

Article

Deep learning model with pathological knowledge for detection of colorectal neuroendocrine tumor

Graphical abstract



Highlights

- Deep learning can discriminate between NET and CRC in biopsied and surgical sections
- The model can identify diagnostically relevant areas akin to experienced pathologists
- Region selection and pre-trained model can improve generalization capability

Authors

Ke Zheng, Jinling Duan, Ruixuan Wang, ..., Yan Sun, Ning Zhang, Muyan Cai

Correspondence

liny36@mail.sysu.edu.cn (Y.L.), sunyan@tjmuch.com (Y.S.), zhangn5@mail.sysu.edu.cn (N.Z.), caimy@sysucc.org.cn (M.C.)

In brief

Zheng et al. develop a deep learning model that accurately distinguishes between NET and CRC using a limited number of diagnostically relevant patches. By utilizing region selection and a pre-trained model, they improve both model performance and generalization ability across various cohorts.

Zheng et al., 2024, Cell Reports Medicine 5, 101785
October 15, 2024 © 2024 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.xcrm.2024.101785

CellPress

Cell Reports Medicine

CellPress
OPEN ACCESS

Article

Deep learning model with pathological knowledge for detection of colorectal neuroendocrine tumor

Ke Zheng,[1,8] Jinling Duan,[1,8] Ruixuan Wang,[2,8] Haohua Chen,[3,8] Haiyang He,[4] Xueyi Zheng,[1] Zihan Zhao,[1] Bingzhong Jing,[3] Yuqian Zhang,[5] Shasha Liu,[6] Dan Xie,[1] Yuan Lin,[4,*] Yan Sun,[6,*] Ning Zhang,[7,*] and Muyan Cai[1,9,*]

[1]Department of Pathology, State Key Laboratory of Oncology in South China, Guangdong Provincial Clinical Research Center for Cancer, Sun Yat-sen University Cancer Center, Guangzhou 510060, China
[2]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China
[3]Artificial Intelligence Laboratory, State Key Laboratory of Oncology in South China, Guangdong Provincial Clinical Research Center for Cancer, Sun Yat-sen University Cancer Center, Guangzhou 510060, China
[4]Department of Pathology, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou 510080, China
[5]Electrical Engineering & Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA
[6]Department of Pathology, Tianjin Medical University Cancer Institute and Hospital, Tianjin 300000, China
[7]Department of Gastroenterology and Hepatology, The First Affiliated Hospital, Sun Yat-Sen University, Guangzhou 510060, China
[8]These authors contributed equally
[9]Lead contact
*Correspondence: liny36@mail.sysu.edu.cn (Y.L.), sunyan@tjmuch.com (Y.S.), zhangn5@mail.sysu.edu.cn (N.Z.), caimy@sysucc.org.cn (M.C.)
https://doi.org/10.1016/j.xcrm.2024.101785

SUMMARY

Colorectal neuroendocrine tumors (NETs) differ significantly from colorectal carcinoma (CRC) in terms of treatment strategy and prognosis, necessitating a cost-effective approach for accurate discrimination. Here, we propose an approach for distinguishing between colorectal NET and CRC based on pathological images by utilizing pathological prior information to facilitate the generation of robust slide-level features. By calculating the similarity between morphological descriptions and patches, our approach selects only 2% of the diagnostically relevant patches for both training and inference, achieving an area under the receiver operating characteristic curve (AUROC) of 0.9974 on the internal dataset, and AUROCs of 0.9724 and 0.9513 on two external datasets. Our model effectively identifies NETs from CRCs, reducing unnecessary immunohistochemical tests and enhancing the precise treatment for patients with colorectal tumors. Our approach also enables researchers to investigate methods with high accuracy and low computational complexity, thereby advancing the application of artificial intelligence in clinical settings.

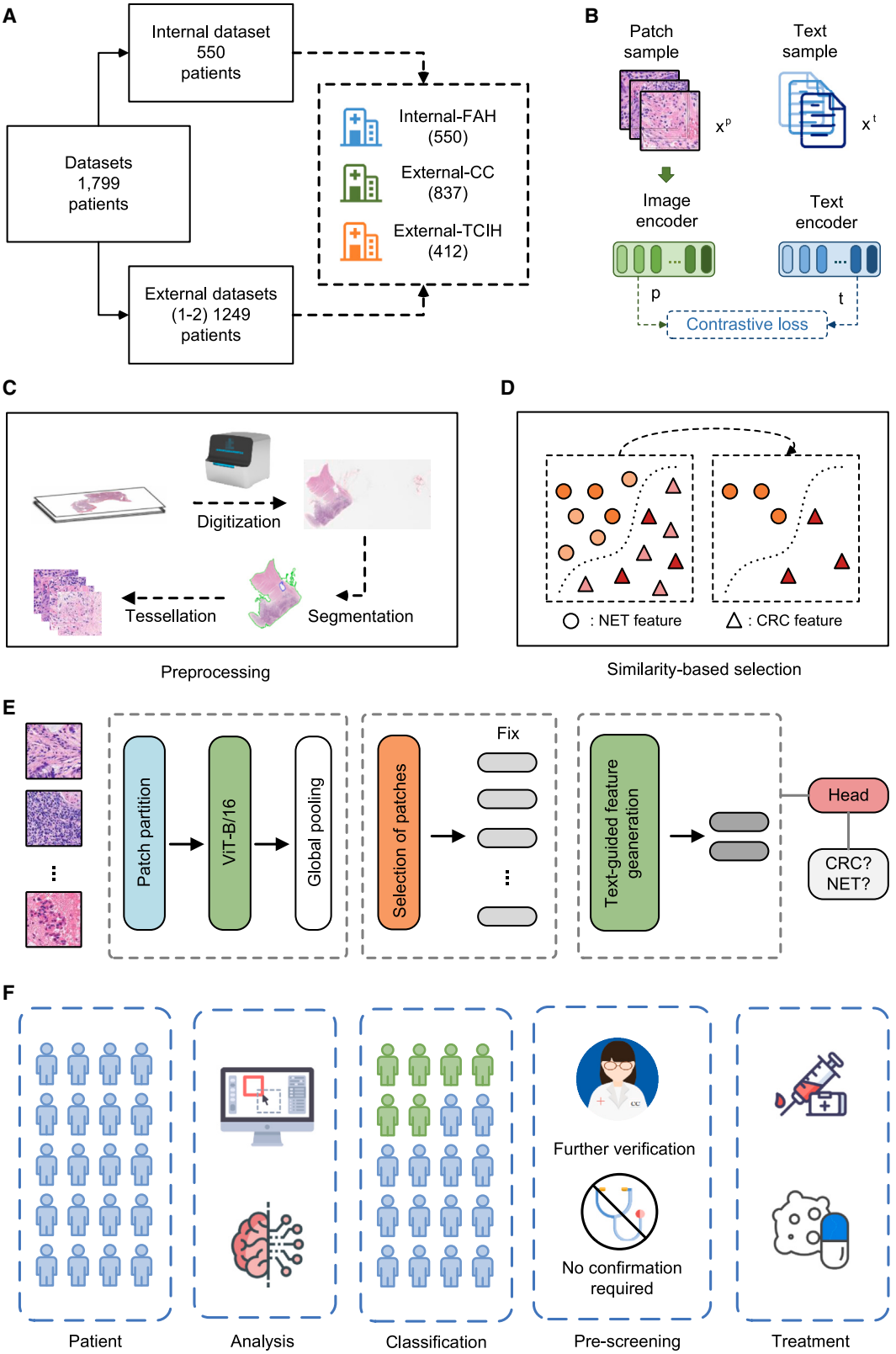
INTRODUCTION

Neuroendocrine tumors (NETs) are a group of rare neoplasms originating from neuroendocrine cells throughout the body, known for producing peptide hormones and biogenic amines.[1–3] The gastrointestinal tract is the most common site of NET occurrence. The frequency of colorectal NETs is increasing, attributed to universal screening colonoscopies and enhanced endoscopic imaging quality.[4] Colorectal NETs differ significantly from colorectal adenocarcinoma in terms of molecular subtypes, treatment strategies, and prognostic outcomes. Consequently, distinguishing colorectal NETs from colorectal adenocarcinoma is crucial for personalized treatment.[5,6]

In clinical settings, differentiation between colorectal NET and adenocarcinoma is typically performed through histopathological examination of hematoxylin and eosin (H&E)-stained sections. Morphologically, NETs often exhibit solid, trabecular, gyriform, or glandular patterns, with uniform nuclei, coarsely stippled chromatin, and finely granular cytoplasm.[7] However, these features can sometimes overlap with those of colorectal adenocarcinoma. Therefore, additional immunohistochemical biomarkers, such as synaptophysin, chromogranin, neural-specific enolase, and CD56, are utilized to identify NETs. Given the rarity of NETs, these approaches are labor-intensive for pathologists and not cost-effective. Hence, effective methods need to be developed to differentiate NETs from colorectal adenocarcinoma.

Deep learning has shown promise in identifying histomorphological patterns and disease-specific features,[8,9] with potential applications for automated biomarkers. Recent studies have revealed that deep learning can classify routine H&E-stained, formalin-fixed, paraffin-embedded digital whole-slide images (WSIs) of colorectal cancer into microsatellite stable and microsatellite instability categories, outperforming board-certified pathologists.[10–12] Furthermore, the incorporation of extensive pre-training in pathology has substantially improved the ability of models to extract morphological characteristics.[13–16] These have sparked interest in using deep learning models to identify additional NET characteristics that may not be readily apparent to pathologists, thus providing an automated screening tool to triage patients for

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).



(legend on next page)

confirmatory testing of NETs. Therefore, we intend to develop and validate a deep learning-based classification method in colorectal samples and extend its application to biopsy tissues, where limited observation fields and diagnostic challenges exist.

In current setups, deep learning approaches for WSI classification typically involve extracting features from all patches and then integrating them into a slide-level feature for final prediction.[17–21] For example, dividing a slide into patches of 256 × 256 pixels at × 40 magnification can result in tens of thousands of patches, many of which may be irrelevant to the tumor. The mixture of tumor-relevant and irrelevant patches often hinders the deep learning model's ability to effectively learn and identify pathological tumor features. Previous studies have attempted to enhance model performance through the implementation of transformer architectures.[11,22] However, the computational complexity of transformers escalates as the number of tokens increases. These approaches differ from standard clinical practice, where pathologists diagnose based on identified tumor regions using their prior knowledge of pathology. Compared to learning from diagnostic reports as supplementary information,[23] prior pathological knowledge is more valuable and easily accessible. Encoding this pathology knowledge as text and integrating it into the deep learning model could enable the model to more accurately locate diagnosis-related regions, thereby improving data efficiency.

In this study, we propose a deep learning approach to distinguishing colorectal NETs from colorectal adenocarcinoma with enhanced generalization capability and improved performance. Departing from the prevailing practice of predicting based on all patches, our approach intelligently selects key patches and generates robust slide-level features based solely on the selected small subset of patches. Specifically, a similarity-based selection method is introduced to exclude diagnosis-irrelevant patches, allowing the model to focus on clinically significant regions. Moreover, the morphological description is considered as a text prototype, uniform and independent of color variations across datasets. We utilize this text prototype as the prior knowledge about each cancer type to guide the generation of slide-level features. Our approach shows great performance in a multi-centric study involving three cohorts of over 1,500 patients with surgical samples, and an additional cohort of biopsy sections.

RESULTS

Patient cohorts

Three cohorts of surgical samples and a cohort of biopsied samples were included in this study, i.e., Internal-The First Affiliated Hospital of Sun Yat-sen University (FAH), External-Sun Yat-sen University Cancer Center (CC), External-Tianjin Medical University Cancer Institute & Hospital (TCIH), and Biopsy-CC (Figure 1A). Each H&E-stained WSI in these cohorts was collected from an individual patient. The Internal-FAH cohort serves as the internal dataset for training the deep learning model, comprising 130 patients with NET and 420 patients with colorectal adenocarcinoma. External-CC and External-TCIH are used as two independent external validation datasets. The External-CC cohort includes 837 patients (706 diagnosed with colorectal adenocarcinoma and 131 with NET), and the External-TCIH cohort consists of 311 patients with colorectal carcinoma (CRC) and 101 patients with NET. Additionally, the Biopsy-CC cohort comprises biopsy data which are used to validate the model's potential for early screening, including 108 slides of NET and 315 slides of colorectal adenocarcinoma.

A novel framework for distinguishing colorectal NET

In this study, the proposed model primarily comprised pre-trained image and text encoders,[13] a diagnostic-related patches selection module, an attention-based aggregation module, and a text-guided slide-level feature generation module. The encoders were pre-trained using a large-scale dataset of pathological images. The pre-trained image encoder was used to extract patch-level features, while the pre-trained text encoder was employed to transform cancer-specific descriptors into text features. The diagnostic-related patches selection module eliminated regions unrelated to diagnosis from the whole slide. Specifically, the similarity between each patch and cancer-specific descriptors was computed, and patches with the highest similarity were selected as inputs for subsequent modules. The design for extracting diagnostically relevant patches was primarily inspired by our observations of the diagnosis process from clinical pathologists, i.e., pathologists can quickly locate key regions of a slide rather than relying on the entire slide for a diagnosis. This approach sharply contrasts with current deep learning pipelines. Then, the attention mechanism and text-prior information were used to aggregate patch features into slide-level features. This aided the model in further focusing on the regions most relevant to the diagnosis.

Twelve deep learning models were constructed, including a baseline model using all patches as input, ten models using varying proportions of patches (subsequently referred to as the similarity-based model), and the text-prior model with 2% patches as input. It is worth noting that the difference between the similarity-based model and the text-prior model was that the text-prior model had a text-guided slide-level feature generation module. The subsequent sections of this part first introduced the three types of models in sequence. Then, the model trained by the

Figure 1. The workflow of the proposed deep learning model

(A) The datasets were collected from three different centers containing more than 1,500 patients. Data from one center were used as an internal dataset for model training, while data from the other two centers were utilized to construct external datasets for testing the model's generalization.

(B) The image encoder and the text encoder employed in our model were trained through contrastive learning on large-scale pathology image-text pairs. This training strategy enhanced the encoders' capabilities, enabling them to capture more robust representations.

(C) After digitizing the slides, the tissue regions were segmented, and the whole-slide images were decomposed into patches.

(D) A similarity-based selection method was used to extract diagnostically relevant patches from the whole-slide images.

(E) The computational flow of the model is mainly divided into three parts: diagnosis-related patch selection, text-guided slide-level feature generation, and prediction.

(F) Our model can be applied in clinical settings for early screening, significantly reducing the workload of pathologists and minimizing the need for additional diagnostic testing.

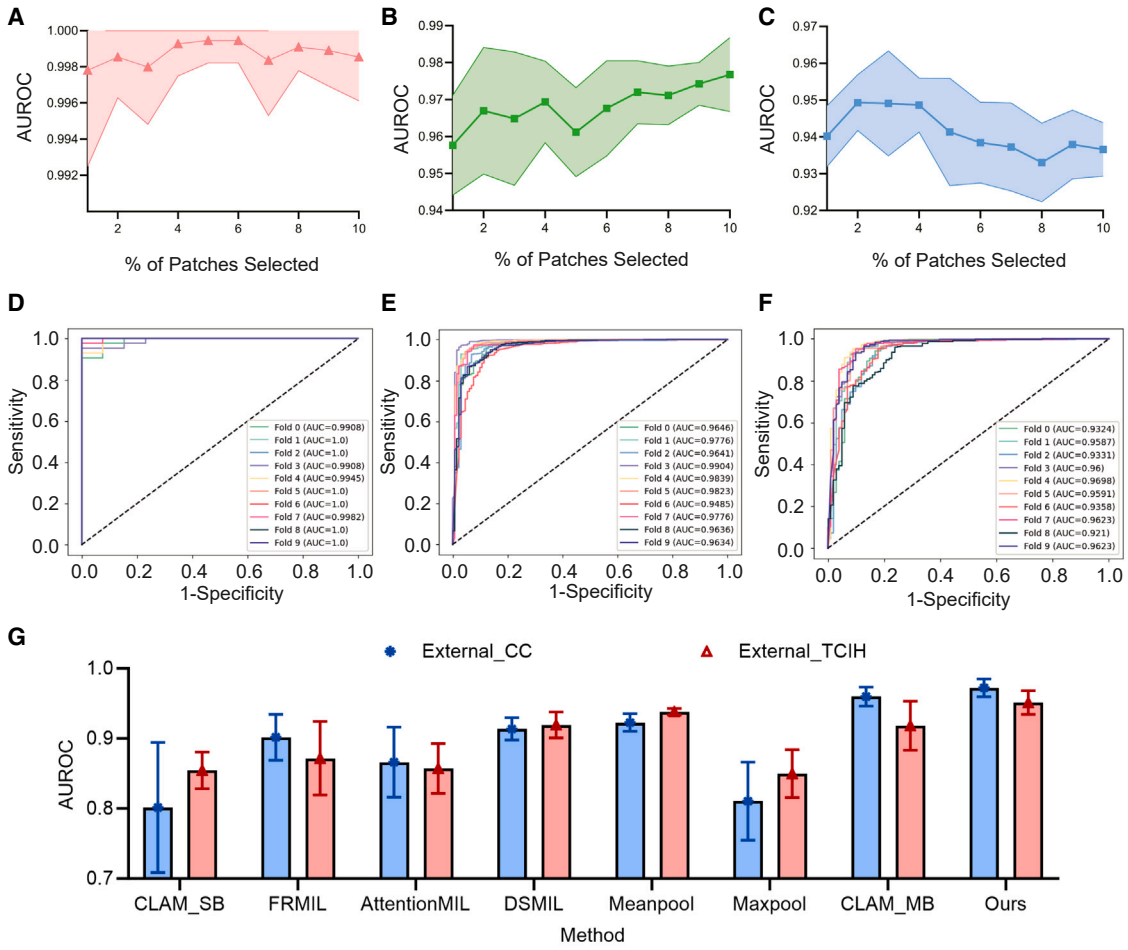


Figure 2. The performance of our models on three datasets: an internal dataset (Internal-FAH) and two external datasets (External-CC and External-TCIH)

(A–C) The performance of the model with varying numbers of selected patches. (A), (B), and (c) correspond to the Internal-FAH, External-CC, and External-TCIH datasets, respectively.

(D) The model achieved an AUROC score of 0.9974 on the internal dataset.

(E) On the External-CC dataset, the model maintained an AUROC of 0.9724.

(F) On the External-TCIH dataset, the model achieved an AUROC of 0.9513.

(G) When compared to existing models on the external datasets, our model consistently outperformed them.

proposed framework was compared to state-of-the-art (SOTA) methods, demonstrating the superiority of the proposed model. Finally, visualization tools were used to interpret the model, and the model was further extended to the biopsy dataset.

Baseline model performance using all patches as input

Initially, the baseline model's performance was assessed using all patches as inputs (Table S1). Evaluation on the internal dataset revealed an area under the receiver operating characteristic curve (AUROC) of 0.9987 (95% confidence interval [CI] 0.9972–1), with a sensitivity of 0.9929 (95% CI 0.9861–0.9996) and a specificity of 0.9615 (95% CI 0.9296–0.9935). Subsequent assessment on the External-CC dataset yielded an AUROC of 0.9325 (95% CI 0.9230–0.9421), a sensitivity of 0.9975 (95% CI 0.9959–0.9990), and a specificity of 0.5718 (95% CI 0.4763–0.6672). Similarly, evaluation on the External-TCIH dataset resulted in an AUROC of 0.9097 (95% CI 0.8993–0.92), with a sensitivity of 0.9997 (95% CI 0.9991–1) and a specificity of 0.5129 (95% CI 0.4075–0.6182). Although the baseline model performed well on the internal dataset, it exhibited limited generalization ability and poor specificity. Due to the limited training data and the inherent complexity of pathology data, deep learning models are prone to overfitting. Consequently, overfitting leads to diminished performance on external datasets, particularly with a corresponding decrease in sensitivity or specificity. We further explored models that offer superior performance and greater computational efficiency.

The effect of diagnosis-related patch extraction on model performance

To improve computational efficiency and performance, a similarity-based approach was introduced. This approach relies on

Table 1. Model performance on the internal dataset and two external datasets

Cohorts	Predictive performance		
	Sensitivity (95% CI)	Specificity (95% CI)	AUROC (95% CI)
Internal-FAH	0.9833 (0.9712–0.9954)	0.9692 (0.9359–1.0000)	0.9974 (0.995–0.9998)
External-CC	0.9402 (0.9018–0.9786)	0.8533 (0.7732–0.9332)	0.9724 (0.9645–0.9801)
External-TCIH	0.9382 (0.8868–0.9896)	0.7547 (0.6765–0.8329)	0.9513 (0.9408–0.9618)

Note: 95% confidence intervals are included in brackets; AUROC, the area under the receiver operating characteristic.

morphological descriptors to pinpoint and prioritize patches that most closely align with the provided description. In this section, we systematically validated the model by selecting the top 1% to top 10% patches with the highest similarity scores, chosen at 1% intervals. On the Internal-FAH dataset (Figure 2A; Table S2), the transition from selecting 10%–1% of the patches yielded AUROCs of 0.9978 (95% CI 0.9945–1) to 0.9995 (95% CI 0.9987–1). Likewise, for the External-CC dataset (Figure 2B; Table S2), AUROC values increased from 0.9577 (95% CI 0.9493–0.9661) to 0.9768 (95% CI 0.9706–0.9830) as the proportion of selected patches decreased from 10% to 1%. Similarly, in the External-TCIH cohort (Figure 2C; Table S2), AUROC values ranged from 0.9331 (95% CI 0.9265–0.9397) to 0.9493 (95% CI 0.9446–0.954) across the same range of selected patches. Corresponding receiver operating characteristic curves (ROCs) are provided in Figure S1. Taken together, the performance of the proposed similarity-based model significantly exceeded that of the baseline model, underscoring the potential of the proposed model to enhance both computational efficiency and diagnostic performance.

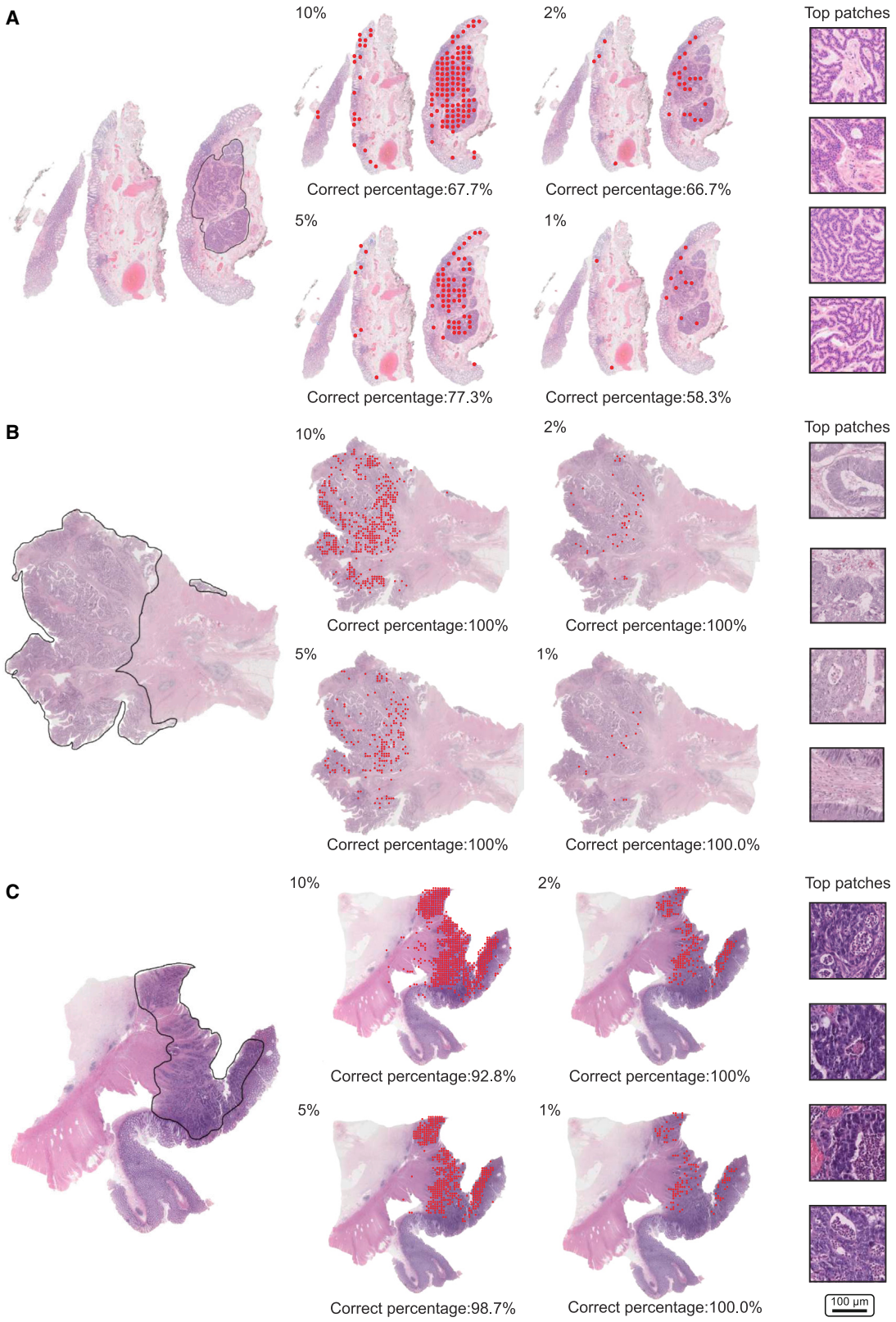
Performance of the text-prior model

The model demonstrates a reduced sensitivity to the patch selection ratio. Considering both computational efficiency and performance, the input was limited to only 2% of diagnostically pertinent patches, subsequently generating text-guided slide-level features with the text-prior model architecture. On the Internal-FAH dataset (Tables 1 and S3; Figure S2), the AUROCs across the testing set in each fold ranged from 0.9908 to 1. Cumulatively across testing folds, the model achieved an AUROC of 0.9974 (95% CI 0.995–0.9998), an accuracy of 0.9800 (95% CI 0.9676–0.9924), an F1-score of 0.9868 (95% CI 0.9787–0.9949), a sensitivity of 0.9833 (95% CI 0.9712–0.9954), and a specificity of 0.9692 (95% CI 0.9359–1). Furthermore, we scrutinized the model's generalization capacity on external datasets. On the External-CC dataset, the model attained an AUROC of 0.9724 (95% CI 0.9645–0.9801), a sensitivity of 0.9402 (95% CI 0.9018–0.9786), and a specificity of 0.8533 (95% CI 0.7732–0.9332). Similarly, on the External-TCIH cohort, the model achieved an AUROC of 0.9513 (95% CI 0.9408–0.9618), a sensitivity of 0.9382 (95% CI 0.8868–0.9896), and a specificity of 0.7547 (95% CI 0.6765–0.8329). ROCs illustrating the model's performance across different datasets are described in Figures 2D–2F. Ablation experiments for each module are detailed in Table S4. These outcomes

Table 2. The comparison of our method's performance and SOTA methods' performance on two external cohorts

Cohorts	Method	Predictive performance			p[a]
		Sensitivity (95% CI)	Specificity (95% CI)	AUROC (95% CI)	
External-CC	our method	0.9402 (0.9018–0.9786)	0.8533 (0.7732–0.9332)	0.9724 (0.9645–0.9801)	NA
	CLAM-SB	0.9956 (0.9879–1.000)	0.1466 (0.0309–0.2622)	0.8014 (0.7438–0.8590)	<0.001
	CALM-MB	0.9975 (0.9963–0.9986)	0.5893 (0.5243–0.6543)	0.9598 (0.9514–0.9682)	0.0606
	AttentionMIL	0.9949 (0.9907–0.9992)	0.5076 (0.3931–0.6222)	0.8664 (0.8354–0.8975)	<0.001
	DSMIL	0.9759 (0.9343–1.0000)	0.7870 (0.6994–0.8747)	0.9137 (0.8978–0.9296)	<0.001
	FRMIL	0.9367 (0.8151–1.0000)	0.7130 (0.6254–0.8006)	0.9017 (0.8690–0.9344)	<0.001
	MeanPooling	0.9999 (0.9994–1.0000)	0.1985 (0.0853–0.3116)	0.9227 (0.9101–0.9354)	<0.001
	MaxPooling	0.9878 (0.9757–0.9999)	0.3626 (0.1827–0.5424)	0.8105 (0.7548–0.8663)	<0.001
External-TCIH	our method	0.9382 (0.8868–0.9896)	0.7547 (0.6765–0.8329)	0.9513 (0.9408–0.9618)	NA
	CLAM-SB	0.9900 (0.9724–1.0000)	0.1991 (0.0615–0.3367)	0.8544 (0.8382–0.8707)	<0.001
	CALM-MB	1.0000 (1.0000–1.0000)	0.5468 (0.4693–0.6243)	0.9182 (0.8966–0.9398)	0.0173
	AttentionMIL	0.9982 (0.9965–0.9999)	0.4367 (0.3259–0.5476)	0.8572 (0.8351–0.8794)	<0.001
	DSMIL	0.9728 (0.9343–1.0000)	0.7041 (0.6994–0.8747)	0.9193 (0.9008–0.9377)	<0.001
	FRMIL	0.8442 (0.6000–1.0000)	0.7481 (0.6664–0.8639)	0.8718 (0.8193–0.9242)	<0.001
	MeanPooling	1.0000 (1.0000–1.0000)	0.1496 (0.0202–0.1888)	0.9382 (0.9337–0.9426)	0.0478
	MaxPooling	0.9836 (0.9553–1.0000)	0.5115 (0.3324–0.7039)	0.8499 (0.8157–0.8841)	<0.001

Note: 95% confidence intervals are included in brackets.
[a]Indicates the comparison of the difference between other models and our method; AUROC, the area under the receiver operating characteristic; NA, not applicable.



(legend on next page)

collectively indicate that the text-prior model, leveraging only 2% of diagnostically relevant patches, exhibits robust performance in diagnosing NET across diverse datasets.

Comparative performance analysis of deep learning methods

For a comprehensive evaluation of our model's performance, a rigorous comparative analysis was conducted against SOTA methods on external datasets. These methods encompassed a variety of approaches, including clustering-constrained attention multiple instance learning network (CLAM)[24] (CLAM-SB and CALM-MB mentioned in the original article), attention based multiple instance learning network (AttentionMIL),[25] dual-stream multiple instance learning network (DSMIL),[18] feature re-calibration based multiple instance learning network (FRMIL),[26] MeanPooling, and MaxPooling. To ensure fair comparison, all models underwent training and assessment via 10-fold cross-validation within a consistent experimental setup. Specifically, we maintained uniformity by employing the same patch-level feature extractor across all methods, while also adhering to consistent training hyperparameters and loss functions for supervision. The results showed that the AUROCs of the SOTA methods ranged from 0.8014 to 0.9598 on the External-CC dataset and from 0.8499 to 0.9382 on the External-TCIH dataset, all of which were surpassed by our proposed method (Figure 2G; Table 2). Furthermore, McNemar's test results indicated a statistically significant difference between our method and other methods (Table S5). We also examined both the training and inference times of the models (Table S6). Our findings indicate that the proposed model requires only two-thirds of the training time needed by other models while also demonstrating superior speed during inference. This dual advantage underscores the model's efficacy in terms of accuracy and efficiency.

Interpretability and feature visualization

To evaluate the efficacy of the patch selection method, the model's selection of varying proportions of patches was visually depicted (Figures 3A–3C and S3). Remarkably, the regions identified by the method closely correspond with tumor regions manually delineated by the pathologists. Furthermore, the patches chosen by the model exhibit a strong concordance with the textual information provided by the pathologists (Figure S4). These localized patches effectively capture representative morphological features specific to CRC or NET, suggesting the model's potential to discern diagnostically significant regions for tumor classification.

To further assess the discriminative nature of different classes of patches at the feature level, a series of analyses were conducted about feature separability and visualization. Initially, 512-dimensional features were extracted using the patch-level feature extractor. To facilitate a more effective comparison of the feature separability between CRC and NET, patches were randomly sampled in proportions approximating the categories in the dataset. Subsequently, the uniform manifold approximation and projection (UMAP) technique[27] was employed to visualize the features. Notably, two sampling approaches were compared, one entailing sampling from all patches, while the other involved sampling from the 2% of patches selected according to the proposed method in this study. When sampling from all patches, a significantly higher degree of feature indistinguishability between the two categories was observed, particularly evident on the Internal-FAH and External-TCIH datasets (Figures 4A and 4C). Conversely, when sampling from the 2% of selected patches as per the described approach, the remaining patches exhibited significantly improved separability in the feature space, with a reduced overlap fraction. This effect was particularly pronounced on the External-CC dataset (Figure 4B), where features from the two categories displayed a higher degree of spatial separability. These results provided evidence to support that the selected patches were more distinguishable at the feature level, thus mitigating the complexity in subsequent classification tasks.

Adaptation of resection-trained models for biopsy analysis

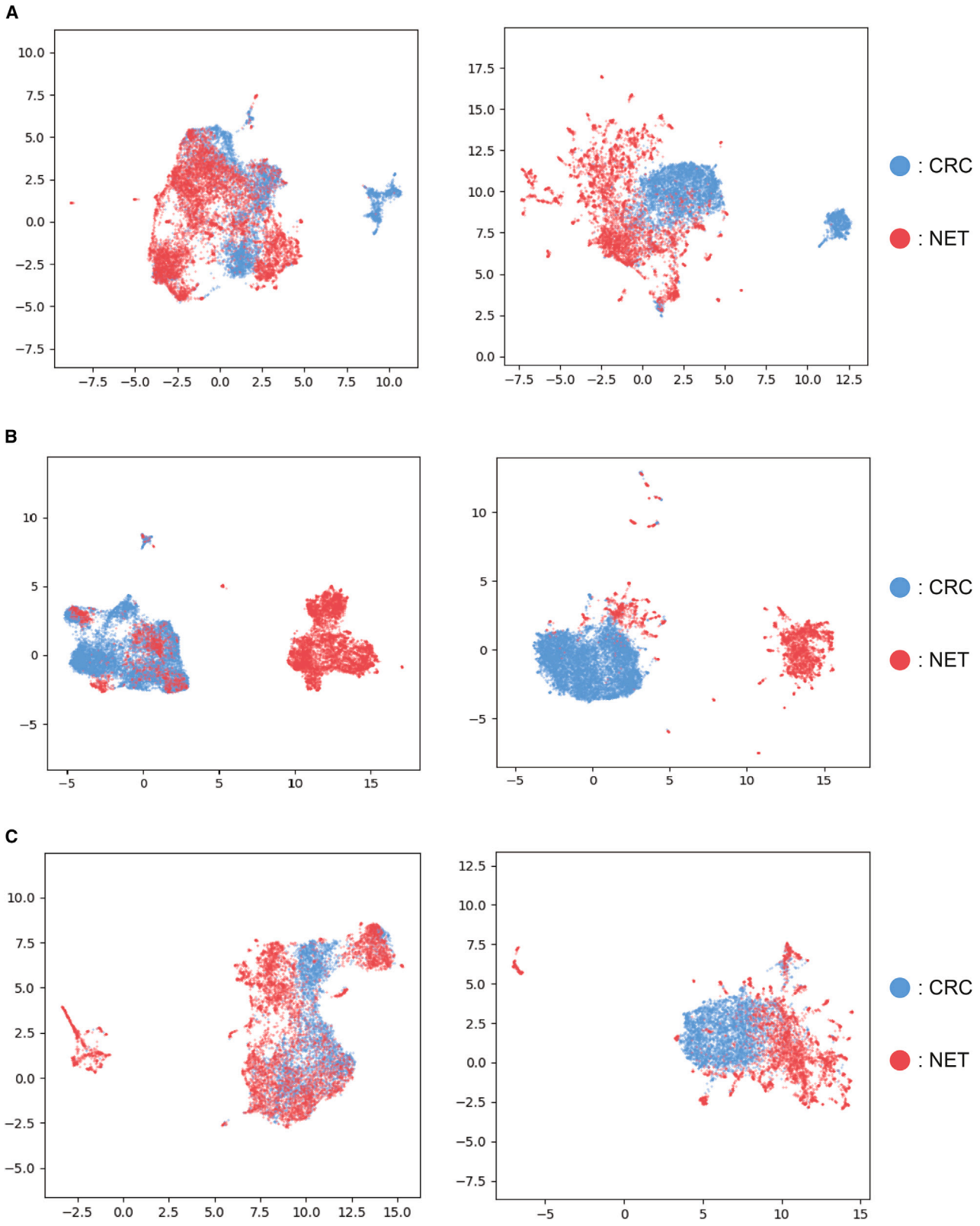
In light of the diagnostic complexities inherent in biopsy clinical settings, the application scope of our model was expanded to include the Biopsy-CC dataset (Figure S5A). Acknowledging the substantial disparity in amount of patches between biopsied and surgical sections, the patches ratio was further redefined within the biopsy cohort. Initially, we assessed various percentages of patches as model inputs (Figures S5B–S5F). Notably, comparable performance levels were observed when utilizing 10%, 15%, and 20% of patches as inputs. However, a noticeable decline in performance was observed when only 2% and 5% of patches were utilized. Based on considerations of performance and computational efficiency, we selected 10% of patches, which represented the most suitable input configuration, for further analysis. Employing this chosen input configuration (Figure S5D; Table S7), our model achieved an AUROC of 0.915 (95% CI 0.9031–0.9269). The data underscored the robust performance and computational efficiency of our deep learning model, even amid the challenges presented by biopsy clinical scenarios.

DISCUSSION

Accurately differentiating between colorectal NETs and colorectal adenocarcinomas is crucial due to the significant differences in treatment strategies and prognoses associated with these conditions. Traditionally, pathologists have relied on the evaluation of H&E-stained sections and the use of immunohistochemical biomarkers to aid in the diagnosis of suspected colorectal NETs. To address the need for a more cost-effective and efficient diagnostic approach, we developed a data-efficient deep learning

Figure 3. Visualization of selected patches

(A–C) The left figure displays the manual outlines performed by the pathologist. The middle figure shows patches selected in different proportions projected onto the image. The right figure presents representative patches chosen based on similarity. Correct proportion indicates the proportions of model-selected patches within the pathologist's labeled region.



(legend on next page)

model that selectively identifies a small number of diagnosis-related patches as inputs. This method not only improves computational efficiency but also enhances diagnosis performance and generalization ability. Moreover, by incorporating cancer-related descriptions at the slide level, we achieved superior performance on both external validation and biopsy datasets.

This study is the first to employ a similarity-based method for selection of diagnosis-related patches. Notably, this approach showed no discernible performance impact on the internal dataset compared to models that use all available patches. However, it led to marked improvements in generalization performance on external datasets. Even with only 1% of the patches selected as input, the model's performance on both external datasets was enhanced. The limited nature of the training data, coupled with the high inherent complexity of pathology data, often leads to overfitting in models trained on WSIs. Our proposed method effectively reduces the inclusion of irrelevant diagnostic patches, thereby decreasing task complexity and mitigating the overfitting. Consequently, the model's generalization ability is strengthened, resulting in better overall performance. This observation underscores the potential of selectively identifying tumor regions to reduce computational demands without compromising diagnostic accuracy. Moreover, our results suggest that reducing the number of patches from 10% to 1% does not significantly impact diagnostic performance, with AUROC differences remaining below 0.02. This indicates that the model is not highly sensitive to the number of selected patches, paving the way for further optimization of patch selection strategies. A particularly noteworthy observation is the disparity in patch numbers between the External-TCIH and Internal-FAH (Table S8). Despite this difference, the model consistently yields satisfactory results on the External-TCIH dataset, underscoring its ability to manage varying quantities of patches effectively.

Existing methods often suffer from performance degradation when applied to data from different centers, primarily due to variations in staining protocols, tissue preparation techniques, and scanning instruments. To address this issue, additional training using generative adversarial networks for color normalization is commonly required.[28–31] Unlike previous networks that relied on models pre-trained on ImageNet,[32] our study employed a foundation model extensively pre-trained on a large-scale dataset of pathological images. This pre-training enabled our model to inherently handle color variations, thereby enhancing its generalization ability across datasets from different centers and eliminating the need for additional color normalization models.

In the field of natural images, studies have shown that textual information can significantly enhance the performance of image-based models.[33–37] While similar attempts have been made in pathology, they have often focused on establishing foundation models[15,16,38–40] using large datasets or extracting insights from diagnostic reports corresponding to the WSIs,[23,41] which often require paired data. Our study takes a different approach by leveraging the diagnostic expertise of pathologists, a resource that remains underutilized. We encapsulate this expertise into cancer-specific descriptors, treating them as prototypes that remain consistent across data from different centers. By integrating a text-guided feature generation module, we align slide-level features more closely with the prototypes, thereby mitigating the influence of color variations. This approach yielded highly satisfactory results, achieving an AUROC of 0.9974 on an internal dataset, and AUROCs of 0.9724 and 0.9513 on two external datasets. Given the model's strong performance and generalization across both biopsy and surgically resected samples, it shows considerable promise for clinical translation. The model can generate diagnostic results with confidence values, allowing pathologists to re-examine only low-confidence cases, or to use the model as an early screening tool in clinical settings.

In conclusion, our study presents a novel model for distinguishing between colorectal NETs and colorectal adenocarcinomas using data from both textual information and WSIs. Given the distinct treatment strategies and prognostic implications associated with these conditions, accurate identification of NETs by our model has the potential to minimize the necessity for supplementary immunohistochemical tests, thereby optimizing the diagnostic workflow. This integration not only enhances diagnostic efficiency but also has the potential to reduce healthcare costs. Our findings contribute to the application of artificial intelligence in clinical settings and open new avenues for further research in this field.

Limitations of the study

Despite these promising results, several limitations of this study should be acknowledged. First, the deep learning model was trained and validated retrospectively, highlighting the need for rigorous and prospective clinical studies to provide more reliable and conclusive evidence. Second, while our approach attempts to select diagnostically relevant patches, it has inherent limitations. Future investigations could explore constructing multi-scale representations by extracting key patches at various magnifications, potentially enhancing the model's ability to capture important features across different scales. Additionally, identifying the optimal textual information that could further improve patch selection remains an open question. Determining specific textual information that complements pathological images offers a promising avenue for further model optimization. Although we enhanced the model's generalization ability by incorporating the pre-trained model and the diagnostic-related patch selection module, further enhancement could be made by including multi-center data during training and developing more robust feature extractors in future research. Moreover, the concept of our method can be

Figure 4. Visualization results using UMAP

(A–C) Results were obtained from Internal-FAH, External-CC, and External-TCIH datasets, respectively. The left-side figures show results after random sampling from all patches, while the right-side figures display results after random sampling from patches selected using the proposed method. The features selected by the proposed method demonstrated greater separability, which enhanced model training.

extended to tasks involving the discrimination of morphologically distinct cancer types. The diagnostic-related patch selection module could be deployed as a plug-and-play component in subsequent models by simply providing relevant category descriptions. This data-effective algorithm holds great potential for practical deployment, thereby enhancing the applicability of AI models in clinical practice.

RESOURCE AVAILABILITY

Lead contact

Requests for further information on software and resources should be directed and will be fulfilled by the lead contact, Muyan Cai (caimy@sysucc.org.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The source code for our model is publicly accessible at https://github.com/LexieK7/wsi_text. Due to patient privacy obligations and institutional regulations, access to the WSIs and annotation data of both internal and external datasets used in this study is restricted. These datasets were obtained with institutional permissions via Institutional Review Board approval and are therefore not publicly available. However, for non-commercial and academic purposes, interested parties may request access to the data supporting the findings of this study directly from the corresponding author. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This work was supported by grants from the National Natural Science Foundation of China (grant no. 62071502, 82172646, 82073189, 82202905, and 82373190), the Science and Technology Planning Project of Guangdong Province (no. 2020B1212060023), and Guangdong Excellent Youth Team Program (grant no. 2023B1515040025).

AUTHOR CONTRIBUTIONS

M.C., D.X., R.W., Y.S., and N.Z. conceived and designed the study. K.Z., J.D., Y.L., H.H., and S.L. collected the samples and acquired the image data. X.Z., J.D., H.H., Y.L., and Z.Z. provided the clinical and pathological data of multiple medical centers. K.Z., R.W., H.C., Y.Z., and B.J. performed the machine learning. M.C., N.Z., Y.S., R.W., D.X., Y.Z., Z.Z., Y.L., and X.Z. conducted the reader study. K.Z. and H.H. did the statistical analyses. All authors vouch for the data, analyses, and interpretations. K.Z., J.D., R.W., H.C., X.Z., and M.C. wrote the first draft of the manuscript, and all authors reviewed, contributed to, and approved the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrm.2024.101785.

Received: June 10, 2024
Revised: August 19, 2024
Accepted: September 19, 2024
Published: October 15, 2024

REFERENCES

1. Dasari, A., Shen, C., Halperin, D., Zhao, B., Zhou, S., Xu, Y., Shih, T., and Yao, J.C. (2017). Trends in the Incidence, Prevalence, and Survival Outcomes in Patients With Neuroendocrine Tumors in the United States. JAMA Oncol. *3*, 1335–1342. https://doi.org/10.1001/jamaoncol.2017.0589.
2. Kooyker, A.I., Verbeek, W.H., van den Berg, J.G., Tesselaar, M.E., and van Leerdam, M.E. (2020). Change in incidence, characteristics and management of colorectal neuroendocrine tumours in the Netherlands in the last decade. United European Gastroenterol. J. *8*, 59–67. https://doi.org/10.1177/2050640619865113.
3. White, B.E., Rous, B., Chandrakumaran, K., Wong, K., Bouvier, C., Van Hemelrijck, M., George, G., Russell, B., Srirajaskanthan, R., and Ramage, J.K. (2022). Incidence and survival of neuroendocrine neoplasia in England 1995-2018: A retrospective, population-based study. Lancet Reg. Health. Eur. *23*, 100510. https://doi.org/10.1016/j.lanepe.2022.100510.
4. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA A Cancer J. Clin. *71*, 209–249. https://doi.org/10.3322/caac.21660.
5. Ni, S.J., Sheng, W.Q., and Du, X. (2010). Pathologic research update of colorectal neuroendocrine tumors. World J. Gastroenterol. *16*, 1713–1719. https://doi.org/10.3748/wjg.v16.i14.1713.
6. Gallo, C., Rossi, R.E., Cavalcoli, F., Barbaro, F., Boškoski, I., Invernizzi, P., and Massironi, S. (2022). Rectal neuroendocrine tumors: Current advances in management, treatment, and surveillance. World J. Gastroenterol. *28*, 1123–1138. https://doi.org/10.3748/wjg.v28.i11.1123.
7. Fleming, M., Ravula, S., Tatishchev, S.F., and Wang, H.L. (2012). Colorectal carcinoma: Pathologic aspects. J. Gastrointest. Oncol. *3*, 153–173. https://doi.org/10.3978/j.issn.2078-6891.2012.030.
8. Sirinukunwattana, K., Domingo, E., Richman, S.D., Redmond, K.L., Blake, A., Verrill, C., Leedham, S.J., Chatzipli, A., Hardy, C., Whalley, C.M., et al. (2021). Image-based consensus molecular subtype (imCMS) classification of colorectal cancer using deep learning. Gut *70*, 544–554. https://doi.org/10.1136/gutjnl-2019-319866.
9. Yu, G., Sun, K., Xu, C., Shi, X.H., Wu, C., Xie, T., Meng, R.Q., Meng, X.H., Wang, K.S., Xiao, H.M., and Deng, H.W. (2021). Accurate recognition of colorectal cancer with semi-supervised deep learning on pathological images. Nat. Commun. *12*, 6311. https://doi.org/10.1038/s41467-021-26643-8.
10. Yamashita, R., Long, J., Longacre, T., Peng, L., Berry, G., Martin, B., Higgins, J., Rubin, D.L., and Shen, J. (2021). Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. Lancet Oncol. *22*, 132–141. https://doi.org/10.1016/s1470-2045(20)30535-0.
11. Wagner, S.J., Reisenbüchler, D., West, N.P., Niehues, J.M., Zhu, J., Foersch, S., Veldhuizen, G.P., Quirke, P., Grabsch, H.I., van den Brandt,

P.A., et al. (2023). Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. Cancer Cell *41*, 1650–1661.e1654. https://doi.org/10.1016/j.ccell.2023.08.002.

12. Niehues, J.M., Quirke, P., West, N.P., Grabsch, H.I., van Treeck, M., Schirris, Y., Veldhuizen, G.P., Hutchins, G.G.A., Richman, S.D., Foersch, S., et al. (2023). Generalizable biomarker prediction from cancer pathology slides with self-supervised deep learning: A retrospective multi-centric study. Cell Rep. Med. *4*, 100980. https://doi.org/10.1016/j.xcrm.2023.100980.
13. Ikezogwo, W.O., Seyfioglu, M.S., Ghezloo, F., Geva, D., Mohammed, F.S., Anand, P.K., Krishna, R., and Shapiro, L.G. (2023). Quilt-1M: One Million Image-Text Pairs for Histopathology. Adv. Neural Inf. Process. Syst. *36*, 37995–38017.
14. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F.K., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al. (2024). Towards a general-purpose foundation model for computational pathology. Nat. Med. *30*, 850–862. https://doi.org/10.1038/s41591-024-02857-3.
15. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., and Zou, J. (2023). A visual–language foundation model for pathology image analysis using medical Twitter. Nat. Med. *29*, 2307–2316. https://doi.org/10.1038/s41591-023-02504-3.
16. Lu, M.Y., Chen, B., Williamson, D.F.K., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al. (2024). A visual-language foundation model for computational pathology. Nat. Med. *30*, 863–874. https://doi.org/10.1038/s41591-024-02856-4.
17. Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., and Tsirigos, A. (2018). Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat. Med. *24*, 1559–1567. https://doi.org/10.1038/s41591-018-0177-5.
18. Li, B., Li, Y., and Eliceiri, K.W. (2021). Dual-stream Multiple Instance Learning Network for Whole Slide Image Classification with Self-supervised Contrastive Learning. In Conference on Computer Vision and Pattern Recognition Workshops. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Workshops 2021, pp. 14318–14328. https://doi.org/10.1109/cvpr46437.2021.01409.
19. Naik, N., Madani, A., Esteva, A., Keskar, N.S., Press, M.F., Ruderman, D., Agus, D.B., and Socher, R. (2020). Deep learning-enabled breast cancer hormonal receptor status determination from base-level H&E stains. Nat. Commun. *11*, 5727. https://doi.org/10.1038/s41467-020-19334-3.
20. Noorbakhsh, J., Farahmand, S., Foroughi pour, A., Namburi, S., Caruana, D., Rimm, D., Soltanieh-ha, M., Zarringhalam, K., and Chuang, J.H. (2020). Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images. Nat. Commun. *11*, 6367. https://doi.org/10.1038/s41467-020-20030-5.
21. Wang, X., Chen, Y., Gao, Y., Zhang, H., Guan, Z., Dong, Z., Zheng, Y., Jiang, J., Yang, H., Wang, L., et al. (2021). Predicting gastric cancer outcome from resected lymph node histopathology images using deep learning. Nat. Commun. *12*, 1637. https://doi.org/10.1038/s41467-021-21674-7.
22. Zhang, Y., Gao, J., Zhou, M., Wang, X., Qiao, Y., Zhang, S., and Wang, D.J.A. (2023). Text-guided Foundation Model Adaptation for Pathological Image Classification. Preprint at arXiv. https://doi.org/10.48550/arXiv.2307.14901.
23. Zhang, Z., Chen, P., Sapkota, M., and Yang, L. (2017). TandemNet: Distilling Knowledge from Medical Images Using Diagnostic Reports as Optional Semantic References. In Medical Image Computing and Computer Assisted Intervention (Springer International Publishing), pp. 320–328.
24. Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Chen, R.J., Barbieri, M., and Mahmood, F. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. Nat. Biomed. Eng. *5*, 555–570. https://doi.org/10.1038/s41551-020-00682-w.
25. Ilse, M., Tomczak, J.M., and Welling, M. (2018). Attention-based Deep Multiple Instance Learning. Preprint at arXiv. https://doi.org/10.48550/arXiv.1802.04712.
26. Chikontwe, P., Nam, S.J., Go, H., Kim, M., Sung, H.J., and Park, S.H. (2022). Feature Re-calibration Based Multiple Instance Learning for Whole Slide Image Classification. In Medical Image Computing and Computer Assisted Intervention (Springer Nature Switzerland), pp. 420–430.
27. McInnes, L., and Healy, J.J.A. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at arXiv. https://doi.org/10.48550/arXiv.1802.03426.
28. Altini, N., Marvulli, T.M., Zito, F.A., Caputo, M., Tommasi, S., Azzariti, A., Brunetti, A., Prencipe, B., Mattioli, E., De Summa, S., and Bevilacqua, V. (2023). The role of unpaired image-to-image translation for stain color normalization in colorectal cancer histology classification. Comput. Methods Progr. Biomed. *234*, 107511. https://doi.org/10.1016/j.cmpb.2023.107511.
29. Barua, B., Bora, K., Kr Das, A., Ahmed, G.N., and Rahman, T. (2023). Stain color translation of multi-domain OSCC histopathology images using attention gated cGAN. Comput. Med. Imag. Graph. *106*, 102202. https://doi.org/10.1016/j.compmedimag.2023.102202.
30. Lahiani, A., Klaman, I., Navab, N., Albarqouni, S., and Klaiman, E. (2021). Seamless Virtual Whole Slide Image Synthesis and Validation Using Perceptual Embedding Consistency. IEEE J. Biomed. Health Inform. *25*, 403–411. https://doi.org/10.1109/jbhi.2020.2975151.
31. Moghadam, A.Z., Azarnoush, H., Seyyedsalehi, S.A., and Havaei, M. (2022). Stain transfer using Generative Adversarial Networks and disentangled features. Comput. Biol. Med. *142*, 105219. https://doi.org/10.1016/j.compbiomed.2022.105219.
32. Deng, J., Dong, W., Socher, R., Li, L.J., Kai, L., and Li, F.-F. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255.
33. Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. (2022). Flamingo: a visual language model for few-shot learning. Preprint at arXiv. https://doi.org/10.48550/arXiv.2204.14198.
34. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Muyan, Z., Zhang, Q., Zhu, X., and Lu, L.J.a.p.a. (2023). Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. Preprint at arXiv. https://doi.org/10.48550/arXiv.2312.14238.
35. Li, J., Li, D., Xiong, C., and Hoi, S.C.H. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2201.12086.
36. Li, J., Selvaraju, R.R., Gotmare, A.D., Joty, S.R., Xiong, C., and Hoi, S.C.H. (2021). Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2107.07651.
37. Liu, H., Li, C., Wu, Q., and Lee, Y. (2024). Visual instruction tuning. Preprint at arXiv. https://doi.org/10.48550/arXiv.2304.08485.
38. Kang, M., Song, H., Park, S., Yoo, D., and Pereira, S. (2023). Benchmarking self-supervised learning on diverse pathology datasets. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3344–3354.
39. Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Liu, S., Mathieu, P., Eck, A.v., Lee, D., Viret, J., et al. (2023). Virchow: A Million-Slide Digital Pathology Foundation Model. Preprint at arXiv. https://doi.org/10.48550/arXiv.2309.07778.
40. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al. (2024). A whole-slide foundation model for digital pathology from real-world data. Nature *630*, 181–188. https://doi.org/10.1038/s41586-024-07441-w.

41. Li, H., Chen, Y., Chen, Y., Yang, W., Ding, B., Han, Y., Wang, L., and Yu, R.J.a.p.a. (2024). Generalizable Whole Slide Image Classification with Fine-Grained Visual-Semantic Interaction. Preprint at arXiv. https://doi.org/10.48550/arXiv.2402.19326.

42. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. (2022). LAION-5B: An open large-scale dataset for training next generation image-text models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2210.08402.

43. Kingma, D.P., and Ba, J.J.a.e.-p. (2014). Adam: A Method for Stochastic Optimization. Preprint at arXiv. https://doi.org/10.48550/arXiv.1412.6980.

CellPress
OPEN ACCESS

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
Our source code	GitHub: https://github.com/LexieK7/wsi_text	N/A
Pretrained encoder	GitHub: https://github.com/wisdomikezogwo/quilt1m	https://doi.org/10.1145/3489517.3530589
Other		
GPU GeForce RTX 2080 Ti	Nvidia Corp., Santa Clara, California.	N/A
CPU Xeon(R) Gold 6240	Intel Corp., Santa Clara, California.	N/A

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Study participants

To conduct model training and validation, we utilized datasets obtained from three medical centers: The First Affiliated Hospital (FAH) of Sun Yat-sen University, the Sun Yat-Sen University Cancer Center (CC), and the Tianjin Medical University Cancer Institute & Hospital (TCIH) (Figure 1A). The Internal-FAH dataset included slides collected from September 2007 to June 2021 (550 patients). The External-CC dataset comprised data from February 2014 to August 2021 (837 patients). The External-TCIH dataset consisted of data from May 2014 to January 2018 (412 patients). Finally, the Biopsy-CC dataset contained data from September 2015 to October 2023 (423 patients). H&E-stained tumor slides in the Internal-FAH, External-CC, and External-TCIH datasets were scanned using the Kfbio KF-PRO-020 scanner, Aperio AT2 scanner, and PHILIPS Ultra Fast Scanner, respectively. Similarly, slides in the Biopsy-CC dataset were scanned using the PHILIPS Ultra Fast Scanner. All patient-related information obtained was ethically approved by the Institutional Ethics Committee (B2021-435-01), and informed consent was waived for this retrospective study.

METHOD DETAILS

Data preprocessing

We followed the method in CLAM[24] to transform the downsampled WSI from RGB to the HSV color space. Initial tissue regions were obtained based on the threshold of the saturation channel, and the small gaps and holes were filled using morphological closing. These initial regions were further filtered based on an area threshold to obtain the final segmentation mask. After segmentation, the foreground regions (containing both tumor and normal tissue) were divided into patches of size 224 × 224 pixels at 10× magnification. Given that the pre-trained image encoder has been trained on a substantial volume of unprocessed data, we believe that it could extract color-agnostic features, thereby obviating the need for supplementary patch processing.

Pre-trained text and image encoders

Ikezogwo et al.[13] employed educational videos sourced from expert pathologists available on YouTube. Key frames were extracted from these videos, followed by conversion of the corresponding speech into text. Subsequently, de-noise techniques were applied to obtain the final set of image-text pairs. This process led to the construction of QUILT, a dataset encompassing 419,780 images aligned with 768,826 text pairs. To further enhance the training data, QUILT-1M was generated by amalgamating data derived from open-source articles in PubMed, pathology images in LAION-5B,[42] and data extracted from Twitter.[15] This augmented dataset was then utilized for the pre-training of the model.

QUILTNET uses the Contrastive Language-Image Pre-training (CLIP) objective to complete the pre-training on QUILT-1M, aiming to enable the model to learn the matching relationship between text and image pairs. QUILTNET consists of two modules: text encoder and image encoder. The image encoder utilizes the ViT-B/16 architecture, while the text encoder employs GPT-2 with a context length of 77. During the training process, for a batch containing N image-text pairs, the model predicts N^2 image-text similarity, which is calculated as the cosine similarity of text embeddings and image embeddings, i.e., the matrix shown in. There are N pairs of positive samples in this matrix, where the text is paired with the image, and the similarity is represented by the diagonal element of the matrix. The remaining $(N^2 - N)$ pairs of samples are considered as negative samples. Therefore, the training objective is to maximize the similarity of the positive samples while minimizing the similarity of the negative samples. Ultimately, the

objective is expressed as:

$$\mathcal{L} = -\frac{1}{2N}\left(\sum_{i=1}^{N}\log\frac{e^{\cos(I_i,T_i)}}{\sum_{j=1}^{N}e^{\cos(I_i,T_j)}}+\sum_{i=1}^{N}\log\frac{e^{\cos(I_i,T_i)}}{\sum_{j=1}^{N}e^{\cos(I_j,T_i)}}\right)$$

where I_i and T_i denote the embeddings of the i-th image and text, respectively. Upon completion of the pre-training phase, the parameters of both the image encoder and text encoder are frozen, and no subsequent fine-tuning is conducted.

Selection of diagnostic-related patches

Whole slide images often contain extensive regions of normal tissue, which we hypothesized may have weak correlations with the diagnostic process, potentially impacting the model's accuracy. Furthermore, capturing tens of thousands of patches at a 10× resolution from the entire slide imposed a substantial computational burden on subsequent modules, especially the attention computation module. As a result, we proposed a diagnostic-related patches selection module based on the similarity to extract the top k% patches that exhibited the highest relevance to the morphological characteristics of the NET and CRC.

Experienced pathologists provided descriptions of NET and CRC by reviewing relevant literature[7] and combining insights from their clinical experience to summarize the key morphological features, as depicted in Table S9. The corresponding textual descriptions of CRC and NET were designated as T_C and T_E, respectively. Initially, the tokenizer pre-processes T_C and T_E, and mapping them to corresponding embedding vectors. These text embedding vectors were then input into the pre-trained text encoder, yielding the respective text features, $\boldsymbol{f}_C$ and $\boldsymbol{f}_E$. Given that the image encoder and text encoder were pre-trained using large-scale text-image pairs, patches exhibiting characteristics aligned with the morphological text descriptions would exhibit high similarity with the aforementioned text features in the feature space. All patches, denoted as $I = \{I_1, \ldots I_N\}$, from a given slide were fed into the image encoder, generating the feature set $F = \{\boldsymbol{f}_1, \ldots \boldsymbol{f}_N\}$. Subsequently, the cosine similarity between the text features and image features was computed, and the top k% of patches with the highest similarity were selected to represent the whole slide, denoted as:

$$P = top\_K\{\cos(\boldsymbol{f}_i, \boldsymbol{f}_j)\}$$

where top_K denotes the operation of selecting the patch with the highest similarity, $\boldsymbol{f}_i \in F, \boldsymbol{f}_j \in \{\boldsymbol{f}_C, \boldsymbol{f}_E\}$, P is the set of obtained patches. Note that all subsequent computations were exclusively performed on this selected set of patches and not on the entirety of patches.

Attention-based aggregation module

The module was designed to aggregate patch features using attention scores and derive slide-level features. Due to the differing regions of interest for each cancer type, two sets of attention scores were calculated separately. This distinction was crucial due to the varied morphological characteristics exhibited by the categories, necessitating the prediction of class-specific attention scores. To facilitate this, we defined two shared fully connected layers with weights $U \in \mathbb{R}^{256\times512}$ and $V \in \mathbb{R}^{256\times512}$ for all classes. Subsequently, we established two parallel attention branches, corresponding to the number of classes, with each branch associated with parameters $W_p \in \mathbb{R}^{1\times256}$, with $i \in \{1,2\}$. Accordingly, the attention score of the q-th patch for the p-th class denoted $a_{q,p}$, is determined as:

$$a_{q,p} = \frac{e^{\left\{W_p\left(\tanh V\boldsymbol{f}_q^T\right)\odot simg\left(U\boldsymbol{f}_q^T\right)\right\}}}{\sum_{j=1}^{N}e^{\left\{W_p\left(\tanh V\boldsymbol{f}_j^T\right)\odot simg\left(U\boldsymbol{f}_j^T\right)\right\}}}$$

Consequently, category-specific slide-level features $\boldsymbol{f}_{slide,p}$ were acquired by aggregating patch features using category-specific attention scores. The slide-level feature for the p-th class was calculated as:

$$\boldsymbol{f}_{slide,p} = \sum_{k=1}^{N} a_{k,p}\boldsymbol{f}_k$$

Text-guided slide-level feature generation

Data obtained from different centers exhibited significant variations in staining due to differences in staining reagents, tissue section thickness, staining conditions, and scanner models (Figure S6). Since we did not perform color normalization on individual patches, models trained solely on internal center data may exhibit poor performance when applied to external data. Despite the visual dissimilarities observed among slides from different centers, their color-independent morphological features demonstrated commonality. By disregarding the color variability, slides belonging to the same class could still be described using identical descriptors. Consequently, we regarded text features as prototypes and utilized them to guide the generation of more robust slide-level features.

The structure and performance of the module are detailed in Figure S7 and Table S10. Specifically, the text features were initially mapped to the same feature space as the slide-level features using a fully connected layer parameterized by $W_{map} \in \mathbb{R}^{256\times512}$.

Subsequently, the mapped text features were integrated with the image features through fusion to obtain the final slide-level feature representation. Two fusion approaches, ADD-Fusion and FC-Fusion were developed for the fusion. In the ADD-Fusion approach, the computation of final slide-level features f_{final} was as follows:

$$f_{final} = f_{slide} + W_{map} f_{text}^{T}$$

where $f_{slide} = [\boldsymbol{f}_{slide,1}, \boldsymbol{f}_{slide,2}]$, $f_{text} = [\boldsymbol{f}_{C}, \boldsymbol{f}_{E}]$, $f_{final} = [\boldsymbol{f}_{final,1}, \boldsymbol{f}_{final,2}]$, And the number of categories is 2 due to the current setting of a binary classification problem.

Conversely, in the FC-Fusion approach, a different procedure is employed for the computation of final slide-level features:

$$f_{final} = W_{f}\left[f_{slide}, W_{map} f_{text}^{T}\right]^{T}$$

where $[\cdot, \cdot]$ denotes the concatenation operation and W_f is the parameter of the fusion module. Finally, the unnormalized slide-level score $\boldsymbol{s}_{slide,p}$ was computed through the classification layer $W_{c,p} \in \mathbb{R}^{1\times 256}$ to obtain:

$$\boldsymbol{s}_{slide,p} = W_{c,p} \boldsymbol{f}_{final,p}^{T}$$

where $p \in \{1,2\}$. For inference, a softmax function is applied to the unnormalized slide-level score to obtain the final predicted probability distribution.

QUANTIFICATION AND STATISTICAL ANALYSIS

Experimental setup and implementation details

We conducted all experiments using 10-fold cross-validation. In this cross-validation variant, internal validation and test sets were split off from the full dataset at the patient level, leaving 8-folds for training, 1-fold for validating, and the remaining 1-fold for testing. To maintain an even distribution of categories among the three datasets, we conducted proportional random sampling from each class to generate data for each respective set. During training, the validation set was used to determine when to stop model training. For external datasets, the models were tested using complete datasets. The models were trained with the Adam[43] optimizer with an L2 weight decay of 1e−5 and a learning rate of 2e−4. All models were trained with a batch size of 1, for a minimum of 50 epochs, extending up to a maximum of 200 epochs if the early stopping criterion was not met. The early stopping criterion was defined as halting the training if the validation loss did not decrease within 20 consecutive epochs. For training time calculation, we iterated the training process 10 times to derive a more dependable average time. Regarding inference time, we computed the inference time across all dataset samples and divided it by the total sample count to determine the average inference time per sample. Note that the encoder employed for patch-level feature extraction is pretrained and consistent across all models, hence not factored into the time evaluation. Segmentation and patching of WSIs were performed on Intel(R) Xeon(R) Gold 6240 Central Processing Units (CPUs), and the models were trained on NVIDIA GeForce RTX 2080 Ti Graphics Processing Units (GPUs).

Quantification and statistical analysis

Our main evaluation metric was the AUROC. Additionally, we employed the F1 score, accuracy, sensitivity, and specificity (with a classification threshold of 0.5) for performance evaluation. For each experiment, we reported the mean and standard deviation of the model's internal and external test performances. To ensure data independence, we selected only one slide per patient, guaranteeing that each patient appeared in only one set for training or validation. The external test sets from different centers, allowing for a more comprehensive assessment of the generalization properties of our algorithms. A p value less than 0.05 was considered statistically significant. Data pre-processing and model development were conducted using Python (version 3.7.0) and the deep learning platform PyTorch (version 1.10).