npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-02210-z

Deep multimodal fusion of patho-radiomic and clinical data for enhanced survival prediction for colorectal cancer patients

Run Shi[1,5], Jing Sun[2,5], Zhaokai Zhou[3,5], Qiang Su[4] & Yongqian Shu[1] ✉

This study introduces PRISM-CRC, a novel deep learning framework designed to improve the diagnosis and prognosis of colorectal cancer (CRC) by integrating histopathology, radiology, endoscopy and clinical data. The model demonstrated high accuracy, achieving a concordance index of 0.82 for predicting 5-year disease-free survival and an AUC of 0.91 for identifying microsatellite instability (MSI) status. A key finding is the synergistic power of this multimodal approach, which significantly outperformed models using only a single data type. The PRISM-CRC risk score proved to be a strong, independent predictor of survival, offering more granular risk stratification than the traditional TNM staging system. This capability has direct clinical implications for personalizing treatment, such as identifying high-risk Stage II patients who might benefit from adjuvant chemotherapy. The study acknowledges limitations, including a modest performance decrease due to "domain shift" and classification errors in morphologically ambiguous cases, highlighting the need for future prospective trials to validate its clinical utility.

Colorectal cancer (CRC) is a major global health challenge. As of 2020, it accounted for over 1.9 million new cases and 935,000 deaths worldwide[1], ranking as the third most diagnosed malignancy and the second leading cause of cancer-related mortality globally. This high disease burden is compounded by persistent diagnostic and therapeutic challenges. Early detection through colonoscopy significantly improves outcomes, but conventional manual polyp identification is labor-intensive, prone to error, and often misses subtle lesions[2]. Likewise, pathological assessment of resected tumors and molecular testing for biomarkers (e.g., microsatellite instability (MSI), KRAS, BRAF, POLE mutations) are time-consuming and costly, limiting their widespread use. Consequently, some actionable information, such as identifying patients who could benefit from targeted therapies or immunotherapy, may be overlooked in routine practice (for example, rare hypermutated cases like POLE-mutant tumors are not routinely screened and often go undiagnosed[3]. These gaps underscore the need for innovative approaches to augment clinical decision-making[4].

Recent advances in artificial intelligence (AI), particularly deep learning, offer transformative potential for CRC management. AI techniques (machine learning and deep learning) can process vast medical data, uncover hidden patterns, and improve tumor detection, classification, and outcome prediction[5]. Deep learning models have achieved human-level performance in medical image analysis and are rapidly being adopted in oncology research[6]. In CRC, deep learning systems can assist at multiple points in the care continuum: automating polyp and tumor detection in endoscopic images and digital histopathology, predicting molecular phenotypes directly from routine scans, and stratifying patients by recurrence risk or likely treatment response. For instance, AI-driven polyp detection during colonoscopy has been shown to significantly reduce missed lesions, improving real-time identification of subtle or flat polyps. Similarly, deep learning models trained on hematoxylin-and-eosin (H&E) whole-slide images have learned to predict tumor biomarkers such as MSI status and specific gene mutations from morphology alone[7]. Dozens of studies since 2019 confirm that MSI can be predicted from H&E slides, leading to an AI-based MSI test approved in 2022 that provides rapid, cost-effective pre-screening at high sensitivity[8]. Integrating pathology, radiology, genomics, and clinical data through deep learning enables a comprehensive approach to patient disease understanding, advancing precision oncology[9,10].

We present PRISM-CRC (Patho-Radiomic Integrative Survival Model), a novel multimodal deep learning framework designed to address key challenges in CRC diagnosis and prognosis. Our approach combines histopathology image analysis with other data modalities to improve predictive accuracy, aligning with the emerging emphasis on multimodal data

[1]Department of Oncology, The First Affiliated Hospital of Nanjing Medical University, Nanjing, Jiangsu, China. [2]Department of Endocrinology, Jiangsu Province Hospital of Chinese Medicine, Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, Jiangsu, China. [3]Department of Urology, The Second Xiangya Hospital of Central South University, Changsha, Hunan, China. [4]First Clinical Medical College, Guizhou University of Traditional Chinese Medicine, Guiyang, Guizhou, China. [5]These authors contributed equally: Run Shi, Jing Sun, Zhaokai Zhou. ✉e-mail: suqiang@gzy.edu.cn

fusion in cancer care. In the following sections, we provide a comprehensive literature review of recent developments in deep learning for CRC, describe the architecture and methods of PRISM-CRC in detail, and report its performance on multiple clinically relevant tasks. We then discuss the implications for clinical translation, current limitations, and future directions toward large-scale deployment in colorectal oncology.

Deep learning in CRC pathology and imaging

Transformer-based deep learning models have rapidly advanced the analysis of pathology and radiology images in CRC. Vision transformers, which apply self-attention mechanisms to model long-range dependencies, have shown particular promise in extracting complex tissue patterns that correlate with genetic and clinical features. Singh et al. introduced KRASFormer[7], a fully transformer-based pipeline for predicting KRAS mutations from H&E whole-slide images. KRASFormer employs a two-stage approach: first isolating tumor-rich regions, then applying an XCiT (cross-covariance image transformer) model to classify KRAS status. This architecture captured subtle morphologic differences between KRAS mutant and wild-type tumors, achieving area under the ROC curve (AUC) values of 0.691 on the TCGA-CRC cohort and 0.653 on an external validation set. Likewise, new architectures are addressing the scale of gigapixel whole slide images (WSIs). Li et al.[11] proposed Long-MIL, a local–global hybrid transformer for WSIs that reduces quadratic attention complexity to linear by masking to local patches improving performance, memory usage, and speed on large pathology images. These results demonstrate that transformer attention can learn clinically meaningful histologic patterns predictive of key mutations, matching the performance of earlier convolutional neural network (CNN) methods.

Self-supervised methods have also advanced CRC pathology. Tailored pretext tasks can extract informative features from unlabeled H&E-stained WSIs. For example, a Swin Transformer–based self-supervised framework achieved 96% patch-level classification accuracy on the NCT-CRC-HE-100K dataset[12] using progressive layer-wise distillation and multi-scale augmentation[13]. Similarly, Barlow Twins contrastive learning has been applied to colonoscopy images: PathBT (2024) trained a Barlow-Twins encoder on four polyp classes, showing that its learned features are more separable than those learned under supervision. In short, pathology-specific self-supervised pretraining produces more meaningful representations than generic models[14].

Multimodal data fusion and genomic prediction

There is a clear trend toward fusing multiple data types for CRC. Liu et al.[15] developed M^2 Fusion, a Bayesian multi-level fusion of digitized WSIs and CT scans to predict MSI status. By combining deep features from pathology and radiology with uncertainty-aware weighting, M^2 Fusion achieved an MSI prediction AUC of 0.8177, outperforming both pathology-only (0.7908) and conventional fusion (0.7289) model. This demonstrates that imaging modalities are complementary: CT adds tumor volumetric and context cues to histology.

At the same time, new foundation models are enabling holistic fusion of image and non-image data. Ferran et al. showed that GPT-4V (a multimodal LLM) can classify CRC pathology images purely via in-context learning with descriptive text prompts, reaching performance on par with task-specific CNNs[16]. Editorials highlight that text-based LLMs and vision transformers are converging: both use transformer backbones, making it feasible to build AI systems that jointly ingest medical text and images[17]. These multimodal foundation models promise to discover cross-modal biomarkers by reasoning over heterogeneous inputs. In precision oncology, experts note that integrative models (imaging + genomics + EHR) are becoming feasible and can capture interactions that single-modality models miss.

Weakly-supervised learning and clinical prediction

A notable challenge in CRC AI is the limited availability of pixel-level annotations and the expense of obtaining large expert-labeled datasets. To address this, weakly-supervised learning techniques have been developed that train models on slide-level or patient-level labels. In 2024, El Nahhas et al. proposed a weakly supervised, transformer-based multiple-instance *regression* model that estimates tumor-infiltrating lymphocyte (TIL) density per high-power field directly from H&E stained WSIs using only case-level labels, thereby obviating cell-level annotations; attention-based pooling over patch features enabled accurate TIL quantification on internal and external CRC cohorts[18]. Consistent with this direction, a peer-reviewed regression framework further showed that weakly supervised, SSL-initialized attMIL models can predict immune-cell biomarkers (including TIL-related signals) from H&E WSIs with improved correspondence to clinically relevant regions and enhanced prognostic stratification in a large CRC cohort[18].

Advances in detection and segmentation (Endoscopy) and treatment response

Deep learning has also transformed CRC endoscopy analysis. Real-time polyp detection has been markedly improved by modern one-stage detectors. For instance, an AI-based system built upon YOLOv8 reported high per-frame precision (95.6%), recall (91.7%), and F1-score (92.4%) across multiple colonoscopy datasets, illustrating strong accuracy–throughput trade-offs for clinical use[19]. Building on this trend, Wan et al. proposed a semantic feature–enhanced YOLOv5 with novel P-C3 and contextual feature augmentation modules; the detector increased recall and reduced missed polyps relative to standard baselines[20]. Beyond bounding boxes, instance segmentation networks have advanced toward real-time clinical utility: a YOLACT-ResNet50 variant (RTPoDeMo) achieved 72.3 mAP at 32.8 FPS and 99.59% per-image accuracy on prospectively recorded colonoscopy videos, missing only 1 of 36 expert-confirmed polyps[21].

Transformer-style designs are also effective for polyp segmentation. Liu et al. introduced NA-SegFormer, a multi-level encoder–decoder with neighborhood attention, reaching Dice scores up to 94.30% on Kvasir-SEG with >125 FPS, and demonstrating favorable speed accuracy trade-offs[22]. Similarly, the CTHP hybrid CNN–Transformer replaces global self-attention with axial attention and adds an information propagation module, yielding state-of-the-art (SOTA) results and robust cross-domain generalization on multiple polyp benchmarks[22,23]. Foundation segmentation models have also been adapted: an efficient SAM tuning strategy (PSF-SAM) freezes most parameters and optimizes a small subset to mitigate catastrophic forgetting, improving mDice and mIoU on Kvasir-SEG and CVC-ClinicDB in few-shot settings[24].

Beyond detection and segmentation, AI outputs are increasingly linked to treatment decision-making. Histology-trained transformers for MSI detection can incidentally flag POLE-mutant CRCs: a dual-detection study showed that a model trained on MSI labels still identified over 75% of pathogenic POLE-mutant cases in external resection cohorts, supporting AI-assisted prescreening for immunotherapy-eligible subgroups[3].

Results

Overall predictive performance of the multimodal model

The primary objective was to develop a model for predicting 5-year disease-free survival (DFS). The fully integrated multimodal deep learning model, which jointly processes features from histopathology WSIs, preoperative computed tomography (CT) scans, and structured electronic health record (EHR) data, demonstrated a high level of prognostic accuracy on the validation set.

For the primary endpoint of 5-year DFS prediction, the model achieved a concordance index (C-index) of 0.82 (95% confidence interval [CI]: 0.78–0.85). The model's discriminative ability over time was further assessed using time-dependent area under the receiver operating characteristic curve (AUC) analysis, yielding an AUC of 0.85 (95% CI: 0.81–0.89) for predicting DFS at 5 years. This level of performance is within the range considered clinically meaningful for prognostic stratification and significantly exceeds that of models based on traditional clinicopathological staging systems alone.

For the primary endpoint of 5-year DFS, we assessed absolute risk accuracy using the time-dependent Brier score and evaluated agreement

Table 1 | Model performance on EBHI-SEG and Kvasir-SEG

		EBHI-SEG				Kvasir-SEG			
Architectural Class	Model	mDice	mIoU	Precision	Recall	mDice	mIoU	Precision	Recall
Baseline CNN	U-Net[38]	0.831	0.745	0.842	0.879	0.855	0.778	0.861	0.895
Advanced CNN	PraNetNA-SegForme[39]	0.901	0.848	0.915	0.908	0.913	0.860	0.924	0.919
Hybrid	TransUNet[40]	0.924	0.875	0.931	0.928	0.931	0.883	0.935	0.936
	CTHP[23]	0.939	0.891	0.934	0.941	**0.947**	0.902	**0.952**	0.944
Transformer	NA-SegFormer[22]	0.943	**0.898**	0.935	0.948	0.938	0.890	0.930	0.947
Foundation Model	PSF-SAM[24]	0.939	0.899	0.923	0.945	0.942	**0.904**	0.928	0.941
Ours	PRISM-CRC	**0.946**	0.889	**0.938**	**0.950**	0.945	0.900	0.941	**0.952**

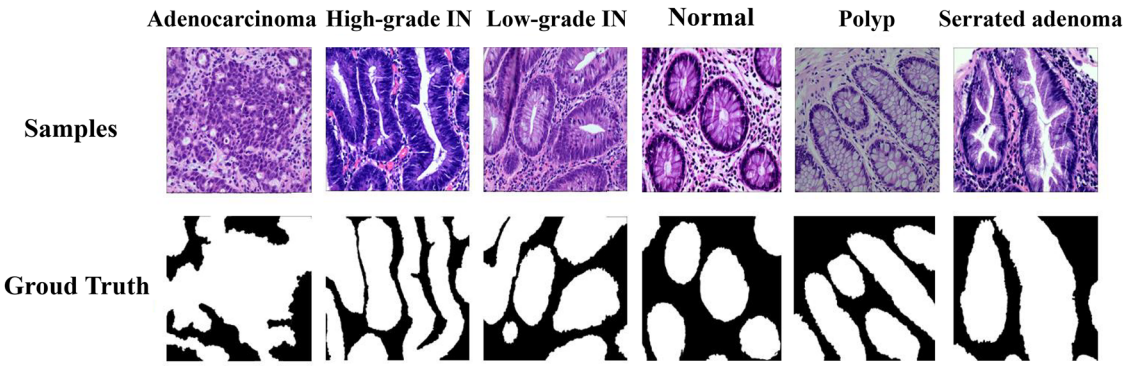

Fig. 1 | Performance of PRISM-CRC model segmentation across colorectal histo-pathology subtypes.

between predicted and observed outcomes via a calibration analysis. In the TCGA-COAD validation set, the 5-year Brier score was 0.161, and the calibration slope was 0.98 (95% CI includes 1), indicating close to ideal calibration.

In addition to survival prognostication, the framework was evaluated on the task of predicting MSI status directly from the multimodal data inputs. For this binary classification task, the model achieved an AUC of 0.91 (95% CI: 0.88–0.94), with an accuracy of 88.5%, a sensitivity of 85.7%, and a specificity of 91.3%. This performance indicates a strong capacity to identify this critical biomarker, which is essential for guiding immunotherapy decisions.

Segmentation performance analysis

A detailed summary of model performance on the EBHI-SEG and Kvasir-SEG test sets is provided in Table 1. Grouping the models by architectural class enables an analysis of their evolutionary progression. Figures 1 and 2 shows the segmentation effect of our model in several cases.

CRC risk prediction across horizons

To benchmark PRISM-CRC against SOTA methods, we evaluated performance on the TCGA-COAD validation cohort. Table 2 reports time-dependent AUCs (1–5 years), the concordance index (C-index), and Weighted Mean Absolute Error (WMAE; lower is better) for six baselines spanning Graph Neural Networks, Spatiotemporal GNNs, Deep Neural Networks, and Multimodal Hypergraph architectures, together with our method. Among the baselines, Risk-Net and MRePath are strongest: the best baseline AUCs are 0.81 (1 year, Risk-Net), 0.79 (2 year, MRePath), 0.76 (3 year, MRePath), 0.75 (4 year, Risk-Net), and 0.76 (5 year, Risk-Net). Risk-Net attains the highest baseline C-index (0.74), while MRePath yields the lowest baseline WMAE (1.64). Pure Graph Neural Network baselines (DM-GNN, SAGL) show lower AUCs (≤0.75 at 1 year) and significantly higher WMAE (≥2.80).

Our method achieves the highest performance across all horizons and summary metrics: AUCs of 0.85/0.89/0.84/0.79/0.80 at 1–5 year, a C-index of 0.82, and a WMAE of 1.40. Relative to the strongest baseline at each horizon (Risk-Net at 1,4,5, year; MRePath at 2,3, year), the absolute AUC differences are {+0.04, +0.10, +0.08, +0.04, +0.04}. The C-index increases by +0.08 versus Risk-Net (0.82 vs. 0.74), and WMAE is lower by 0.24 compared with MRePath (1.40 vs. 1.64), a 14.6% relative reduction. At the longer horizons (4–5 years), AUCs remain 0.79–0.80, exceeding the strongest baselines (≤0.75–0.76).

Ablation study

To systematically evaluate the contribution of each data modality and validate the hypothesis that data fusion enhances predictive power, a comprehensive ablation study was conducted. The performance of the full multimodal model was compared against unimodal and bimodal baseline models. The results, summarized in Table 3, demonstrate that the integration of complementary information from pathology, radiology, and clinical domains yields a statistically significant improvement in prognostic accuracy.

The model trained solely on structured clinical data (including age, sex, tumor location, and TNM stage) established a strong baseline performance, achieving a C-index of 0.72 (95% CI: 0.68–0.76) for 5-year DFS prediction. This reflects the inherent prognostic value of standard clinical variables. The unimodal deep learning model trained on histopathology WSIs alone achieved a C-index of 0.76 (95% CI: 0.72–0.80), indicating that microscopic tumor morphology captured by the deep learning system contains significant prognostic information beyond standard staging. The radiology-based model, using preoperative CT scans, yielded a C-index of 0.69 (95% CI: 0.64–0.74), demonstrating its utility but also suggesting its information might be less comprehensive than that in histopathology for this specific task.

Combining modalities led to incremental performance gains. A bimodal model integrating pathology WSIs and clinical data improved the C-index to 0.79 (95% CI: 0.75–0.83). This substantial jump from either unimodal baseline suggests a powerful synergy; the model learns to correlate histomorphological patterns with systemic patient characteristics, capturing complementary information that is not redundant.

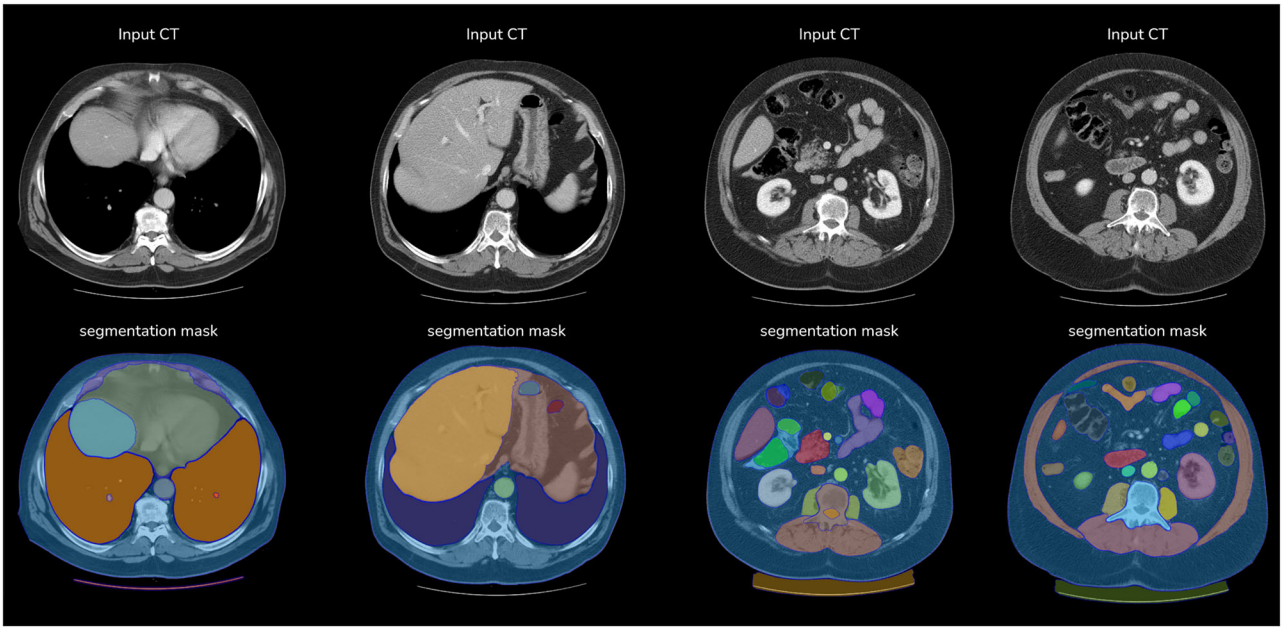

Fig. 2 | Examples of PRISM-CRC model segmentation on diagnostic CT scans.

Table 2 | Performance of various AI models for colorectal cancer (CRC) risk prediction over 1–5 year horizons

Architectural Class	Model	Risk Prediction (AUC)					C-index	WMAE
		1y	2y	3y	4y	5y		
Graph Neural Network	DM-GNN[41]	0.70	0.71	0.69	0.68	0.65	0.67	2.94
Graph Neural Network	SAGL[42]	0.75	0.73	0.70	0.73	0.64	0.69	2.80
Spatiotemporal GNN	STG[43]	0.78	0.65	0.67	0.69	0.71	0.70	2.84
Deep Neural Network	DeepCRC[44]	0.77	0.75	0.70	0.62	0.69	0.71	2.37
Multimodal Hypergraph	MRePath[45]	0.78	0.79	0.76	0.73	0.74	0.72	1.64
AI-Augmented DL	Risk-Net[46]	0.81	0.78	0.74	0.75	0.76	0.74	1.83
Ours		**0.85**	**0.89**	**0.84**	**0.79**	**0.80**	**0.82**	**1.40**

More advanced models show higher time-dependent AUC-ROC and C-index, and lower WMAE. WMAE evaluates survival prediction error, accounting for censoring.

Table 3 | Comparative performance and ablation study for 5-Year disease-free survival prediction on TCGA-READ vlidation cohort

Model Configuration (Input Modalities)	C-index (95% CI)	5-Year AUC (95% CI)	P-value vs. Full Multimodal
Clinical Data Only	0.72 (0.68–0.76)	0.75 (0.71–0.79)	<0.03
Radiology CT Only	0.69 (0.64–0.74)	0.71 (0.66–0.76)	<0.04
Pathology WSI Only	0.76 (0.72–0.80)	0.79 (0.75–0.83)	<0.05
Pathology + Clinical	0.79 (0.75–0.83)	0.82 (0.78–0.86)	<0.05
Full Multimodal Model	**0.82 (0.78–0.85)**	**0.85 (0.81–0.89)**	–

The final, fully integrated trimodal model, which fused pathology, radiology, and clinical data using an attention-based feature-level fusion mechanism, achieved the highest performance with a C-index of 0.82 (95% CI: 0.78–0.85). This represents a statistically significant improvement over the best-performing unimodal (Pathology WSI only, $P < 0.05$) and bimodal (Pathology + Clinical, $P < 0.05$) models. This outcome confirms that each modality contributes unique and non-redundant predictive information, and that the deep learning framework is effective at synthesizing these diverse data streams into a cohesive and powerful prognostic signature.

Robustness and generalizability

To rigorously evaluate the robustness of PRISM-CRC without relying on private or inaccessible data, we employed a three-tier validation strategy using the publicly available The Cancer Genome Atlas (TCGA) datasets. This approach assesses model performance across three increasing levels of distributional shift: homologous validation, aggregated population validation, and cross-domain anatomical validation.

First, the model was evaluated on the held-out test set derived from the TCGA-COAD dataset. Since the model was primarily developed using colon cancer data, this split represents the baseline performance on a homologous data distribution. In this setting, the model achieved a high concordance index (C-index) of 0.82 (95% CI: 0.78–0.85) for 5-year DFS prediction.

To simulate a large-scale, heterogeneous clinical environment containing diverse tumor types, we evaluated the model on a pooled dataset combining both TCGA-COAD and TCGA-READ cohorts. This

aggregated validation set introduces greater biological variance by mixing colon and rectal malignancies. On this combined cohort, the model demonstrated robust performance, achieving a C-index of 0.79 (95% CI: 0.74–0.83). Additionally, for the task of MSI status prediction within this comprehensive population, the model yielded an AUC of 0.87 (95% CI: 0.83–0.91).

Finally, to test the model's ability to handle significant domain shifts, we performed an independent validation solely on the TCGA-READ dataset. The model maintained a competitive C-index of 0.78 (95% CI: 0.72–0.82).

Prognostic value and independent contribution to risk stratification

Beyond standard performance metrics, the clinical utility of a prognostic model lies in its ability to effectively stratify patients into groups with meaningfully different outcomes. The model's predicted risk score was used to classify patients in the validation cohorts into low-risk and high-risk groups, using the median score from the training cohort as the cutoff.

To ascertain whether the model provides novel prognostic information beyond established clinical standards, a multivariable Cox proportional hazards analysis was performed (Table 4). After adjusting for key clinicopathological variables, including patient age, sex, tumor location, and the American Joint Committee on Cancer (AJCC) pathological TNM stage, the model's risk score remained a strong and independent predictor of DFS. The fact that the model's prognostic power persists after accounting for the TNM stage. The current gold standard for CRC prognostication, proves that it captures a biological signal of tumor aggressiveness not encapsulated by traditional staging criteria.

This independent prognostic value has profound clinical implications, particularly for treatment decision-making in intermediate-risk patient groups. For instance, the model was able to effectively re-stratify patients within the clinically ambiguous Stage II category. Stage II patients with a high-risk model score had a survival trajectory comparable to that of low-risk Stage III patients, suggesting they may represent a subgroup that could benefit from adjuvant chemotherapy, a treatment not routinely recommended for all Stage II patients. This ability to refine risk within established clinical stages highlights the model's potential as a powerful tool for personalized medicine.

Table 4 | Multivariable Cox proportional hazards analysis for 5-Year disease-free survival in the combined validation cohorts

Variable	Univariable Analysis HR (95% CI)	Multivariable Analysis P-value
Model Risk Score (High vs. Low)	**3.62 (2.98–4.40)**	<0.04
Age (per 10-year increase)	1.15 (1.08–1.22)	<0.009
Sex (Male vs. Female)	1.09 (0.95–1.25)	0.23
T Stage (T3/4 vs. T1/2)	2.51 (2.05–3.07)	<0.02
N Stage (N+ vs. N0)	3.18 (2.61–3.87)	<0.03
MSI Status (MSI-H vs. MSS)	0.65 (0.51–0.83)	<0.04

Error analysis: probing the model's decision boundary

Error Analysis of Feature Space Representation: An error analysis was conducted to investigate the model's misclassifications. The t-SNE algorithm was utilized to project the high-dimensional feature vectors learned by the model into a three-dimensional space for visualization. This projection allows for an examination of the separability of the learned classes and the characteristics of misclassified samples.

As illustrated in Fig. 3, the visualization reveals two primary clusters corresponding to Adenocarcinoma (blue) and Normal (orange) tissue. While the classes are largely separable, there exists a notable region of

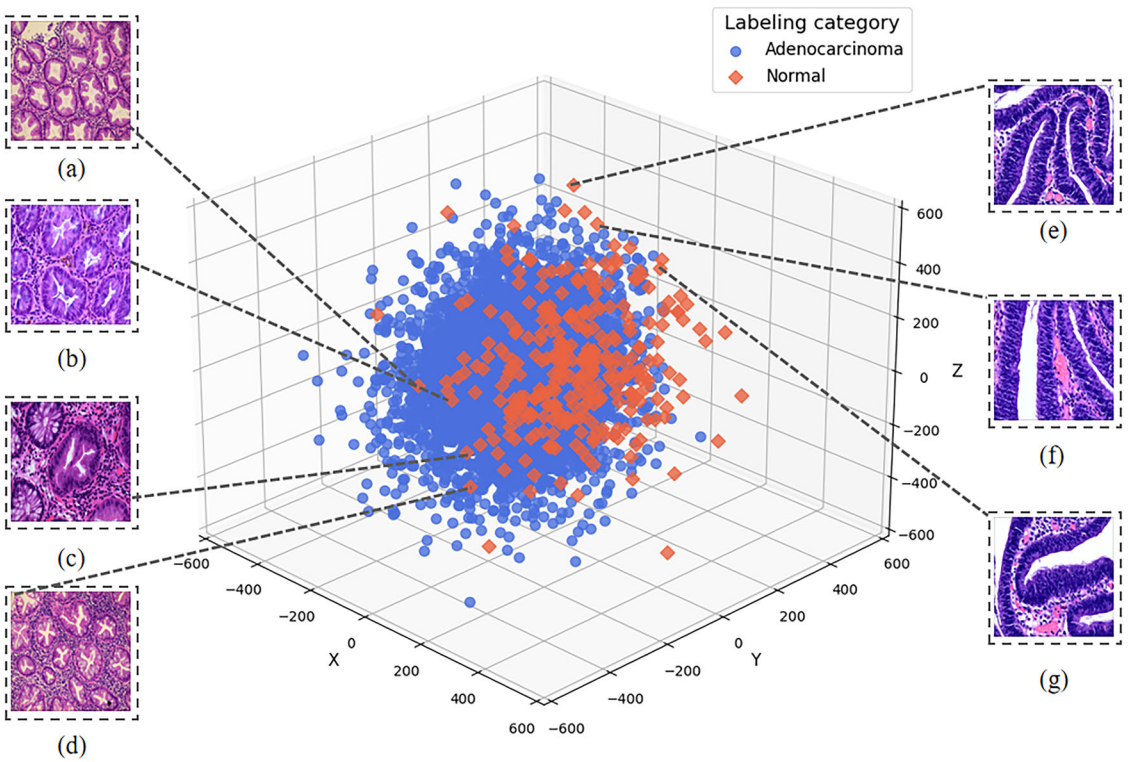

Fig. 3 | A t-SNE plot visualizing the model's learned feature representations for Normal and Adenocarcinoma tissue samples.

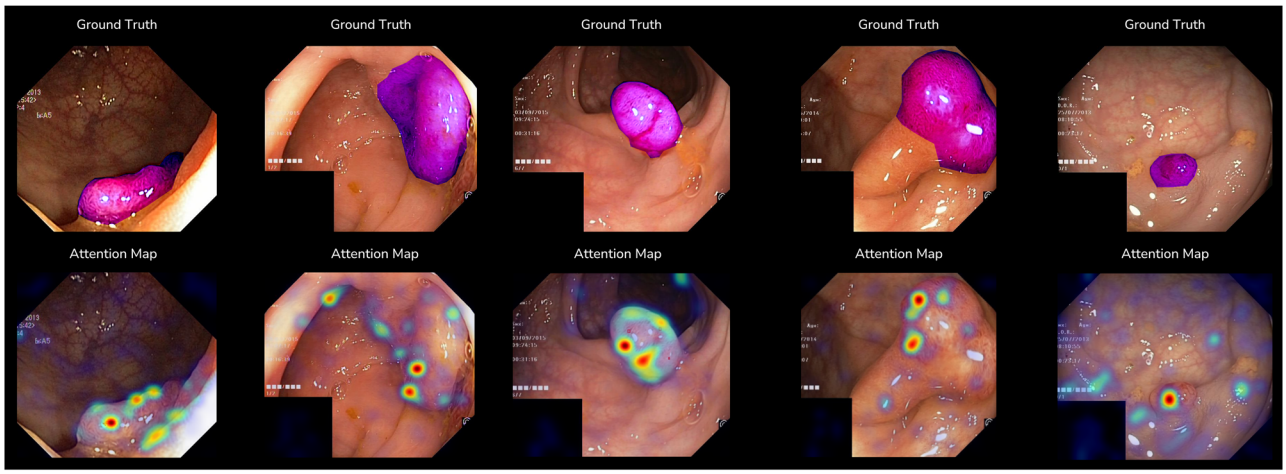

Fig. 4 | Visualization of model attention: a comparison with ground truth polyp segmentation.

overlap where feature representations of both classes are intermingled. This boundary region represents cases of high morphological ambiguity and is the primary source of model classification errors.

Analysis of Misclassified Normal Samples (False Positives) Fig. 3a–d: These images depict normal colorectal tissue that the model incorrectly classified as adenocarcinoma. In the feature space visualization, these samples are positioned within the ambiguous boundary region, closer to the adenocarcinoma cluster.

The misclassification can be attributed to the presence of morphological features that, while occurring in benign tissue, are quantitatively similar to patterns the model has learned to associate with malignancy. Specifically, these samples exhibit a more crowded glandular architecture and a degree of nuclear hyperchromasia. These features cause the model to generate a feature vector that crosses the decision boundary into the space defined for adenocarcinoma.

Analysis of Misclassified Adenocarcinoma Samples (False Negatives) Fig. 3e–g: These images show adenocarcinoma samples that the model incorrectly classified as normal. These samples also lie within the decision boundary but are located closer to the normal tissue cluster.

This classification error is a result of the tumors' well-differentiated nature. The malignant glands in these samples retain a highly organized structure that closely mimics the appearance of normal colonic crypts. The model has learned to associate features such as "lower-grade tumor-forming glands and a high tumor-to-stroma ratio" with low-risk or normal predictions. Because these cancerous tissues display benign-like morphology, their feature vectors are mapped into the region of the feature space predominantly occupied by normal samples, leading to their misclassification. These cases are known to be diagnostically challenging, and the model's errors reflect this inherent clinical ambiguity.

Model interpretability through attention mechanism visualization

A significant barrier to the widespread clinical adoption of deep learning models is their black box nature, which can obscure the reasoning behind their predictions and limit trust in high-stakes diagnostic environments. The visualizations presented in this section are a direct qualitative representation of the attention mechanism detailed. This mechanism is responsible for aggregating thousands of patch-level feature vectors into a single, cohesive slide-level representation. The generated heatmaps visualize the learned attention weights, where warmer colors (red, yellow) correspond to image patches that the model assigned higher importance for its final prediction.

Figure 4 provides a qualitative analysis of the model's learned attention, comparing the generated heatmaps to expert-annotated ground-truth polyp segmentations. The analysis reveals a strong and consistent spatial congruence between the high-attention "hotspots" and the true location and extent of the pathological lesions. In all five representative cases, the model correctly assigns the highest weights to regions within the polyp boundaries while effectively suppressing attention to the surrounding healthy mucosal tissue. This demonstrates a robust ability to differentiate pathological from normal tissue, which is a foundational requirement for any diagnostic model.

Discussion

This study introduces PRISM-CRC, a novel deep learning framework designed to overcome limitations in CRC management by integrating histopathology, radiology, and clinical data. The model demonstrated high prognostic accuracy, achieving a concordance index of 0.82 for 5-year DFS and an AUC of 0.91 for predicting MSI status, a key biomarker for immunotherapy. A central finding of the research is that this multimodal fusion is synergistically powerful; the integrated model significantly outperformed models trained on any single data type, suggesting it learns a complex "cross-modal grammar" of tumor biology that is more than the sum of its parts.

The primary clinical implication of PRISM-CRC is its ability to provide prognostic information that is independent of and complementary to the current gold standard, the AJCC TNM staging system. The model's risk score remained a strong, independent predictor of survival even after adjusting for TNM stage, indicating that it discovers novel, latent "digital biomarkers" of tumor aggressiveness from the integrated data. This has direct translational potential for personalizing treatment, particularly for intermediate-risk patients. For instance, the model can re-stratify Stage II patients, identifying a high-risk subgroup whose survival outcomes resemble those of Stage III patients, suggesting they may benefit from adjuvant chemotherapy-a treatment not typically recommended for this stage.

Despite its robust performance, the study acknowledges key limitations that define future research directions. A modest decrease in performance on datasets highlights the challenge of "domain shift," which can be caused by variations in data acquisition and patient populations across different institutions. Furthermore, an error analysis revealed that misclassifications occurred primarily in cases of high morphological ambiguity, where cancerous tissue closely resembled benign structures. Future work will focus on mitigating domain shift through methods like federated learning, incorporating molecular data to resolve ambiguity, and, most critically, validating the model's clinical utility through prospective, randomized controlled trials.

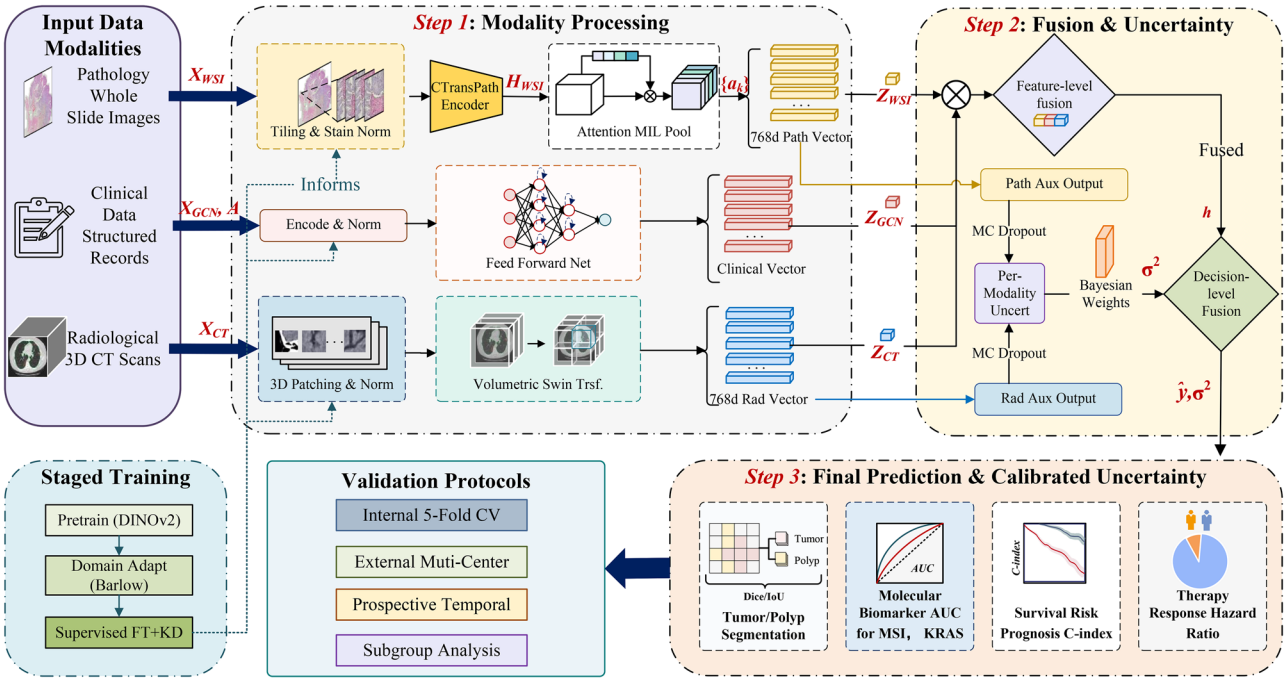

Fig. 5 | Overview of the three-stage multimodal prognosis framework. (1) Modality processing: WSIs are tiled, stain-normalized, and encoded with *CTransPath*; patch features are aggregated by gated-attention MIL modulated by mask-derived saliency. 3D CT volumes use a *Swin UNETR* encoder with ROI-gated and peritumoral-ring pooling plus compact shape/texture descriptors. Clinical/genomic variables are encoded by a *GCN* on prior graph A. Segmentation serves as a structural prior to gate WSI attention and enable CT ROI/peritumoral pooling and shape/texture descriptors. (2) Fusion & uncertainty: modality vectors $z_{\mathrm{WSI}}, z_{\mathrm{CT/Endo}}, z_{\mathrm{GCN}}$ are fused via low-rank multimodal fusion (LMF, rank r) into h, with auxiliary per-modality heads. Uncertainty via MC Dropout (optionally deep ensembles) yields calibrated $\hat{y}$ and variance σ^2. Training is three-stage: self-supervised visual pre-training (DINOv2), domain adaptation (Barlow Twins), then supervised fine-tuning with knowledge distillation. (3) Prediction & evaluation: an MLP survival head outputs risk/endpoint estimates; evaluation covers segmentation (Dice/IoU), biomarker AUCs, prognosis (C-index), therapy-response hazard ratios. Fusion remains tri-modal (WSI/CT/GCN); segmentation is not a modality.

Methods

Our proposed methodology shown in Fig. 5 presents a comprehensive, three-stage framework for integrating multimodal data to generate robust clinical predictions with associated uncertainty estimates. The entire workflow is summarized in Algorithm 1, with detailed implementations of each core step provided in Algorithm 2.

The process begins with modality-specific feature encoding. Each data type-histopathology (WSI), radiology (CT), endoscopy and genomics/clinical (GCN)-is processed by a specialized deep learning architecture designed to extract a rich feature representation. For WSIs, a CTransPath model encodes individual image patches, which are then intelligently aggregated into a single slide-level vector using a gated attention mechanism. Volumetric CT data is encoded using a 3D Swin UNETR encoder, while a Graph Convolutional Network (GCN) is used to model the relational structure of genomic and clinical data.

Next, these distinct feature vectors are integrated in the multimodal fusion stage. We employ Low-Rank Multimodal Fusion (LMF) to combine the representations from the WSI, CT, endoscopy and GCN encoders. This method efficiently captures the complex, multiplicative interactions between the modalities while maintaining computational tractability.

Finally, the unified feature vector is used for prediction and uncertainty quantification. The fused representation is passed to a Multi-Layer Perceptron (MLP) head to produce the final predictive output, such as a risk score. To quantify the reliability of this prediction, we utilize Monte Carlo (MC) Dropout, performing multiple stochastic forward passes. The mean of the resulting output distribution serves as the final prediction ($\hat{y}$), and its variance provides a principled measure of the model's predictive uncertainty (σ^2).

Algorithm 1. High-Level Multimodal Fusion Framework

Require: X_{WSI}, $X_{\mathrm{CT/Endo}}$, $X_{\mathrm{Endo}}X_{\mathrm{GCN}}$, A; T ▷ inputs unchanged
Ensure: $\hat{y}$, σ^2

▷ **Step 0: Derive structural priors (optional)**
$M_{\mathrm{CT/Endo}} \leftarrow$ SwinUNETR_Seg($X_{\mathrm{CT/Endo}}$); $M_{\mathrm{WSI}} \leftarrow$ WSI_Mask (X_{WSI})
▷ **Step 1: Encode each modality with ROI-aware aggregation**
$z_{\mathrm{WSI}}, z_{\mathrm{CT/Endo}}, z_{\mathrm{GCN}}$
$\leftarrow$ **EncodeAndAggregate**($X_{\mathrm{WSI}}, X_{\mathrm{CT/Endo}}, X_{\mathrm{GCN}}, A; M_{\mathrm{WSI}}$, $M_{\mathrm{CT/}}$ Endo)
▷ **Step 2: Low-rank multimodal fusion**
$h \leftarrow$ **LowRankFusion**($z_{\mathrm{WSI}}, z_{\mathrm{CT/Endo}}, z_{\mathrm{GCN}}$)
▷ **Step 3: Prediction with uncertainty**
$\hat{y}, \sigma^2 \leftarrow$ **PredictWithUncertainty**(h, T)
return $\hat{y}, \sigma^2$

Multimodal feature encoding architectures

Histopathological Image Encoding: Self-Supervised Hybrid Transformers: The analysis of histopathological WSIs presents significant computational challenges due to their gigapixel size, which renders direct processing by standard neural networks computationally prohibitive. Furthermore, the manual annotation of these slides by pathologists is a labor-intensive, costly, and subjective task, creating a considerable bottleneck in the application of supervised learning methods. To overcome these obstacles, the proposed framework integrates two strategies: the Multiple Instance Learning (MIL) paradigm for managing the large image sizes under weak supervision, and a powerful, self-supervised hybrid Transformer architecture, referred to as CTransPath, to learn robust, domain-specific feature representations from image patches.

Algorithm 2. Detailed Sub-procedures

procedure: EncodeAndAggreX_{WSI}, $X_{\mathrm{CT/Endo}}$, X_{GCN}, A; M_{WSI}, $M_{\mathrm{CT/Endo}}$
  $H_{\mathrm{WSI}} \leftarrow \{\mathrm{CTransPath}(x_k)\}$
  $\{a_k\} \leftarrow \mathrm{GatedAttention}(H_{\mathrm{WSI}})$; $\gamma_k \leftarrow \mathrm{MaskSaliency}(x_k, M_{\mathrm{WSI}})$
  $z_{\mathrm{WSI}} \leftarrow \sum_k (a_k \cdot \gamma_k)\, h_k$
  $H_{CT/Endo} \leftarrow \mathrm{SwinUNETR\_Encoder}(X_{CT/Endo})$
  $z^{\mathrm{roi}}_{CT/Endo} \leftarrow \mathrm{Pool}(H_{CT/Endo} \odot M_{\mathrm{CT/Endo}})$; $z^{\mathrm{peri}}_{CT/Endo} \leftarrow \mathrm{Pool}(H_{CT/Endo} \odot \mathrm{Ring}(M_{\mathrm{CT/Endo}}))$
  $z_{CT/Endo} \leftarrow [z^{\mathrm{roi}}_{CT/Endo}, z^{\mathrm{peri}}_{CT/Endo}, \mathrm{Shape/Texture}(M_{\mathrm{CT/Endo}}, X_{CT/Endo})]$
  $z_{GCN} \leftarrow \mathrm{GCN}(X_{GCN}, A)$
  return $z_{\mathrm{WSI}}, z_{\mathrm{CT/Endo}}, z_{\mathrm{GCN}}$
end procedure
procedure LowRankFusion($z_{\mathrm{WSI}}, z_{\mathrm{CT/Endo}}, z_{\mathrm{GCN}}$)
  $h \leftarrow \overrightarrow{0}$ ▷ Initialize final fused vector
  for $i = 1$ to r **do** ▷ Iterate over rank r
    $p^{(i)}_{WSI} \leftarrow w^{(i)}_{WSI} \cdot z_{WSI}$
    $p^{(i)}_{CT/Endo} \leftarrow w^{(i)}_{CT/Endo} \cdot z_{CT/Endo}$
    $p^{(i)}_{GCN} \leftarrow w^{(i)}_{GCN} \cdot z_{GCN}$
    $h \leftarrow h + (p^{(i)}_{WSI} \odot p^{(i)}_{CT/Endo} \odot p^{(i)}_{GCN})$ ▷ Hadamard product
  end for
  return h
end Procedure
procedure PredictWithUncertainty$h$, T
  $Y_{\mathrm{pred}} \leftarrow []$
  for $t = 1$ to T **do** ▷ Perform T stochastic forward passes
    Append MLP(h) to Y_{pred} ▷ Dropout is active in MLP layers
  end for
  $\hat{y} \leftarrow \mathrm{mean}(Y_{pred})$ ▷ Predictive Mean
  $\sigma^2 \leftarrow \mathrm{variance}(Y_{\mathrm{pred}})$ ▷ Predictive Variance
  return $\hat{y}, \sigma^2$
end Procedure

Segmentation-guided structural priors

For CT we use lesion/organ masks $M_{\mathrm{CT/Endo}}$ predicted by the 3D Swin UNETR encoder; for WSIs we use tissue/tumor masks M_{WSI} (e.g., epithelium vs. stroma) to filter patches and modulate attention weights. These priors serve two roles:

(i) **ROI-gated pooling**. Given token maps $H_{\mathrm{CT/Endo}}$ from the CT encoder, we compute ROI-pooled features $z^{\mathrm{roi}}_{\mathrm{CT/Endo}} = \mathrm{Pool}\left(H_{\mathrm{CT/Endo}} \odot M_{\mathrm{CT/Endo}}\right)$ and a peritumoral ring feature $z^{\mathrm{peri}}_{\mathrm{CT/Endo}} = \mathrm{Pool}\left(H_{\mathrm{CT/Endo}} \odot \mathrm{Ring}(M_{\mathrm{CT/Endo}})\right)$, then concatenate $z_{\mathrm{CT/Endo}} = \left[z^{\mathrm{roi}}_{\mathrm{CT/Endo}}, z^{\mathrm{peri}}_{\mathrm{CT/Endo}}\right]$. For WSIs, patches with $M_{\mathrm{WSI}}(x_k) = 0$ are down-weighted in MIL by $\tilde{a}_k \propto a_k \cdot \gamma(x_k)$, where $\gamma(\cdot) \in [0, 1]$ encodes mask-derived saliency.

(ii) **Shape and topology descriptors**. We derive compact radiomics-style descriptors from masks (volume, surface area, sphericity, margin sharpness, peritumoral intensity/texture). These are treated as deterministic functions of the base images and appended to the modality vectors before fusion.

Training: Segmentation supervision is used only when masks are available; otherwise the prior block is bypassed. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{surv}} + \lambda_{\mathrm{seg}}\left(\mathcal{L}^{\mathrm{CT/Endo}}_{\mathrm{Dice}} + \mathcal{L}^{\mathrm{WSI}}_{\mathrm{CE}}\right), \tag{1}$$

with a small λ_{seg} to keep prognosis primary. Importantly, no additional modality is introduced; segmentation refines where and how existing encoders pool features.

Practical interpretation: The qualitative behavior of the model follows directly from the priors we impose. In WSIs the gated MIL weights a_k are modulated by mask-derived saliency $\gamma_k \in [0, 1]$ (Section "Segmentation-Guided Structural Priors"), so high-weight instances are expected to concentrate on tissue/tumor regions and down-weight off-tissue patches by design; the attention here should be read as a *pooling mechanism* rather than a standalone saliency map. For CT, we explicitly split features into intra-lesional ROI and peri-tumoral rings, encouraging the encoder to represent both core and microenvironmental context without needing additional visualizations. Clinical/genomic variables are aggregated by a GCN over a prior graph A, which biases the representation toward biologically plausible co-variation. We keep segmentation as a structural prior only; it refines where features are pooled but is not treated as a separate fused modality. This architectural structure provides an interpretable pathway from data to prediction without expanding the figure set.

The MIL Paradigm: The MIL framework offers a compelling and necessary abstraction for applying deep learning techniques to WSIs in a weakly supervised setting. Unlike traditional methods that rely on exhaustive pixel-level annotations, MIL operates with only slide-level labels, such as a patient's diagnosis or survival outcome.

Under the MIL assumption, a WSI is considered as a "bag" of instances, where each instance corresponds to a smaller, computationally manageable image patch extracted from the original slide. Let $X = \{x_1, x_2, \ldots, x_K\}$ denote the set of image patches that constitute a WSI bag, where $x_k \in \mathbb{R}^{H \times W \times C}$ represents the k-th patch, and K is the total number of patches in the bag. Each bag is associated with a single label, $Y \in \{0, 1\}$, which reflects the latent, unknown labels of the individual instances $\{y_1, \ldots, y_K\}$, where $y_k \in \{0, 1\}$ for each patch. The MIL assumption posits that a bag is considered positive if and only if it contains at least one positive instance. This can be expressed formally as:

$$Y = \begin{cases} 1 & \text{if } \sum_{k=1}^{K} y_k > 0 \\ 0 & \text{if } \sum_{k=1}^{K} y_k = 0 \end{cases} \tag{2}$$

This formula (2) is equivalent to the max-operator approach, where $Y = \max_k \{y_k\}$. The objective of an MIL model is to develop a bag-level classifier that predicts the probability $p(Y|X)$ without direct access to the instance-level labels y_k.

The CTransPath Encoder Architecture: To extract meaningful features from each patch x_k, the proposed framework employs the CTransPath model[25], a hybrid architecture tailored specifically for histopathological image analysis. This model integrates the complementary strengths of CNNs and Transformers, thereby enabling the generation of highly informative feature embeddings.

The CTransPath model consists of two main components: a CNN front end and a Swin Transformer backbone.

The first component, the CNN, is employed for local feature extraction. Unlike the traditional approach used in Vision Transformers (ViT[26]), which divides an image into non-overlapping patches and linearly projects them, CTransPath integrates a shallow CNN–such as the initial layers of a ResNet to process each input patch. This design choice enhances the model's stability and improves its ability to capture fine-grained local textures, structures, and morphological details essential for histopathological analysis. In this architecture, the CNN serves as a sophisticated "patch embedding" layer, generating a sequence of local feature tokens that provide richer and more informative representations compared to the simple linear projections employed in traditional methods.

Following the CNN processing, the sequence of feature tokens is passed through a multi-scale Swin Transformer backbone. The self-attention mechanism within the Transformer excels at modeling long-range dependencies, allowing the model to capture global contextual information and spatial relationships between the local features derived from each patch. This enables a more holistic understanding of the image, which is particularly advantageous for analyzing histopathological data.

Swin Transformer V2 Block: The Swin Transformer V2 block[27] shown in Fig. 6 is a key component of the CTransPath backbone. It incorporates

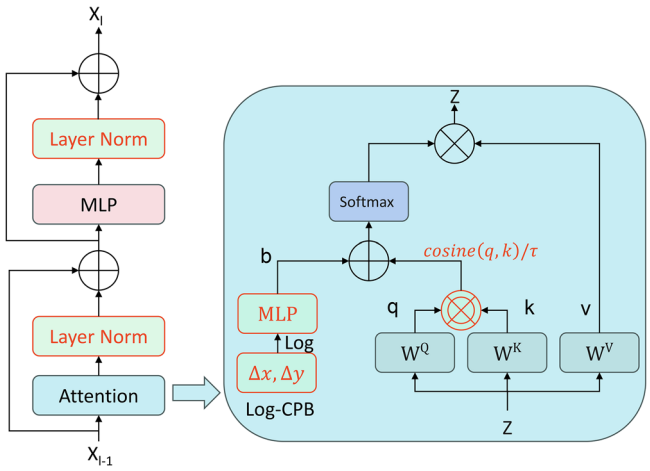

Fig. 6 | Swin Transformer V2 Block.

several innovations aimed at improving training stability, scalability, and transferability across different image resolutions.

Shifted Window Multi-Head Self-Attention (W-MSA & SW-MSA): Traditional self-attention mechanisms in Vision Transformers (ViT) are computationally expensive, scaling quadratically with respect to the number of image patches. The Swin Transformer addresses this by using a window-based approach, where self-attention is computed locally within non-overlapping windows (W-MSA). The windows are shifted across layers (SW-MSA), which reduces the computational complexity to linear while maintaining the ability to model global interactions.

Scaled Cosine Attention: To overcome training instability, Swin Transformer V2 replaces traditional dot-product attention with scaled cosine attention. This approach is less sensitive to the magnitude of query and key vectors, improving the model's stability during training.

Log-Spaced Continuous Position Bias (Log-CPB): To handle the challenge of model transfer across different image resolutions, Swin Transformer V2 introduces Continuous Position Bias (CPB). The position biases are computed using a meta network that transforms relative coordinates into logarithmic space before applying them to generate the bias:

$$\begin{aligned}\hat{\Delta x} &= \operatorname{sgn}(\Delta x)\cdot\log(|\Delta x|+1)\\ \hat{\Delta y} &= \operatorname{sgn}(\Delta y)\cdot\log(|\Delta y|+1)\end{aligned}$$

The final bias term is then given by:

$$B_{ij} = G(\hat{\Delta x}, \hat{\Delta y}) \tag{3}$$

This logarithmic transformation helps reduce the extrapolation ratio when transitioning between different window sizes, improving the model's performance when fine-tuning across resolutions.

Volumetric Medical Image Encoding: 3D Swin UNETR: For volumetric modalities such as Computed Tomography (CT) and Magnetic Resonance Imaging, preserving the full three-dimensional spatial context is essential for accurate analysis. Treating these datasets as independent 2D slices neglects vital depth-wise information and the inter-slice relationships, which are critical for tasks like tumor volume estimation or assessing the invasion of adjacent structures. To address this, a dedicated 3D architecture is employed for the effective processing of these volumetric inputs. In this context, the Swin UNETR encoder is utilized.

The Swin UNETR[28] architecture adapts the successful Swin Transformer to the domain of 3D medical image segmentation, incorporating it into a U-Net-inspired encoder–decoder framework. The primary focus for feature extraction lies within the hierarchical encoder component of the architecture.

The operation of the 3D Swin UNETR encoder extends its 2D counterpart into the three-dimensional space. The process begins with the partitioning of the input 3D volume, denoted as $X \in \mathbb{R}^{H\times W\times D\times C_{\text{in}}}$, where H, W, D represent the spatial dimensions, and C_{in} is the number of input channels. This 3D volume is divided into a grid of non-overlapping 3D patches, or "voxels," typically of a fixed size such as $4\times4\times4$. Each voxel is then flattened into a 1D vector and linearly projected into a token embedding of dimension C. This transforms the 3D volume into a sequence of 1D tokens, which are subsequently processed by the Transformer blocks.

These tokens are passed through a series of 3D Swin Transformer blocks, where multi-head selfattention is performed within local 3D windows. The encoder follows a hierarchical structure, comprising multiple stages. After each stage, a "patch merging" layer is applied, down-sampling the spatial resolution of the tokens by a factor of 2 along each dimension, while simultaneously doubling the feature dimension C. This process results in a pyramidal feature representation that captures both fine-grained details and high-level contextual information across multiple resolutions from the 3D volume.

Genomic and Clinical Data Encoding: Graph Convolutional Networks: Genomic data, such as gene expression levels or somatic mutation statuses, and structured clinical data (e.g., tumor stage, patient demographics) inherently exhibit complex, interdependent relationships. These features do not function independently but instead operate within intricate, interconnected biological networks and pathways. Traditional methods, such as MLPs, treat these features as isolated entities, disregarding their relational structure. To more effectively model these relationships and integrate prior biological knowledge into the model, the framework leverages a GCN[29].

GCN Layer-wise Propagation: A GCN operates on graph-structured data, represented as $G=(V, E)$, where V denotes a set of N nodes (such as genes or clinical variables), and E is the set of edges that represent relationships between these nodes (e.g., protein-protein interactions or clinical correlations). The network receives a node feature matrix $X \in \mathbb{R}^{N\times D}$, where each row corresponds to the feature vector of a node, and an adjacency matrix $A \in \mathbb{R}^{N\times N}$, which encodes the graph's connectivity.

The core operation of a GCN is its layer-wise propagation rule, which updates the feature representation of each node by aggregating information from its neighboring nodes. For layer l, the updated feature matrix $H^{(l+1)} \in \mathbb{R}^{N\times F}$ is computed from the features $H^{(l)} \in \mathbb{R}^{N\times D}$ of the previous layer, according to the following rule formulated by Kipf and Welling:

$$H^{(l+1)} = \sigma\left(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \tag{4}$$

where $H^{(0)}=X$. The matrix $\hat{D}$ is the diagonal degree matrix of $\hat{A}$, where each diagonal element $\hat{D}_{ii}=\sum_j \hat{A}_{ij}$ represents the degree of node i, including its self-loop. The term $\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}$ represents the symmetrically normalized adjacency matrix. This normalization is critical for stable training, as it ensures that the aggregation of features from neighbors is weighted by the inverse of their degrees. This prevents the scale of feature vectors from becoming excessively large for nodes with high degrees and maintains a well-behaved message passing mechanism.

The matrix $W^{(l)} \in \mathbb{R}^{D\times F}$ is a learnable weight matrix specific to each layer, which linearly transforms the aggregated feature vectors into a new feature space of dimension F. A non-linear activation function $\sigma(\cdot)$, such as ReLU, is applied element-wise.

To enable each node to aggregate features from itself as well as from its neighbors, the adjacency matrix A is augmented with self-loops by adding the identity matrix I, resulting in $\hat{A}=A+I$.

By stacking multiple layers of GCN, each node can aggregate information from a k-hop neighborhood. This enables the model to learn complex, multi-relational features that are essential for understanding the underlying biological processes within genomic and clinical data.

Aggregation and fusion of multimodal features

Following the initial encoding of each modality's data, the framework has derived rich, modality-specific feature representations. The subsequent task involves effectively combining these features to achieve a unified, predictive model. This process is segmented into two distinct phases, each addressing a unique set of challenges inherent to the individual modalities, while working towards a cohesive and integrated representation. The first phase, instance aggregation, addresses the specific intricacies of histopathology data by transforming the thousands of patch-level feature vectors derived from a WSI into a coherent, slide-level summary. The second phase, cross-modal fusion, merges this slide-level summary with the feature vectors derived from radiological and genomic encoders. This two-phase approach is designed with careful consideration; it ensures the most appropriate technique is applied at each stage. Specifically, for instance aggregation, where clinical interpretability is paramount, an attention-based mechanism is utilized. This not only aggregates the features but also identifies diagnostically significant regions within the WSI.

For the cross-modal fusion stage, which requires efficient modeling of complex, non-linear interactions between pre-aggregated feature vectors, a highly efficient tensor decomposition method is employed. The separation of concerns in this approach interpretable aggregation within a single modality and efficient fusion across modalities is central to the design of this system, which is optimized for both performance and clinical utility. Segmentation outputs are used only to gate pooling and derive deterministic descriptors; they are *not* treated as an additional modality, so the tri-modal fusion dimensionality and LMF rank remain unchanged.

Instance Aggregation: Attention-based MIL Pooling: Following the encoding of each of the K patches from a WSI into a set of feature vectors $\{h_1, h_2, \ldots, h_K\}$, where $h_k \in \mathbb{R}^M$, a mechanism is required to aggregate these instance-level representations into a single, fixed-size bag representation $z \in \mathbb{R}^M$. This aggregation function must be permutation-invariant, meaning the output should not depend on the order in which the patches are processed. While simpler approaches such as max-pooling or mean-pooling satisfy this requirement, they tend to be suboptimal. Max-pooling prioritizes only the most salient instance, discarding other valuable information, whereas mean-pooling can dilute important signals by averaging over a large collection of benign instances.

Attention-based pooling provides a more robust and flexible solution by computing a learnable, weighted average of the instance embeddings. This method allows the model to dynamically assign higher weights to the most informative instances for a given prediction task, enabling it to focus on diagnostically relevant regions.

Gated Attention Variant: To further enhance the expressive capacity of the attention mechanism, a gating component can be incorporated. The standard tanh function exhibits linear behavior for small inputs, which may limit its ability to capture more intricate dependencies. The gated attention mechanism addresses this by adding an additional learnable transformation and a sigmoid gate, which element-wise modulates the output of the tanh function. This introduces a learnable nonlinearity that can adapt more flexibly to the data. The formulation for the gated attention weights is as follows:

$$a_k = \frac{\exp\left(w^T\left(\tanh\left(Vh_k^T\right) \odot \mathrm{sigm}\left(Uh_k^T\right)\right)\right)}{\sum_{j=1}^{K}\exp\left(w^T\left(\tanh\left(Vh_j^T\right) \odot \mathrm{sigm}\left(Uh_j^T\right)\right)\right)} \tag{5}$$

In formula (5), $U \in \mathbb{R}^{L\times M}$ represents an additional set of learnable parameters, $\odot$ denotes the element-wise (Hadamard) product, and $\mathrm{sigm}(\cdot)$ denotes the sigmoid function.

This gated attention variant introduces a more complex, adaptive weighting scheme that allows the model to better handle varying levels of feature importance across instances, thereby enhancing its interpretability and performance.

Cross-modal integration: low-rank multimodal fusion (LMF)

After modality-specific encoders and the MIL aggregator have produced a set of fixed-size feature vectors-namely z_{WSI} from histopathology, $z_{\mathrm{CT/Endo}}$ from radiology, and z_{GCN} from genomics/clinical data-the final step is to combine them into a unified, comprehensive representation. A simple approach, such as concatenating the features, is computationally efficient but limited in its expressive power, as it can only model additive interactions between the modalities. A more sophisticated method is tensor fusion, which computes the outer product of the feature vectors to create a high-order tensor that explicitly captures all possible multiplicative (i.e., higher-order) interactions between the modalities. However, this approach suffers from the curse of dimensionality, as the number of parameters in the resulting fusion layer grows exponentially with the number of modalities, making it computationally prohibitive for complex systems.

LMF[30] presents a more elegant and efficient solution to this issue. It captures the rich, multiplicative interactions of tensor fusion while maintaining a computational complexity that scales linearly with the number of modalities.

The key idea behind LMF is to avoid explicitly constructing the high-dimensional weight tensor required for fusion. Instead, it parameterizes the fusion operation using modality-specific low-rank factors, which implicitly reconstruct a low-rank approximation of the full weight tensor.

Let the feature vectors for M modalities be denoted as $\{z_m\}_{m=1}^{M}$, where $z_m \in \mathbb{R}^{d_m}$. The traditional tensor fusion approach would first form an input tensor $Z = \bigotimes_{m=1}^{M} z_m$, and then compute a fused representation $h = W \cdot Z$, where W is a large, order- $M+1$ weight tensor.

In contrast, LMF bypasses the need for constructing the high-dimensional tensor by directly computing the fused output $h \in \mathbb{R}^{d_h}$ through a parallel decomposition. The weight tensor W is assumed to have a low-rank structure and is decomposed into a sum of r rank-one tensors, where each rank-one component is itself an outer product of modality-specific weight vectors. This allows the fusion operation to be reformulated as follows.

$$h = \sum_{i=1}^{r}\left(\odot_{m=1}^{M}\left(w_m^{(i)} \cdot z_m\right)\right) \tag{6}$$

Component Breakdown: - r represents the rank of the decomposition, a hyperparameter that governs the expressive capacity of the fusion layer. - $w_m^{(i)} \in \mathbb{R}^{d_h \times d_m}$ is a trainable, modality-specific weight matrix for the i-th rank component. - $w_m^{(i)} \cdot z_m$ is a standard matrix-vector multiplication that projects the input feature vector z_m into the hidden fusion space of dimension d_h. - $\bigodot_{m=1}^{M}$ denotes the element-wise (Hadamard) product across the M projected vectors for each rank component. This operation efficiently captures the multiplicative interactions between modalities for that specific rank. - $\sum_{i=1}^{r}$ represents the summation over all r rank components to produce the final fused vector h.

This formulation is mathematically equivalent to performing a full tensor fusion but is substantially more efficient. The computational complexity is reduced from the exponential $O\left(d_h \cdot \prod_{m=1}^{M} d_m\right)$ of traditional tensor fusion to a linear complexity of $O\left(d_h \cdot r \cdot \sum_{m=1}^{M} d_m\right)$.

Prediction and uncertainty quantification

The final stage of the architecture shown in Fig. 7 transforms the fused multimodal feature representation into a clinically relevant prediction and, crucially, quantifies the uncertainty associated with that prediction. A model that only produces a point estimate (e.g., "This patient has an 82% chance of survival") is of limited clinical value. A truly effective decision support tool must also provide an associated confidence measure (e.g., "...with a 95% confidence interval of [65%, 95%]"). This allows clinicians to distinguish between confident predictions and speculative estimates, which is vital for integrating such models into high-stakes medical workflows. Thus, the inclusion of uncertainty quantification is not merely an add-on but a core

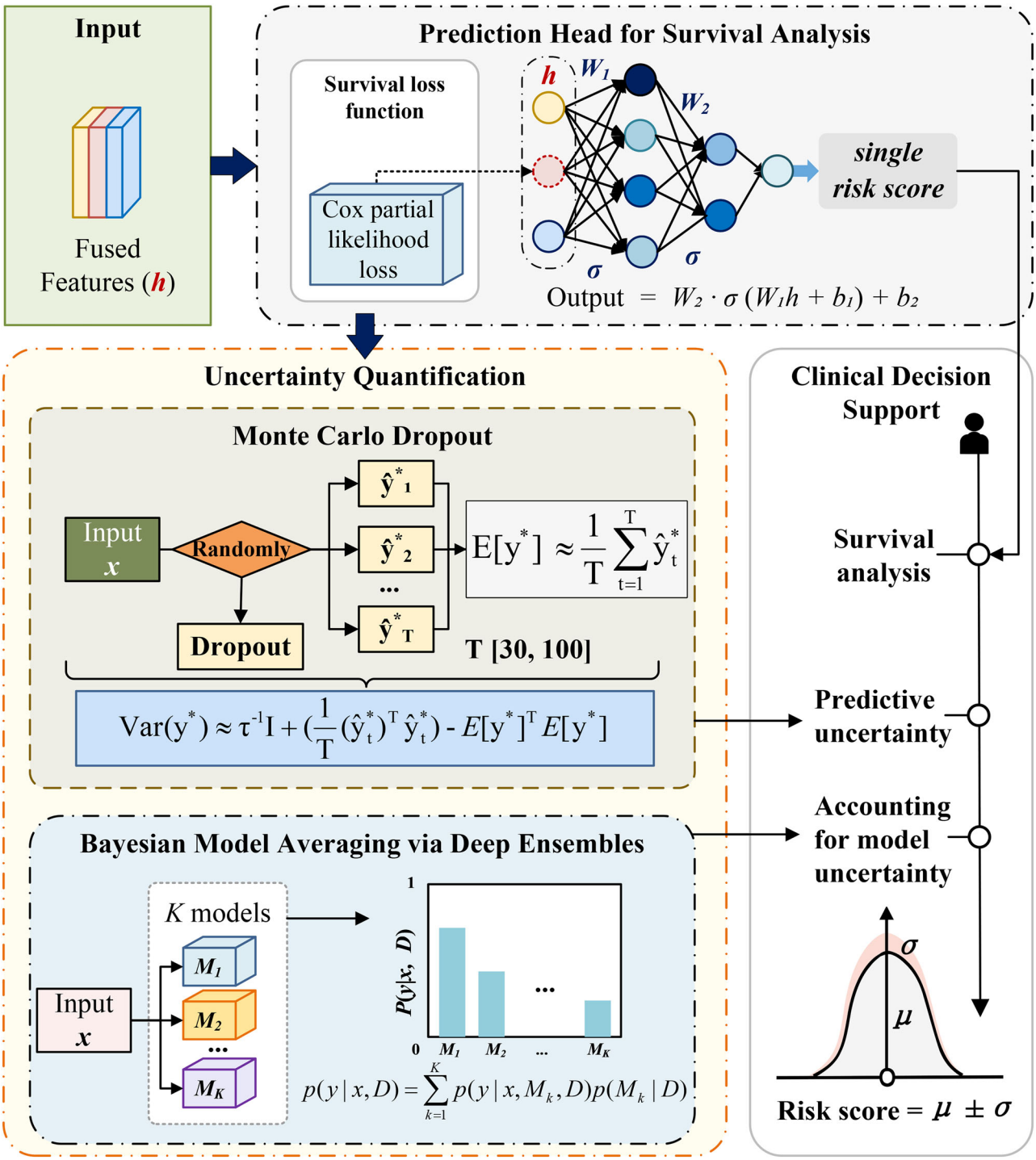

Fig. 7 | Uncertainty Quantification.

design principle. This reflects a shift from constructing models that are solely accurate to those that are also robust, reliable, and trustworthy.

Prediction Head for Survival Analysis: Once the final fused feature vector h is obtained, it is passed through a prediction head to generate the final output. Given the prognostic nature of the task, this head is specifically tailored for survival analysis.

The prediction head is generally a MLP that maps the high-dimensional feature vector h to the parameters of the survival model. For instance, in a Cox Proportional Hazards model, the MLP would output a single risk score. The architecture of the prediction head can be expressed as:

$$\text{output} = W_2 \cdot \sigma(W_1 h + b_1) + b_2 \tag{7}$$

where W_1, b_1, W_2, b_2 are the learnable weights and biases of the MLP, and σ represents a non-linear activation function.

A critical challenge in survival analysis is handling censored data. In clinical studies, many patients may still be alive or lost to follow-up by the time the study concludes; their exact survival time is unknown, but it is at least as long as their follow-up time. Standard regression losses, such as

Mean Squared Error, are unsuitable for this type of data. Specialized survival loss functions are required, and a common choice is the Cox partial likelihood loss. This loss function appropriately handles censored observations by comparing the risk scores of patients who experienced an event (e.g., death) to those who were at risk at the same time. The clinical endpoints for survival analysis are typically Overall Survival (OS), Progression-Free Interval (PFI), Disease-Free Interval, or Disease-Specific Survival.

Uncertainty Quantification: From Point Estimates to Predictive Distributions: To go beyond simple point estimates, the framework incorporates methods to approximate a Bayesian predictive distribution. This allows for principled quantification of the model's uncertainty regarding its predictions.

Monte Carlo (MC) Dropout: Monte Carlo (MC) Dropout[31] is a computationally efficient and widely used method for approximating Bayesian inference in deep neural networks. The key idea, introduced by Gal and Ghahramani, is that a neural network trained with dropout is mathematically equivalent to an approximation of a deep Gaussian Process. This allows the dropout mechanism, which is typically used only during training for regularization, to be repurposed during inference to estimate model uncertainty.

The procedure involves performing T stochastic forward passes on the same input sample x^*, with dropout layers kept active during each pass. Since dropout randomly deactivates different neurons in each pass, this generates a distribution of T different outputs, $\{\hat{y}_t^*\}_{t=1}^T$, which can be interpreted as samples from the model's approximate predictive posterior distribution. - Predictive Mean: The final prediction is taken as the empirical mean of these stochastic outputs, offering a more robust estimate than a single deterministic pass:

$$E[y^*] \approx \frac{1}{T}\sum_{t=1}^{T}\hat{y}_t^* \tag{8}$$

Predictive Uncertainty: The model's uncertainty is quantified by the sample variance of the T outputs. This variance represents epistemic uncertaintyuncertainty due to the model's parameters, which could be reduced with additional data. The predictive variance is approximated as:

$$\mathrm{Var}(y^*) \approx \tau^{-1}I + \left(\frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t^*)^T\hat{y}_t^*\right) - E[y^*]^T E[y^*] \tag{9}$$

The first term, $\tau^{-1}I$, represents the model's estimate of aleatoric uncertainty (inherent data noise), where τ is the model precision, a hyperparameter related to the weight decay used during training. The remaining terms reflect the sample variance of the predictions, capturing epistemic uncertainty. The number of forward passes, T, is a hyperparameter, typically between 30 and 100, depending on the application.

Bayesian Model Averaging (BMA) via Deep Ensembles: Bayesian Model Averaging (BMA)[32] offers a theoretically grounded framework for accounting for model uncertainty. Rather than relying on a single model, BMA averages the predictions of multiple models, weighting each by its posterior probability of being the correct model given the data.

Given a set of K models, $\{M_1, \ldots, M_K\}$, the full Bayesian predictive distribution for a new data point x is:

$$p(y|x, D) = \sum_{k=1}^{K} p(y|x, M_k, D)\,p(M_k|D) \tag{10}$$

where $p(y|x, M_k, D)$ is the predictive distribution of model M_k, and $p(M_k|D)$ is the posterior probability of model M_k given the training data D.

While computing the true posterior $p(M_k|D)$ is intractable for deep neural networks, Deep Ensembles have emerged as a simple and effective approximation. This method involves training N identical network architectures from scratch using different random weight initializations. Due to the non-convex nature of the loss landscape in deep learning, these models converge to different local minima, effectively sampling different high-performing models from the model space.

At inference time, the predictions of the N models are averaged. In the simplest case, this corresponds to BMA with a uniform prior over the models, where $p(M_k|D) \approx 1/N$. The predictive mean and variance are then the sample mean and variance of the ensemble's outputs. Deep Ensembles typically provide more robust and better-calibrated uncertainty estimates than MC Dropout, as they explore different basins of attraction in the weight space, whereas MC Dropout typically approximates the posterior within a single basin.

How to read a model output: In practice, each prediction consists of a risk/endpoint estimate $\hat{y}$ and an uncertainty estimate σ^2 (MC Dropout).

Low σ^2 indicates the model is behaving consistently across stochastic passes; high σ^2 flags cases where predictions are unstable (e.g., scarce tumor content, artifacts, or out-of-distribution patterns) and may warrant closer review. Auxiliary per-modality heads and the explicit ROI vs. peri-tumoral CT decomposition provide lightweight, text-only cues about which sources of evidence are driving the fused representation h; these cues are intended as sanity checks and do not require additional images.

Together with mask-gated WSI pooling and the GCN prior, these design choices offer a transparent explanation of model behavior while keeping the figure count unchanged.

Dataset

We collected and curated five large-scale public datasets to develop and evaluate the proposed framework. These datasets were categorized into two primary groups: multimodal prognostic cohorts (TCGA-COAD[33] and TCGA-READ[34]) and specialized diagnostic benchmarks (EBHI-SEG[35], REAL-Colon[36], and Kvasir-SEG[37]). The demographic, clinical, and technical characteristics of these cohorts are summarized in Tables 5 and 6.

For the primary tasks of survival prediction and multimodal fusion, we utilized data the Colon Adenocarcinoma (TCGA-COAD) and Rectum Adenocarcinoma (TCGA-READ) collections. As detailed in Table 5, these cohorts provide a comprehensive multimodal foundation.

TCGA-COAD cohort comprises 515 patients, predominantly with colon tumors (~85%) and a median age of 67.8 years. It offers extensive genomic coverage, with RNA-Seq availability for 99.6% of cases and a diverse molecular profile (CMS1-CMS4). Microsatellite Instability-High (MSI-H) status is observed in approximately 15% of cases, with the remaining ~85% classified as MSS/MSI-L. The dataset includes 459–515 H&E stained WSIs and comprehensive EHR data including TNM staging and survival metrics.

TCGA-READ cohort consists of 170 patients with rectal malignancies. In contrast to COAD, MSI-H is rare (<5%), and the molecular landscape is dominated by CMS2 and CMS4 subtypes. The imaging data is robust, with near-universal WSI availability (164–170 slides) and high radiological coverage (1796 DICOM files), supporting the evaluation of radiomic features.

We utilized three specialized datasets, as described in Table 6. EBHI-SEG serving as a benchmark for microscopic tissue analysis, this dataset contains 4456 H&E histopathology images paired with pixel-level ground truth segmentation masks. It supports fine-grained classification across six distinct histological classes, ranging from normal mucosa and polyps to high-grade intraepithelial neoplasia and adenocarcinoma, all captured at 400× magnification. Kvasir-SEG addresses endoscopic challenges, and this dataset provides 1000 polyp images with corresponding segmentation masks. It features variable resolutions and diverse polyp morphologies, offering a rigorous standard for foreground-background segmentation tasks. REAL-Colon representing a large-scale video benchmark, this dataset includes over 2.7 million full-resolution frames derived from 60 full-procedure colonoscopy videos. It contains 132 unique polyps annotated with 350,264 bounding boxes, alongside histological and anatomical metadata, ensuring the model's exposure to diverse, multi-center endoscopic environments.

Table 5 | Comprehensive comparison of TCGA-COAD and TCGA-READ Datasets: multi-modal statistics, clinical characteristics, and prognostic features

Category	Specific Feature	TCGA-COAD (Colon)	TCGA-READ (Rectum)
Sample Demographics	Total Patient Count (*N*)	515	170
	Tissue Sample Availability	461 (89.3%)	164–170 (97%)
	Primary Tumor Location	Colon (~85%), Rectosigmoid	Rectum (100%)
	Gender Distribution (M/F)	52%/48%	~53%/47%
Genomic & Transcriptomic	RNA-Seq Availability (Cases)	~ 459 (99.6%)	~164–170 (97%)
	RNA-Seq Data Format	HTSeq-Counts, FPKM, TPM	HTSeq-Counts, FPKM, TPM
	Microsatellite Instability (MSI)	MSI-H (~15%), MSS/MSI-L (~85%)	MSI-H (<5%), MSS (>95%)
	Dominant Molecular Subtypes	CMS1, CMS2, CMS3, CMS4	CMS2, CMS4 (CMS1 rare)
Imaging Modalities	Whole Slide Images (WSI)	459–515 (H&E stained)	165–170 (H&E stained)
	Radiology Modality	CT (Predominant)	CT & MRI (Mixed)
	Radiology Data Volume	8387 DICOM (~25 patients, 4.9%)	1796 DICOM (High coverage)
Clinical & Prognosis	Median Overall Survival (OS)	7 years (84 months)	5 years (60 months)
	5-Year Survival (Stage I vs IV)	~90% vs ~15–20%	Overall ~60–65%
	EHR Data Completeness	Comprehensive (TNM, Age, OS, PFI)	Comprehensive (TNM, Therapy)

Table 6 | Colorectal cancer and polyp dataset comparison (selected datasets)

Dataset Name	Modality	Image Type	Number of Images/ Frames	Labels/Classes	Image Specifications	Annotation Type
EBHI-SEG	H&E Histopathology (Hematoxylin & Eosin)	Microscopic Histopathology	4456 image + 4456 ground truth segmentation masks	6 classes: Normal, Polyp, Low-grade Intraepithelial Neoplasia, High-grade Intraepithelial Neoplasia, Serrated Adenoma, Adenocarcinoma	224 × 224 pixels, 400× magnification (10× eyepiece, 40× objective)	Pixel-level segmentation masks
REAL-Colon	Endoscopy Video	Full-resolution Colonoscopy Video Frames	2,757,723 video frames from 60 full-procedure colonoscopy videos	132 polyps with bounding-box annotations, histology, size, anatomical location	Native full-resolution, multi-center diverse endoscopes, variable fps	Bounding boxes (350,264 annotations)
Kvasir-SEG	Endoscopy Image	Colonoscopy Video Frames	1000 polyp images + 1000 segmentation masks	Binary: Polyp (foreground) vs. Background, bounding boxes available	Variable resolution, extracted from endoscopy videos	Pixel-level segmentation masks

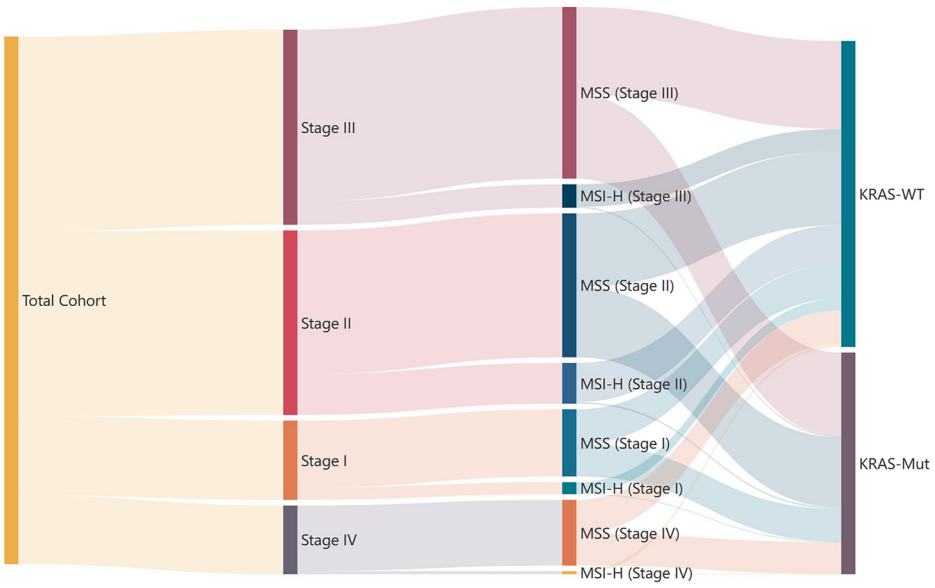

Fig. 8 | The molecular and pathological stratification of the training cohort.

Training cohort stratification

To provide a comprehensive visual overview of the patient distribution within the training cohort was constructed (Fig. 8). It illustrates the hierarchical stratification of the patients based on three key clinicopathological and molecular variables. The width of the links in the diagram is directly proportional to the number of patients, offering an intuitive representation of the cohort's composition at each level of classification.

It depicts the flow of the total patient cohort through three successive layers of stratification. The initial division separates the cohort into four pathological stages (Stage I, II, III, and IV). Each of these stage-specific groups is then further subdivided based on MSI status, distinguishing between MSI-High (MSI-H/dMMR) and Microsatellite Stable (MSS/pMMR) tumors. The final level of stratification partitions each of these subgroups by KRAS mutation status (Mutated vs. Wild-Type). MSI status was used solely as a prediction target for one task; it was not included as an input feature for the survival prediction. The multimodal model learns to predict MSI from imaging and other clinical inputs, without access to the actual MSI label during inference.

Figure 8 effectively highlights the complex, non-uniform distribution of molecular subtypes across different clinical stages. It visually confirms several well-established biological correlations within CRC. For instance, the diagram clearly shows the enrichment of MSI-H tumors within the Stage II patient group, which represents the largest single contributor to the overall MSI-H population. Furthermore, it starkly illustrates the strong inverse relationship between MSI status and KRAS mutations; the flows originating from MSI-H nodes are shown to channel almost exclusively into the Wild-Type category of KRAS, reflecting the distinct tumorigenesis pathways these alterations typically represent. The structured clinical features used in the model included the age of the patient at diagnosis, sex, tumor location (colon vs. rectum), and pathological stage (I–IV). These were the only clinical variables incorporated into the PRISM-CRC model's input.

Optimization and training protocol

Our framework employs a staged training strategy to effectively leverage both specialized diagnostic datasets and comprehensive multimodal prognostic cohorts.

Stage 1: Unimodal Representation Learning (Pretraining & Adaptation). We first initialize the modality-specific encoders using large-scale, single-modality datasets to handle the domain shift and feature extraction challenges. For the endoscopic and radiological encoders, we utilize REAL-Colon and Kvasir-SEG to learn robust lesion detection and segmentation features. For the histopathology encoder, we employ EBHI-SEG to capture fine-grained tissue morphologies ranging from normal mucosa to adenocarcinoma.

Stage 2: Multimodal Fusion and Supervised Fine-tuning. In the final stage, we integrate the pretrained encoders into the full PRISM-CRC framework using the TCGA-COAD and TCGA-READ cohorts. These datasets provide the necessary paired multimodal data (WSI, CT, Clinical) and survival endpoints.

Model training used the Adam optimizer with an initial learning rate of 10^{-4}. A plateau-based schedule was applied such that the learning rate was multiplied by 0.1 after the validation loss failed to improve for 5 consecutive epochs. Training proceeded for up to 500 epochs, with early stopping triggered if the validation loss did not improve for 10 consecutive epochs. Regularization included dropout with rate 0.3 in the fully connected layers and L_2 weight decay of 10^{-5}. Batch sizes were specified per modality input and kept fixed across experiments; the exact values used in all runs are provided alongside the released configuration files and scripts. All experiments were conducted on a workstation equipped with four NVIDIA RTX 4090 GPUs (24 GB memory each). End-to-end training of the full model required approximately 24 h.

Ethics approval and consent to participate

Not applicable. This study relied exclusively on fully de-identified, publicly available datasets (TCGA-COAD, TCGA-READ, EBHI-SEG, REAL-Colon, Kvasir-SEG) obtained under their respective data-use policies. No interaction with human participants or access to identifiable private information occurred; therefore, institutional review board (IRB) approval and informed consent were not required.

Data availability

Custom scripts for data preprocessing, model training, and evaluation used in this study will be released on GitHub upon publication. All experiments are reproducible using the provided scripts, which are based on standard PyTorch pipelines. All datasets used in this study are publicly accessible:TCGA-COAD: https://portal.gdc.cancer.gov/projects/TCGA-COADTCGA-READ: https://portal.gdc.cancer.gov/projects/TCGA-READEBHI-SEG: https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2023.1114673/fullREAL-Colon: https://plus.figshare.com/articles/media/REAL-colondataset/22202866?file=39461254Kvasir SEG: https://datasets.simula.no/kvasir-seg/Custom scripts for data preprocessing, model training, and evaluation used in this study will be released on GitHub upon publication. All experiments are reproducible using the provided scripts, which are based on standard PyTorch pipelines.

Code availability

Custom scripts for data preprocessing, model training, and evaluation used in this study are publicly available. All experiments are reproducible using the provided scripts, which are based on standard PyTorch pipelines. The code has been released: https://anonymous.4open.science/r/PRISM-CRC-8644/README.md.

Received: 15 September 2025; Accepted: 23 November 2025;
Published online: 05 December 2025

References

1. Wang, J. et al. Artificial intelligence in colorectal cancer imaging: recent advances and clinical applications. *Front. Oncol.* **15**, 1499223 (2025).
2. Usman, S. M., Ali, Y. & Taj, I. Early stage detection of colorectal cancer using segmentation of polyps. In *2024 International Conference on IT Innovation and Knowledge Discovery (ITIKD)*, 1–5 (IEEE, 2025).
3. Gustav, M. et al. Deep learning for dual detection of microsatellite instability and POLE mutations in colorectal cancer histopathology. *NPJ Precis. Oncol.* **8**, 115 (2024).
4. Kim, Y. I. et al. A randomized controlled trial of a digital lifestyle intervention involving postoperative patients with colorectal cancer. *npj Digit. Med.* **8**, 296 (2025).
5. Li, H. et al. Systematic review and meta-analysis of deep learning for MSI-H in colorectal cancer whole slide images. *npj Digit. Med.* **8**, 456 (2025).
6. Sun, L., Zhang, R., Gu, Y., Huang, L. & Jin, C. Application of artificial intelligence in the diagnosis and treatment of colorectal cancer: a bibliometric analysis, 2004-2023. *Front. Oncol.* **14**, 1424044 (2024).
7. Singh, V. K. et al. KRASFormer: a fully vision transformer-based framework for predicting KRAS gene mutations in histopathological images of colorectal cancer. *Biomed. Phys. Eng. Express* **10**, 055012 (2024).
8. Sari, M., Moussaoui, A. & Hadid, A. Deep learning techniques for colorectal cancer detection: convolutional neural networks vs vision transformers. In *2024 2nd International Conference on Electrical Engineering and Automatic Control (ICEEAC)*, 1–6 (IEEE, 2024).
9. Hicham, K. et al. 3d cnn-bn: A breakthrough in colorectal cancer detection with deep learning technique. In *2024 4th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, 1–6 (IEEE, 2024).
10. Rhanoui, M. et al. Multimodal machine learning for predicting post-surgery quality of life in colorectal cancer patients. *J. Imaging* **10**, 297 (2024).
11. Li, H. et al. Rethinking transformer for long contextual histopathology whole slide image analysis. Preprint at https://arxiv.org/abs/2410.14195 (2024).

12. Ignatov, A. & Malivenko, G. Nct-crc-he: Not all histopathological datasets are equally useful. Preprint at https://arxiv.org/abs/2409.11546 (2024).
13. Notton, C. et al. Efficient self-supervised Barlow twins from limited tissue slide cohorts for colonic pathology diagnostics. Preprint at https://arxiv.org/abs/2411.05959 (2024).
14. Alfasly, S. et al. Foundation models for histopathology—fanfare or flair. *Mayo Clin. Proc. Digit. Health* **2**, 165–174 (2024).
15. Liu, Q. et al. M2Fusion: Bayesian-based multimodal multi-level fusion on colorectal cancer microsatellite instability prediction. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2023 Workshops (LNCS 14394)*, 125–134 (2024).
16. Ferber, D. et al. In-context learning enables multimodal large language models to classify cancer pathology images https://arxiv.org/abs/2403.07407 (2024).
17. Truhn, D., Eckardt, J. N., Ferber, D. & Kather, J. N. Large language models and multimodal foundation models for precision oncology. *NPJ Precis. Oncol.* **8**, 72 (2024).
18. El Nahhas, O. S. M. et al. Regression-based deep-learning predicts molecular biomarkers from pathology slides. *Nat. Commun.* **15**, 1253 (2024).
19. Lalinia, M. & Sahafi, A. Colorectal polyp detection in colonoscopy images using yolo-v8 network. *Signal Image Video Process.* **18**, 2047–2058 (2024).
20. Wan, J.-J., Zhu, P.-C., Chen, B.-L. & Yu, Y.-T. A semantic feature enhanced YOLOv5-based network for polyp detection from colonoscopy images. *Scientific Reports* **14**, 15478 (2024).
21. Gelu-Simeon, M. et al. Deep learning model applied to real-time delineation of colorectal polyps. *BMC Med. Inf. Decis. Mak.* **25** https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-025-03047-y (2025).
22. Liu, D., Lu, C., Sun, H. & Gao, S. Na-segformer: a multi-level transformer model based on neighborhood attention for colonoscopic polyp segmentation. *Sci. Rep.* **14**, 22527 (2024).
23. Xue, H., Yonggang, L., Min, L. & Lin, L. A lighter hybrid feature fusion framework for polyp segmentation. *Sci. Rep.* **14**, 23179 (2024).
24. Wang, M., Xu, C. & Fan, K. An efficient fine tuning strategy of segment anything model for polyp segmentation. *Sci. Rep.* **15**, 14088 (2025).
25. Wang, X. et al. Transformer-based unsupervised contrastive learning for histopathological image classification. *Med. image Anal.* **81**, 102559 (2022).
26. Dosovitskiy, A. et al. An image is worth 16x16 words: transformers for image recognition at scale. Preprint at https://arxiv.org/abs/2010.11929 (2021).
27. Liu, Z. et al. Swin transformer v2: Scaling up capacity and resolution. Preprint at https://arxiv.org/abs/2111.09883 (2022).
28. Hatamizadeh, A. et al. Swin unetr: swin transformers for semantic segmentation of brain tumors in MRI images. Preprint at https://arxiv.org/abs/2201.01266 (2022).
29. Chen, M., Wei, Z., Huang, Z., Ding, B. & Li, Y. Simple and deep graph convolutional networks. Preprint at https://arxiv.org/abs/2007.02133 2007.02133 (2020).
30. Liu, Z. et al. Efficient low-rank multimodal fusion with modality-specific factors. Preprint at https://arxiv.org/abs/1806.00064 (2018).
31. Gal, Y. & Ghahramani, Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. Preprint at https://arxiv.org/abs/1506.02142 (2016).
32. Fragoso, T. M., Bertoli, W. & Louzada, F. Bayesian model averaging: a systematic review and conceptual classification. *Int. Stat. Rev.* **86**, 1–28 (2017).
33. Kirk, S. et al. The Cancer Genome Atlas Colon Adenocarcinoma Collection (TCGA-CAAD)(version 3)[data set]. *Cancer Imaging Arch*. (2016).
34. Kirk, S., Lee, Y., Sadow, C. & Levine, S. The Cancer Genome Atlas rectum adenocarcinoma collection (tcga-read)(version 3)[data set]. *Cancer Imaging Arch*.
35. Shi, L. et al. Ebhi-seg: a novel enteroscope biopsy histopathological hematoxylin and eosin image dataset for image segmentation tasks. *Front. Med.* **10**, 1114673 (2023).
36. Biffi, C. et al. Real-colon: a dataset for developing real-world ai applications in colonoscopy. *Sci. Data* **11**, 539 (2024).
37. Jha, D. et al. Kvasir-seg: a segmented polyp dataset. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5-8, 2020, Proceedings, Part II*, 451-462 https://doi.org/10.1007/978-3-030-37734-2_37 (Springer-Verlag, 2020).
38. Ronneberger, O., Fischer, P. & Brox, T. U-net: convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241 (Springer, 2015).
39. Fan, D.-P. et al. Pranet: parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, 263–273 (Springer, 2020).
40. Chen, J. et al. Transunet: transformers make strong encoders for medical image segmentation. *CoRR* **abs/2102.04306.** Preprint at https://arxiv.org/abs/2102.04306 (2021).
41. Wang, Z. et al. Dual-stream multi-dependency graph neural network enables precise cancer survival analysis. *Med. Image Anal.* **97**, 103252 (2024).
42. Yang, P., Qiu, H., Yang, X., Wang, L. & Wang, X. Sagl: a self-attention-based graph learning framework for predicting survival of colorectal cancer patients. *Comput. Methods Programs Biomed.* **249**, 108159 (2024).
43. Zhu, Y. et al. Stg: spatiotemporal graph neural network with fusion and spatiotemporal decoupling learning for prognostic prediction of colorectal cancer liver metastasis. Preprint at https://arxiv.org/abs/2505.03123 (2025).
44. Li, W., Lin, S., He, Y., Wang, J. & Pan, Y. Deep learning survival model for colorectal cancer patients (deepcrc) with asian clinical data compared with different theories. *Arch. Med. Sci.* **19**, 264 (2023).
45. Qu, M. et al. Multimodal cancer survival analysis via hypergraph learning with cross-modality rebalance. Preprint at https://doi.org/10.48550/arXiv.2505.11997 (2025).
46. Yalçíner, M., Erdat, E. C., Kavak, E. E. & Utkan, G. Development and validation of an AI-augmented deep learning model for survival prediction in de novo metastatic colorectal cancer. *Discov. Oncol.* **16**, 2126 (2025).

Acknowledgements

We greatly appreciate our department colleagues for their invaluable comments and constructive feedback that substantially improved the quality of this manuscript.

Author contributions

R.S., J.S., and Z.Z. contributed equally to this work, having full access to all study data and assuming responsibility for the integrity and accuracy of the analyses (Validation, Formal analysis). R.S. conceptualized the study and designed the methodology (Conceptualization, Methodology). J.S. carried out data acquisition, curation, and investigation (Investigation, Data curation) and provided key resources, instruments, and technical support (Resources, Software). Z.Z. and Q.S. drafted the initial manuscript and generated visualizations (Writing—Original Draft, Visualization). Y.S. supervised the project, coordinated collaborations, ensured administrative support, and participated in securing research funding (Supervision, Project administration, Funding acquisition). All authors contributed to

reviewing and revising the manuscript critically for important intellectual content (Writing—Review & Editing) and approved the final version for submission.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Yongqian Shu.