ARTICLE Artificial Intelligence

OPEN ACCESS

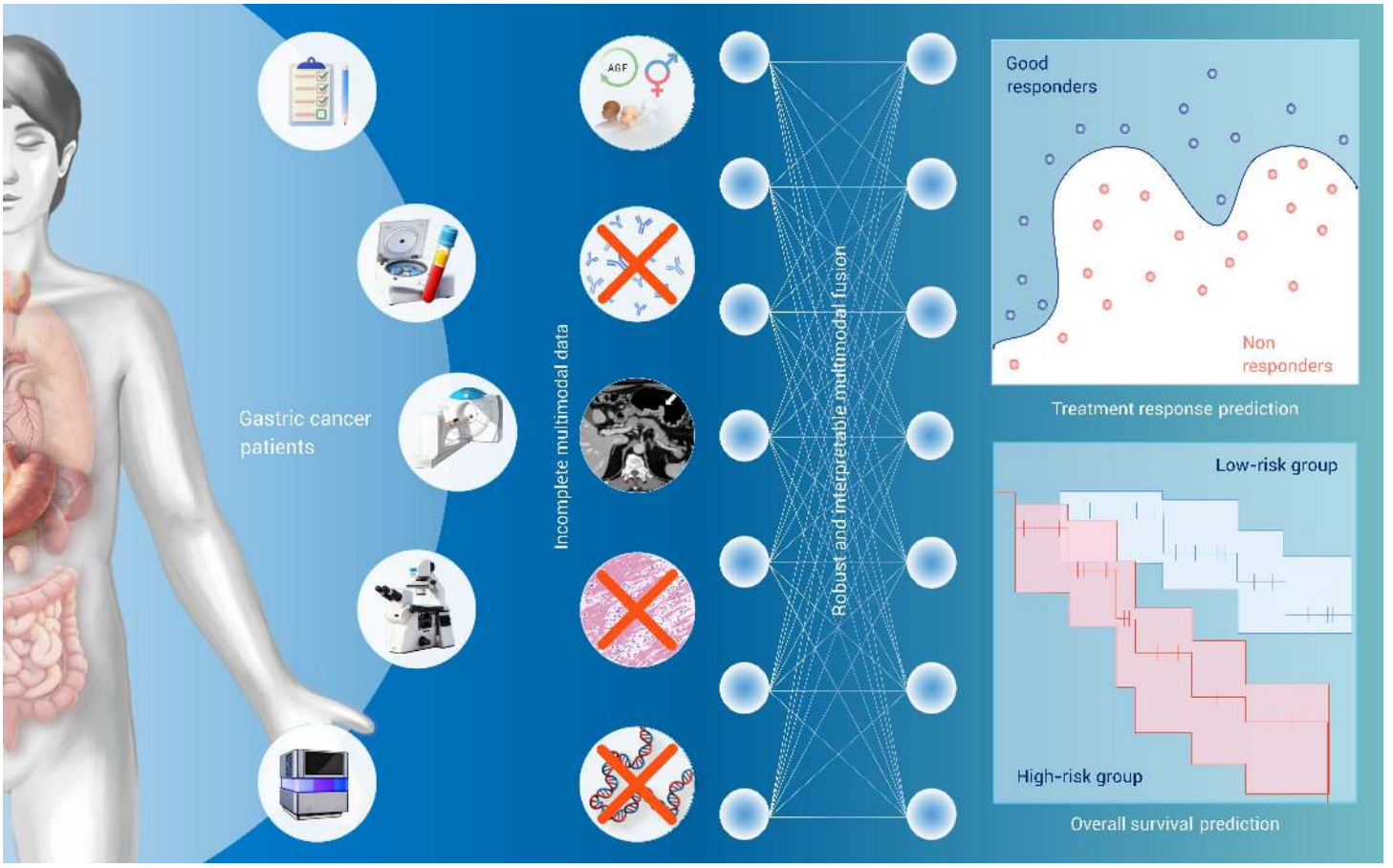
Interpretable and robust multimodal data integration for precise treatment response and survival prediction in gastric cancer

Fengtao Zhou,[1] Yingxue Xu,[1] Yanfen Cui,[2] Shenyan Zhang,[3] Yun Zhu,[4] Weiyang He,[5] Dafu Zhang,[6] Jiguang Wang,[7] Xin Wang,[8] Ronald CK Chan,[9] Louis Ho Shing Lau,[10] Chu Han,[11] Zhenhui Li,[6,*] and Hao Chen[1,7,12,13,14,*]

*Correspondence: lizhenhui@kmmu.edu.cn (Z.L.); jhc@cse.ust.hk (H.C.)

Received: September 16, 2025; Accepted: October 27, 2025; Published Online: November 6, 2025; https://doi.org/10.59717/j.xinn-inform.2025.100011

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- This study proposes an interpretable and robust multimodal fusion framework for gastric cancer analysis.
- This model robustly integrates multiple data sources to predict treatment response and patient survival.
- This model shows reliable performance across multiple datasets and can handle incomplete data scenarios.
- Its interpretability identifies key factors, providing trustworthy insights for treatment personalization.

OPEN ACCESS

Artificial Intelligence ARTICLE

The Innovation Informatics

Interpretable and robust multimodal data integration for precise treatment response and survival prediction in gastric cancer

Fengtao Zhou,[1] Yingxue Xu,[1] Yanfen Cui,[2] Shenyan Zhang,[3] Yun Zhu,[4] Weiyang He,[5] Dafu Zhang,[6] Jiguang Wang,[7] Xin Wang,[8] Ronald CK Chan,[9] Louis Ho Shing Lau,[10] Chu Han,[11] Zhenhui Li,[6,*] and Hao Chen[1,7,12,13,14,*]

[1]Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR 999077, China
[2]Department of Radiology, Shanxi Cancer Hospital/Shanxi Hospital Affiliated to Cancer Hospital, Taiyuan 030013, China
[3]Department of Pathology, The Sixth Affiliated Hospital of Sun Yat-sen University, Guangzhou 510655, China
[4]Department of Radiology, The First Affiliated Hospital of Kunming Medical University, Kunming 650032, China
[5]Department of Gastrointestinal Surgery, Sichuan Province Cancer Hospital, University of Electronic Science and Technology of China, Chengdu 610054, China
[6]Department of Radiology, The Third Affiliated Hospital of Kunming Medical University, Yunnan Cancer Hospital, Kunming 650118, China
[7]Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR 999077, China
[8]Department of Surgery, Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong SAR 999077, China
[9]Department of Anatomical and Cellular Pathology, Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong SAR 999077, China
[10]Department of Medicine & Therapeutics, Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong SAR 999077, China
[11]Department of Radiology, Guangdong Provincial People's Hospital, Southern Medical University, Guangzhou 510000, China
[12]Division of Life Science, The Hong Kong University of Science and Technology, Hong Kong SAR 999077, China
[13]HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Shenzhen 518000, China
[14]State Key Laboratory of Nervous System Disorders, The Hong Kong University of Science and Technology, Hong Kong SAR 999077, China
*Correspondence: lizhenhui@kmmu.edu.cn (Z.L.); jhc@cse.ust.hk (H.C.)

Received: September 16, 2025; Accepted: October 27, 2025; Published Online: November 6, 2025; https://doi.org/10.59717/j.xinn-inform.2025.100011

Citation: Zhou F., Xu Y., Cui Y., et al. (2025). Interpretable and robust multimodal data integration for precise treatment response and survival prediction in gastric cancer. *The Innovation Informatics* **1**:100011.

Gastric cancer (GC) was the fifth most prevalent cancer in 2022. Integrating multimodal data enables significant improvement in precise GC analysis. However, the heterogeneity and incompleteness of multimodal data pose challenges for effective integration. The lack of interpretability in models hinders their trustworthiness in clinical practice. Therefore, this study presented an interpretable and robust multimodal data integration framework for GC analysis (iMD4GC), facilitating effective integration of diverse data sources to predict treatment response and overall survival. Three multimodal datasets were collected for model evaluation: GastricRes (n=698) for response prediction, GastricSur (n=801) for survival analysis, and TCGA-STAD (n=400) for further survival analysis. The iMD4GC achieved an area under the receiver operating characteristic curve (AUC) of 0.803 (95% CI 0.680-0.925) on GastricRes, a concordance index (c-index) of 0.715 (95% CI 0.646-0.784) and a time-dependent AUC of 0.739 (95% CI 0.644-0.833) on GastricSur, and a c-index of 0.661 (95% CI 0.504-0.818) and a time-dependent AUC of 0.690 (95% CI 0.500-0.881) on TCGA-STAD. The comparative analysis underscores the reliability of iMD4GC in integrating diverse multimodal data and addressing missing modalities. Further interpretability analysis identified pivotal factors across various data modalities, validating the alignment with clinical expertise and providing essential insights for informed decision-making in clinical contexts. The promising performance of iMD4GC highlights its potential in advancing clinical decision-making, personalizing treatment strategies, and improving patient prognosis in GC. Inherent interpretability fosters transparent decision analysis and provides valuable insights for clinicians, promoting its trustworthiness and acceptance in clinical settings.

INTRODUCTION

Gastric cancer (GC) represents a significant global health challenge as the fifth most common malignancy in terms of both incidence and mortality rates. According to world cancer statistics reported by the World Health Organization in 2022,[1] there were 968,350 new cases and 659,853 deaths worldwide. The asymptomatic nature of GC contributes to a delayed diagnosis, with approximately two-thirds of all cases being identified only at advanced stages.[2,3] To improve treatment outcomes for patients with locally advanced GC (LAGC), neoadjuvant chemotherapy (NACT) has emerged as a promising pre-operative therapeutic strategy. NACT aims to reduce tumor size and prevent their spread, thereby facilitating subsequent surgical interventions (Figure 1A).[4,5] However, the response to NACT treatment varies among LAGC patients, with a notable proportion exhibiting resistance.[6,7] Ineffective NACT treatment will result in adverse side effects as well as the loss of an optimal treatment period, ultimately affecting survival rates. Therefore, accurate predictions of treatment response and overall survival are essential for personalizing treatment strategies and optimizing therapeutic outcomes for LAGC patients.

Recent advancements in artificial intelligence (AI) have significantly enhanced the capabilities of doctors in diagnosing cancer,[8-11] predicting treatment response,[12-14] and analyzing survival prognosis[15-17] for GC patients. In the specific domain of NACT response prediction, a growing body of research has demonstrated the potential of AI. For instance, several studies have developed radiomics models based on CT images to predict pathological complete response in gastric cancer.[18-20] Similarly, in other cancer types where NACT is common, such as breast cancer, deep learning models have achieved remarkable success in predicting response from ultra-sound images,[21,22] multi-parametric MRI,[23] and pathology images.[24] However, a common limitation among many of these prior studies is their primary reliance on a single data modality for predictions. While these approaches have shown promise, they fail to fully capture the multifaceted nature of cancer, often resulting in suboptimal performance and limited generalizability.

Throughout the medical trajectory of GC patients, clinicians will gather a diverse range of medical data to make informed decisions and devise personalized treatment strategies. Specifically, clinical information contains patient demographics, laboratory findings, medical histories, and other diagnosis results. Radiological imaging, in turn, provides pivotal insights into tumor morphology and spatial distribution from a macroscopic perspective. Histopathological examination, characterized by detailed analyses of cellular and tissue characteristics, offers microscopic insights into the structural and compositional attributes of malignancy. Molecular data, such as transcriptome profiles, can further elucidate the underlying biological mechanisms of GC. Integrating these diverse data modalities can provide a more comprehensive understanding of GC patients, enabling AI models to capture the intricate interactions between different aspects of the disease. This holistic approach empowers AI models to make more accurate diagnoses, predict treatment responses, and improve survival predictions for patients with GC.

While some researchers have proposed integrating multimodal data for GC analysis,[25-27] these endeavors have often focused on a limited set of modalities, such as demographics and radiomics features. The extraction of radiomics features necessitates the manual delineation of the region of interest (ROI). The integration of radiological images with other data modalities remains underexplored. Incorporating additional modalities could provide supplementary insights, enhancing prediction accuracy. However, the inherent

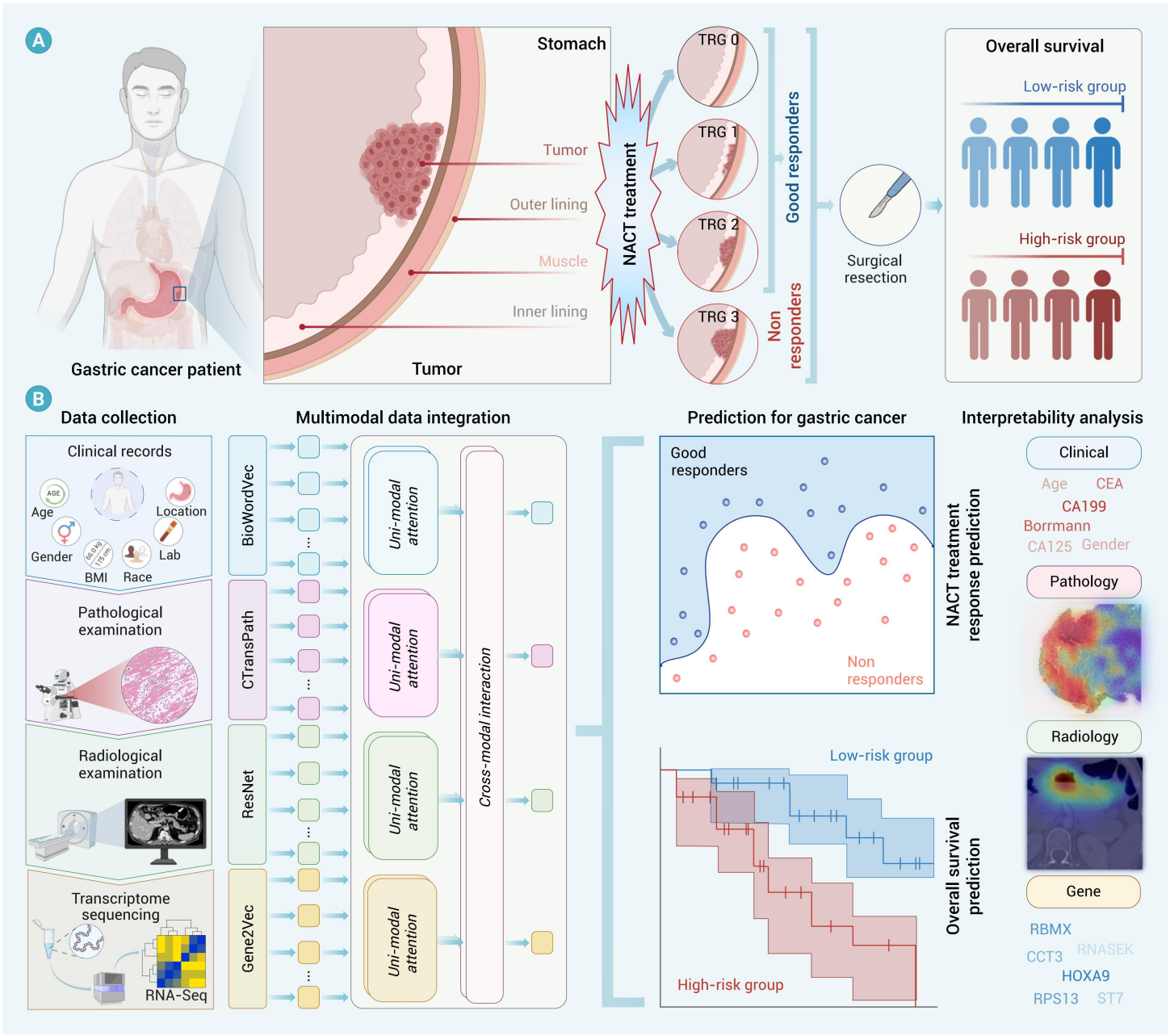


Figure 1. Overview of integrating multimodal data to predict treatment response and overall survival for GC patients (A) The LAGC patients undergo NACT treatment first, which is expected to reduce tumor size and extent. Then, the patients receive surgical resection to remove the remaining tumor. (B) The proposed multimodal fusion framework integrates clinical records, pathology images, radiology images, and transcriptome profiles for predictions of NACT treatment response and overall survival in GC patients. The interpretability analysis is conducted to provide insights into the model's decision-making process, enhancing trustworthiness and acceptance in clinical practice.

heterogeneity of multimodal data poses significant challenges for seamless integration and thorough analysis. Furthermore, current studies typically assume that all modalities are readily available for GC analysis, which may not reflect real-world clinical scenarios where data incompleteness can arise from factors like corruption, privacy concerns, technical limitations, or other constraints. Moreover, the lack of interpretability in deep learning models poses a critical barrier to their trustworthiness and acceptance in clinical practice.

To address these challenges, this study presented an interpretable and robust multimodal data integration framework for GC studies (iMD4GC) to predict NACT response and overall survival via heterogeneous and incomplete multimodal data, including clinical information, pathology images, radiology images, and transcriptomics profiles (Figure 1B). Specifically, iMD4GC employs specific data tokenization and embedding strategies to encode the multimodal data into a unified representation space, mitigating the heterogeneity of data sources and boosting the fusion of modalities. Subsequently, the framework introduces unimodal attention layers and cross-modal interaction layers to capture intra-modal information and explore potential inter-modal interactions, respectively. This enables the framework to explore complementary information across modalities, thereby compensating for missing modalities and addressing the challenges posed by incomplete multimodal data. To enhance the robustness of iMD4GC in handling severely incomplete multimodal data, knowledge distillation is further introduced, distilling the knowledge learned from all available modalities to reduced sets.

MATERIALS AND METHODS

Problem formulation

Let $\mathbb{X} = \{X_1, X_2, \cdots, X_N\}$ denote the set of all cases diagnosed with GC where N represents the total number of cases. Each individual case can be described by a 4-tuple $X_i = \{\mathcal{C}_i, \mathcal{R}_i, \mathcal{P}_i, \mathcal{G}_i\}$, in which $\mathcal{C}_i$, $\mathcal{R}_i$, $\mathcal{P}_i$, and $\mathcal{G}_i$ correspond to the clinical information, radiology image (CT scans), pathology image (WSIs), and transcriptomics profiles (RNA-seq), respectively. Let $\mathbb{Y} = \{Y_1, Y_2, \cdots, Y_N\}$ denote the set of all labels, where Y_i represents the label of X_i. This study encompasses two primary tasks related to GC: NACT

response prediction and survival prediction. In the context of NACT response prediction, the label Y takes binary values, specifically 0 to indicate the non-responder and 1 to indicate good responder. The main objective is to estimate the response probability $\hat{Y}$. For survival prediction, the label Y is represented as $Y=(c,t_{os})$, where $c\in\{0,1\}$ signifies the right uncensored status, and $t_{os}\in R^{+}$ represents the overall survival time measured in months. In this case, our primary goal is to estimate the hazard function $f_{hazard}(T=t|T\geq t,X)\in[0,1]$. The hazard function measures the instantaneous risk or death event occurring at a specific time point t. Instead of directly estimating t_{os}, survival models output an ordinal risk value obtained by leveraging the cumulative distribution function $f_{surv}(T\geq t,X)=\prod_{u=1}^{t}(1-f_{hazard}(T=t|T\geq t,X))$.

In practical scenarios, it is common to encounter incomplete modality cases where certain modalities are missing for specific patients. For instance, $X=\{\mathcal{C},\mathcal{R},\mathcal{P}\}$ indicates the absence of transcriptomics profiles, while $X=\{\mathcal{C},\mathcal{R}\}$ suggests the lack of both pathological images and transcriptomics profiles. In such cases, the multimodal learning model F is expected to make accurate predictions based on the available modalities. The multimodal learning model F should possess the capability to handle both complete and incomplete cases. In this study, we propose a novel multimodal learning framework that effectively integrates all available modalities within X to estimate the response probability $\hat{Y}$ and the hazard function f_{hazard} for both complete and incomplete cases.

Patient cohorts

For the analysis of GC, we assembled three multimodal datasets from multiple hospitals and public databases. In the context of predicting response to NACT, patients with LAGC who received NACT followed by surgical resection were retrospectively enrolled from four medical centers (Yunnan Cancer Hospital, Shanxi Cancer Hospital, Sichuan Cancer Hospital, and the Sixth Affiliated Hospital of Sun Yat-sen University). Pathological response to NACT was evaluated using the Tumor Regression Grade (TRG) system, with patients categorized as good responders (TRG 0-2) or non-responders (TRG 3). The dataset integrates three types of data: clinical information, histopathological whole-slide images (WSIs), and radiological computed tomography (CT) scans. Owing to variations in clinical practice and data collection protocols, not all patients contributed complete data across all three modalities, resulting in an inherently incomplete multimodal dataset. In this study, clinical information was treated as the baseline modality, while CT scans and WSIs were considered optional modalities that could be missing for certain patients. Individuals lacking serological test results were excluded. Ultimately, a total of 698 patients were included in the GastricRes dataset for NACT response prediction.

For survival prediction, GC patients who underwent surgical resection were recruited from two independent hospitals (the First Affiliated Hospital of Kunming Medical University and Yunnan Cancer Hospital). Patients with clinical follow-up from December 2012 to June 2020 were included. Those without serological test results were excluded. Overall survival (OS) was recorded for each patient. Similar to the NACT response cohort, not all patients had complete multimodal data, leading to another inherently incomplete multimodal dataset. A total of 801 patients were included in the GastricSur dataset for survival analysis. During the follow-up period, 288 deaths were recorded, while 513 patients were censored. To further assess the generalizability of our model, we also incorporated the TCGA-STAD dataset from The Cancer Genome Atlas for survival prediction. This dataset comprises clinical information, WSIs, and transcriptomic RNA-Seq data. A total of 400 GC patients were included in the TCGA-STAD dataset. During the follow-up period, 162 patients experienced death, while 238 patients were censored. Table 1 provides a summary of the patient demographics and hospital characteristics for these datasets.

Data processing and tokenization

The overall framework of our proposed multimodal learning model is illustrated in Figure 2A. This model consists of four primary components: data tokenization, unimodal attention layers, cross-modal interaction layers, and multimodal feature fusion. To facilitate subsequent modelling tasks, it is necessary to convert heterogeneous multimodal data into a unified format. To this end, we employ different tokenization strategies for different modalities to transform the multimodal data into a set of tokens. Specifically, clinical information $\mathcal{C}$ comprise patient-specific information. These tabular records encompass a diverse array of data types, including discrete, continuous, and collective data. Each clinical record in $\mathcal{C}$ is represented as a 2-tuple (k,v) with k denoting the record's name and v signifying the corresponding value. To explore semantic correlations among these records, we employ BioWordVec to obtain d dimensional word embeddings for each record,[28] with d representing the embedding dimension. Furthermore, a learnable embedding network is utilized to map the records' values to d dimensional vectors. By integrating word embeddings and value embeddings, clinical records can be tokenized into a set $\mathcal{C}=\{c_1,c_2,\cdots,c_k\}$, with each $c_i\in R^d$ representing the sum of word embedding and value embedding of the i-th record.

The radiology images $\mathcal{R}$ are represented as 3D-CT images. However, the stomach only occupies a small portion of the image, with surrounding organs and tissues being irrelevant to prediction tasks. To eliminate extraneous information and focus on the stomach region, we employ TotalSegmentator to locate and crop the stomach region from the entire volume.[29] The intensity values of the cropped volume are then normalized to the range of [0, 1] using min-max normalization (HU values are clipped to the range of $[-100, 200]$ before normalization). Each slice is then fed into ResNet pre-trained on ImageNet to obtain d dimensional image embedding.[30] To maintain spatial correlation among slices, we follow Transformer to generate positional embeddings for each slice.[31] The radiology image can be tokenized into a set $\mathcal{R}=\{r_1,r_2,\cdots,r_m\}$, where each $r_i\in R^d$ represents the sum of image embedding and positional embedding of the i-th slice.

In computational pathology, whole slide images (WSIs) are typically represented as a bag data structure due to their ultra-high resolution. Following this widely used setting, we crop each WSI into a series of non-overlapping 512×512 patches at $40\times$ magnification level. Subsequently, each patch is fed into CTransPath to obtain d-dimensional patch embeddings.[32] Then, pathology image can be tokenized into a set $\mathcal{P}=\{p_1,p_2,\cdots,p_n\}$, with $p_i\in R^d$ representing the embedding of the i-th patch. Similar to clinical records, the transcriptomics profiles can also be regarded as a kind of tabular data, in which each element represents the expression level of a specific gene. The expression value of each gene is normalized by log2(TPM+1) transformation. To investigate potential co-expression among genes, we adopt Gene2Vec to generate d-dimensional gene embeddings for each gene.[33] For expression levels, there is a learnable embedding network to map them to d-dimensional vectors. By integrating gene embeddings and expression embeddings, genomic profiles can be tokenized into a set $\mathcal{G}=\{g_1,g_2,\cdots,g_l\}$, with each $g_i\in R^d$ representing the sum of gene embedding and expression embedding of the i-th gene.

Network architecture

First, we leverage unimodal attention layers tailored for each modality, enabling the extraction of discriminative representations from the tokenized data. The formulation of the unimodal attention layer can be expressed as follows,

$$\begin{aligned}
\left(c_*^{(1)},c_1^{(1)},c_2^{(1)},\cdots,c_k^{(1)}\right)&=F_{uni}^{c}\left(c_*^{(0)}\,||\,\mathcal{C}^{(0)}\right)\\
\left(r_*^{(1)},r_1^{(1)},r_2^{(1)},\cdots,r_k^{(1)}\right)&=F_{uni}^{r}\left(r_*^{(0)}|\mathcal{R}^{(0)}\right)\\
\left(p_*^{(1)},p_1^{(1)},p_2^{(1)},\cdots,p_k^{(1)}\right)&=F_{uni}^{p}\left(p_*^{(0)}|\mathcal{P}^{(0)}\right)\\
\left(g_*^{(1)},g_1^{(1)},g_2^{(1)},\cdots,g_k^{(1)}\right)&=F_{uni}^{g}\left(g_*^{(0)}|\mathcal{G}^{(0)}\right)
\end{aligned}$$

where $\mathcal{C}^{(0)}$, $\mathcal{R}^{(0)}$, $\mathcal{P}^{(0)}$, and $\mathcal{G}^{(0)}$ represent the initial tokens of each modality. $c_*^{(0)}$, $r_*^{(0)}$, $p_*^{(0)}$, and $g_*^{(0)}$ denote the learnable class tokens defined for clinical records, radiology images, pathology images, and genomic profiles, respectively. These class tokens serve as placeholders when a particular modality is missing, facilitating subsequent cross-modal information aggregation.

Following the conventional layers in Transformer, the unimodal attention layer consists of several primary components: a multi-head self-attention module, a feed-forward network, and layer normalization. This configuration allows the model to effectively capture and process relationships within each modality. However, due to the substantial sequence length of pathology and genomic tokens, the native self-attention mechanism becomes computationally expensive and memory-intensive. To address this challenge, we

Table 1. Patient demographics and hospital characteristics of the collected datasets.

Clinical Information		GastricRes (*n*=698)	GastricSur (*n*=801)	TCGA-STAD (*n*=400)
Gender	Female	179 (25.6%)	260 (32.5%)	137 (34.3%)
	Male	519 (74.4%)	541 (67.5%)	263 (65.8%)
Age	≤65	548 (78.5%)	647 (80.8%)	183 (45.8%)
	>65	150 (21.5%)	154 (19.2%)	211 (52.8%)
	N/A	-	-	6 (1.5%)
Differentiation	Well	16 (2.3%)	28 (3.5%)	-
	Moderately	92 (13.2%)	246 (30.7%)	-
	Poorly	367 (52.6%)	601 (75.0%)	-
	N/A	223 (31.9%)	48 (8.0%)	-
Borrmann	Elevated	65 (9.3%)	62 (7.7%)	-
	Ulcerative	361 (51.7%)	114 (14.2%)	-
	Infiltrative	171 (24.5%)	369 (46.1%)	-
	Diffuse	98 (14.0%)	167 (20.8%)	-
	N/A	3 (0.4%)	89 (11.1%)	-
Lauren	Intestinal	119 (17.0%)	172 (21.5%)	-
	Diffuse	133 (19.1%)	289 (36.1%)	-
	Mixed	177 (25.4%)	162 (20.2%)	-
	Other	11 (1.6%)	1 (0.1%)	-
	N/A	258 (37.0%)	177 (22.1%)	-
CEA	≤5	497 (71.2%)	610 (76.2%)	-
	>5	161 (23.1%)	123 (15.4%)	-
	NA	40 (5.7%)	68 (8.5%)	-
CA199	≤20	455 (65.2%)	533 (66.5%)	-
	>20	201 (28.8%)	160 (20.0%)	-
	NA	42 (6.0%)	108 (13.5%)	-
CA125	≤35	523 (74.9%)	389 (48.6%)	-
	>35	29 (4.2%)	14 (1.7%)	-
	NA	146 (20.9%)	398 (49.7%)	-
Clinical T Stage	T1	4 (0.6%)	-	19 (4.8%)
	T2	24 (3.4%)	-	87 (21.8%)
	T3	279 (40.0%)	-	187 (46.8%)
	T4	388 (55.6%)	-	102 (25.5%)
	Other	1 (0.1%)	-	-
	N/A	2 (0.3%)	-	5 (1.2%)
Clinical N Stage	N0	103 (14.8%)	-	119 (29.8%)
	N1	142 (20.3%)	-	112 (28.0%)
	N2	239 (34.2%)	-	79 (19.8%)
	N3	195 (27.9%)	-	80 (20.0%)
	Other	17 (2.4%)	-	-
	N/A	2 (0.3%)	-	10 (2.5%)
Clinical M Stage	M0	630 (90.3%)	-	360 (90.0%)
	M1	68 (9.7%)	-	26 (6.5%)
	N/A	-	-	14 (3.5%)
NACT Response	TRG0 (Good)	57 (8.2%)	-	-
	TRG1 (Moderate)	55 (7.9%)	-	-
	TRG2 (Minimal)	213 (30.5%)	-	-
	TRG3 (Poor)	373 (53.4%)	-	-
Overall Survival	Death	-	288 (36.0%)	162 (40.5%)
	Censored	-	513 (64.0%)	238 (59.5%)

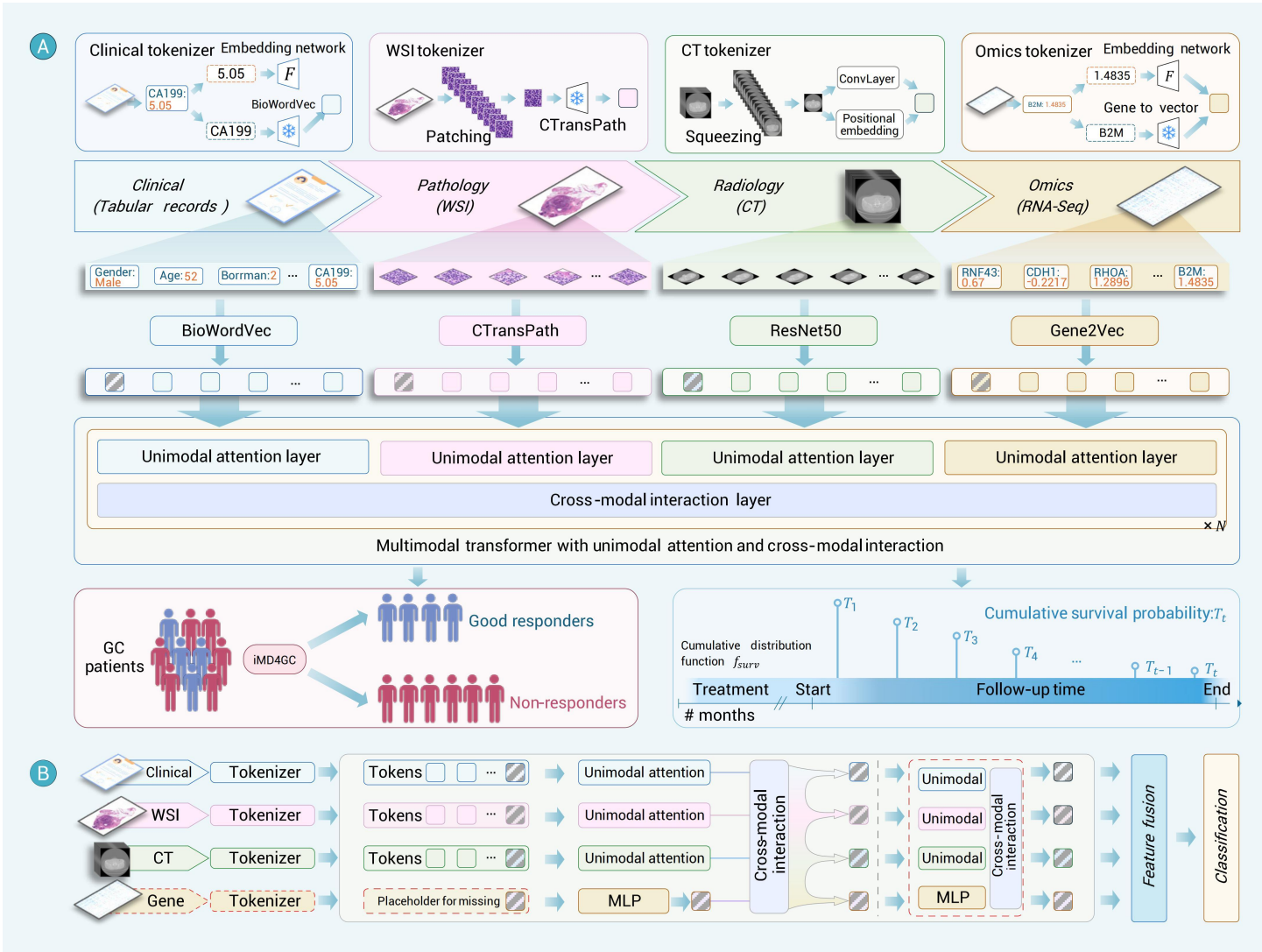


Figure 2. Overall framework (A) The pipeline of our proposed multimodal learning framework for NACT response prediction and survival analysis. (B) The pipeline of proposed framework using incomplete multimodal data, which explores related information contained in the available modalities and compensates for the missing modalities.

adopt the linear attention mechanism proposed in Nystromer, which can effectively reduce the time and space complexity of self-attention while maintaining the overall performance.

The inclusion of unimodal attention layers in our model allows for the extraction of discriminative representations from each modality. However, these representations are inherently limited to their respective modalities and fail to capture the intricate interactions that exist among different modalities. To address this limitation and harness the unique strengths of each modality while integrating information from multiple sources, we propose the incorporation of cross-modal interaction layers in our framework. These interaction layers follow the unimodal attention layers and enable the exploration of inter-modal interactions and the aggregation of inter-modal information. To ensure the integrity of the information and mitigate any contamination arising from the heterogeneity of multimodal data, we leverage the class tokens $c_*^{(1)}$, $r_*^{(1)}$, $p_*^{(1)}$, $g_*^{(1)}$ as the query tokens. These class tokens serve as the information bridge for exploring inter-modal interactions and aggregating inter-modal information. The formulation of the cross-modal interaction layer can be expressed as follows,

$$\begin{aligned}
\widehat{c}_*^{(1)} &= F_{cross}^{c}\left(c_*^{(1)} \,||\, \mathcal{R}^{(1)} \,||\, \mathcal{P}^{(1)} \,||\, \mathcal{G}^{(1)} \,||\right) \\
\widehat{r}_*^{(1)} &= F_{cross}^{r}\left(r_*^{(1)} \,||\, \mathcal{C}^{(1)} \,||\, \mathcal{P}^{(1)} \,||\, \mathcal{G}^{(1)} \,||\right) \\
\widehat{p}_*^{(1)} &= F_{cross}^{p}\left(p_*^{(1)} \,||\, \mathcal{C}^{(1)} \,||\, \mathcal{R}^{(1)} \,||\, \mathcal{G}^{(1)} \,||\right) \\
\widehat{g}_*^{(1)} &= F_{cross}^{g}\left(g_*^{(1)} \,||\, \mathcal{C}^{(1)} \,||\, \mathcal{R}^{(1)} \,||\, \mathcal{P}^{(1)} \,||\right)
\end{aligned}$$

where $\mathcal{C}^{(1)}$, $\mathcal{R}^{(1)}$, $\mathcal{P}^{(1)}$, and $\mathcal{G}^{(1)}$ represent the outputs of the previous unimodal attention layers. It is worth noting that we only need to calculate the attention between the class token and the other tokens, which significantly reduces the computational complexity compared to the unimodal attention layer. Hence, we can directly employ the naive multi-head attention mechanism. By incorporating both types of attention layers, our model can generate comprehensive representations that encompass the unique characteristics of each modality and the intermodal relationships between them.

When certain modalities are unavailable, we leverage the power of cross-modal interaction layers to effectively compensate for the missing modalities. As depicted in Figure 2B, the class token serves as a placeholder when a specific modality is missing. In such cases, the corresponding unimodal attention layer is degraded to a linear layer. Consequently, the placeholder is seamlessly integrated into the cross-modal interaction layer, enabling the capture of complementary information from the available modalities. This process empowers our model to compensate for the missing modalities, thereby preserving the accuracy and reliability of its predictions. By leveraging the available information and exploiting the intricate relationships among different modalities, our framework could surmount the limitations imposed by incomplete multimodal data, ultimately enhancing the robustness and efficacy of our predictions.

Knowledge distillation

To further enhance the performance of our framework when dealing with severely incomplete multimodal data, we introduce a knowledge distillation, which aims to distil knowledge from all available modalities to the reduced

Table 2. Quantitative results for the predictions of treatment response and overall survival on the GastricRes, GastricSur, and TCGA-STAD datasets, respectively. The above values represent the mean performance of each method on the corresponding dataset. The below values represent 95% confidence interval of the corresponding metrics.

Method	GastricRes			GastricSur		TCGA-STAD	
	AUC	Accuracy	F1-Score	C-Index	AUC	C-Index	AUC
MFN	0.768 (0.654-0.882)	0.742 (0.642-0.842)	0.709 (0.563-0.855)	0.617 (0.523-0.710)	0.614 (0.487-0.740)	0.630 (0.500-0.760)	0.665 (0.494-0.836)
MulT	0.761 (0.639-0.882)	0.733 (0.624-0.841)	0.677 (0.516-0.839)	0.579 (0.461-0.696)	0.566 (0.425-0.708)	0.599 (0.445-0.752)	0.628 (0.446-0.809)
MMD	0.742 (0.600-0.884)	0.721 (0.610-0.831)	0.668 (0.486-0.851)	0.608 (0.494-0.721)	0.635 (0.485-0.785)	0.592 (0.450-0.735)	0.608 (0.413-0.803)
Performer	0.749 (0.621-0.877)	0.738 (0.630-0.846)	0.695 (0.548-0.841)	0.623 (0.532-0.715)	0.614 (0.492-0.736)	0.618 (0.489-0.748)	0.646 (0.477-0.815)
Nystromer	0.775 (0.663-0.886)	0.744 (0.647-0.842)	0.719 (0.582-0.855)	0.652 (0.564-0.741)	0.664 (0.531-0.796)	0.637 (0.488-0.787)	0.649 (0.452-0.845)
Ours	**0.803** (0.680-0.925)	**0.759** (0.662-0.855)	**0.732** (0.607-0.858)	**0.715** (0.646-0.784)	**0.739** (0.644-0.833)	**0.661** (0.504-0.818)	**0.690** (0.500-0.881)

sets (power set of X). The strategy consists of three steps: (1) training a teacher model F_t using all available modalities, (2) constructing a student model F_s that only utilizes a subset of modalities as input, and (3) transferring knowledge from the teacher model to the student model. Within this work, two types of knowledge distillation are employed: feature-level distillation and response-level distillation (logits-level). Feature-level distillation ensures that the student model generates representations similar to those of the teacher model, while response-level distillation focuses on ensuring that the student model produces predictions comparable to those of the teacher model. The knowledge distillation process can be formulated as follows:

$$\mathcal{L}_{fea} = \sum_{S \in P^*(X)} D_{KL}\left(F_X, F_S\right)$$
$$\mathcal{L}_{res} = \sum_{S \in P^*(X)} D_{KL}\left(\widehat{Y}_X, \widehat{Y}_S\right)$$

where $S \in P^*(X)$ represents the power set of X excluding the empty set. D_{KL} denotes the Kullback-Leibler divergence loss function. F_X and F_S are the representations generated by the teacher model and the student model, respectively. $\widehat{Y}_X$ and $\widehat{Y}_S$ denote the predictions made by the teacher model and the student model, respectively. To mitigate heavy computation requirements and reduce the risk of overfitting, we employ an exponential moving average (EMA) to update the parameters of the teacher model.[34] Additionally, the student model is updated while being constrained by classification loss. Incorporating classification loss with knowledge distillation loss, the student model can learn from the teacher model and achieve comparable performance even in scenarios where more modalities are missing. This strategy enhances the effectiveness of our framework in handling severely incomplete multimodal data.

RESULTS

Experimental design and statistical analysis

We employed 5-fold cross-validation on the collected datasets to evaluate the performance of our framework. To estimate 95% confidence intervals for the model performance on each fold, we conducted non-parametric bootstrapping with 1000 bootstrap replicates. To demonstrate the effectiveness and superiority of our framework, we reproduced various models for comparison, including MFN,[35] MulT,[36] MMD,[37] Performer,[38] and Nystromer.[39] These baseline methods can take incomplete multimodal data as input during both training and testing phases, ensuring a fair comparison with our model. For the prediction of treatment response, we utilized three evaluation metrics: AUC (area under the receiver operating characteristic curve), accuracy, and F1-score, in which AUC is the primary metric. For survival analysis, we employed the C-index (concordance index) and time-dependent AUC[40,41] as the evaluation metrics. The C-index measures the concordance between predicted survival probabilities and actual survival outcomes, while the time-dependent AUC evaluates the model's performance at different time points.

Quantitative analysis on NACT response prediction

The experimental results of various methods are summarized in Table 2. The heterogeneity and incompleteness of multimodal data would disrupt the synergistic interaction among different modalities, leading to the loss of valuable information and subsequent performance degradation. For instance, the MMD method only yielded an AUC of 0.742 (95% CI 0.600-0.884), an accuracy of 0.721 (95% CI 0.610-0.831), and a F1-score of 0.668 (95% CI 0.486-0.851) on the GastricRes dataset. The Performer, known for its superior performance in capturing long-range dependencies and its ability to handle incomplete multimodal data, achieved an AUC of 0.749 (95% CI 0.621-0.877), an accuracy of 0.738 (95% CI 0.630-0.846), and a F1-score of 0.695 (95% CI 0.548-0.841). Despite the existence of dedicated methods for handling incomplete multimodal data, their performance often falls short of meeting the requirements of clinical practice due to the impact of heterogenous multimodal data, as seen with methods like MFN (AUC: 0.768, 95% CI 0.654-0.882; Accuracy: 0.742, 95% CI 0.642-0.842; F1-score: 0.709, 95% CI 0.563-0.855) and MulT (AUC: 0.761, 95% CI 0.639-0.882; Accuracy: 0.733, 95% CI 0.624-0.841; F1-score: 0.677, 95% CI 0.516-0.839). In contrast, our framework demonstrated superior performance, achieving an AUC of 0.803 (95% CI 0.680-0.925), an accuracy of 0.759 (95% CI 0.662-0.855), and an F1-score of 0.732 (95% CI 0.607-0.858), markedly surpassing other methods in NACT response prediction. Compared to the second-best method, Nystromer, our framework showed superiority by 2.8% in AUC, 1.5% in accuracy, and 1.3% in F1-score. These results substantiated the effectiveness of our proposed framework in predicting the NACT response, particularly when dealing with incomplete multimodal data. The ROC curves for each fold and the mean ROC curve are visually represented in Figure 3A.

Quantitative analysis on overall survival prediction

Similar to the observations made in NACT response prediction, heterogeneity and incompleteness of multimodal data will also significantly affect the prediction of multimodal learning models on survival prediction tasks. The MMD method only achieved a C-index of 0.608 (95% CI 0.494-0.721) and a time-dependent AUC of 0.635 (95% CI 0.485-0.785) on the GastricSur dataset, and a C-index of 0.592 (95% CI 0.450-0.735) and a time-dependent AUC of 0.608 (95% CI 0.413-0.803) on the TCGA-STAD dataset. And the Performer obtained a C-index of 0.623 (95% CI 0.532-0.715) and a time-dependent AUC of 0.614 (95% CI 0.492-0.736) on the GastricSur dataset, and a C-index of 0.618 (95% CI 0.489-0.748) and a time-dependent AUC of 0.646 (95% CI 0.477-0.815) on the TCGA-STAD dataset. In contrast, our framework exhibited a C-index of 0.715 (95% CI 0.646-0.784) and a time-dependent AUC of 0.739 (95% CI 0.644-0.833) on the GastricSur dataset, demonstrating a 6.3% and 7.5% improvement over the second-best method, Nystromer, respectively. Furthermore, on the TCGA-STAD dataset, our framework achieved a C-index of 0.661 (95% CI 0.504-0.818) and a time-dependent AUC of 0.690 (95% CI 0.500-0.881), outperforming Nystromer by 2.4% in C-index and 4.1% in AUC. These results validate the effectiveness of our proposed framework in survival prediction when utilizing heterogeneous multimodal data, even in the presence of missing modalities. To further validate the effectiveness of our framework in predicting overall survival, we stratified all patients into low-risk and high-risk groups based on the median of the predicted risk scores. Subsequently, we employed Kaplan-Meier curves to

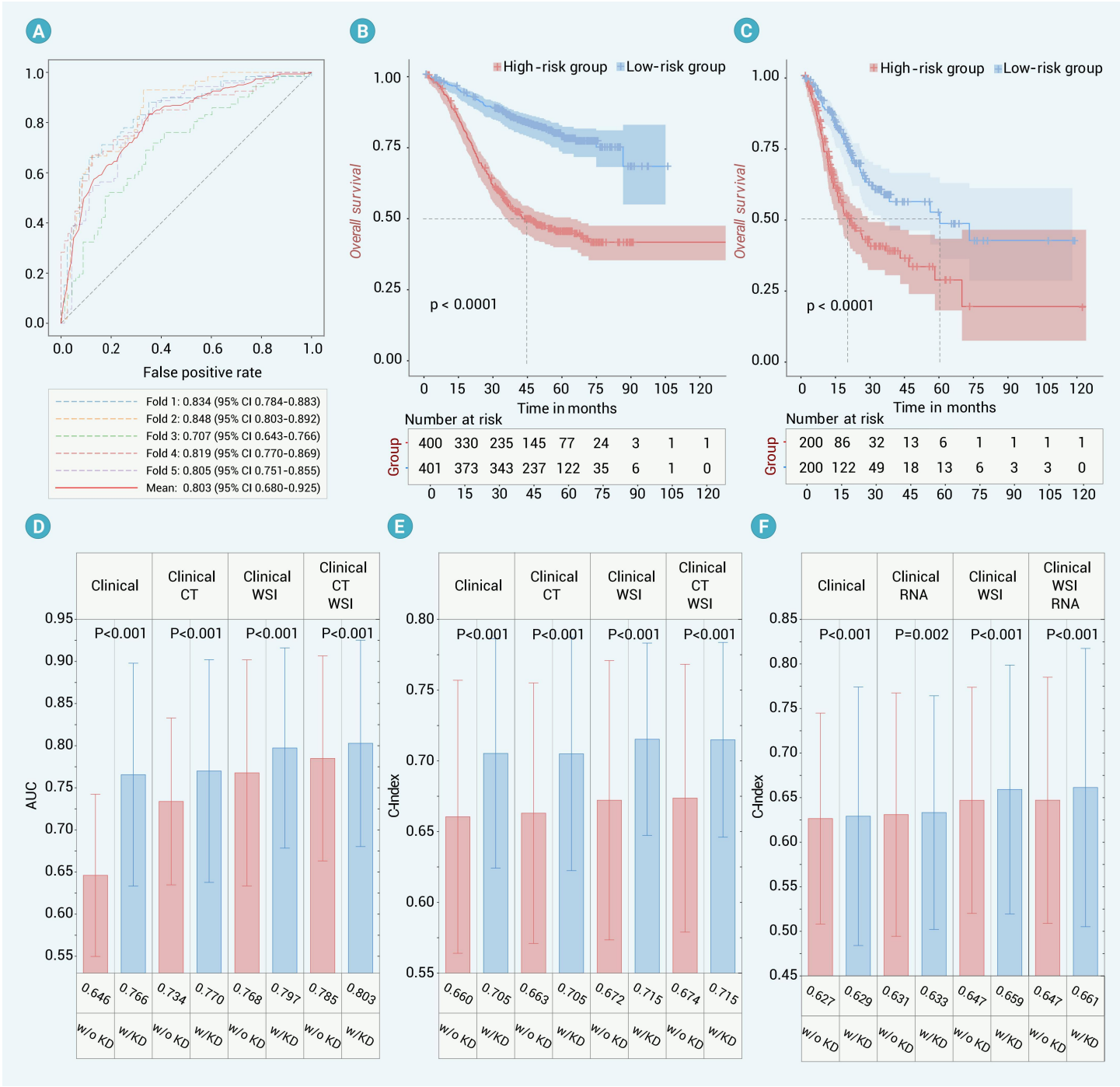


Figure 3. ROC curves for treatment prediction, Kaplan-Meier curves for survival analysis, and comparative analysis for knowledge distillation (A) ROC curves of different folds on the GastricRes dataset are shown in different colors, with the mean ROC curve in red. (B-C) Kaplan-Meier curves for survival analysis on the GastricSur and TCGA-STAD datasets, respectively. All patients are stratified into low-risk group and high-risk group according to the predicted risk scores. (D-F) Performance comparison of knowledge distillation. The top row shows the combinations of modalities used in the experiments, while the bottom rows present the corresponding performance with (w/ KD) and without (w/o KD) knowledge distillation.

visualize the survival events of all patients. The analysis results were presented in Figures 3B-C. Additionally, the Logrank test (p-value) was employed to measure the statistical significance between the low-risk group and the high-risk group. The p-values associated with our framework (4.50×10^{-23} for GastricSur and 3.17×10^{-6} for TCGA-STAD) were significantly lower than 0.05. These findings further reinforced the effectiveness of our proposed framework in survival analysis.

Comparative analysis on multimodal data integration

To further enhance the performance of our framework when dealing with severely incomplete multimodal data, we employed knowledge distillation, aiming to transfer the knowledge acquired from all available modalities to the reduced subset. That is, during training, the teacher model has access to all three modalities, while the student model is trained on a subset of these modalities. The distillation process encourages the student model to learn from the richer representations of the teacher model, thereby improving its performance even when some modalities are missing.

To investigate the potential of multimodal data integration and evaluate the impact of knowledge distillation, we conducted comparative experiments across three collected datasets, as depicted in Figures 3D-F. It is worth noting that clinical information serves as the fundamental information in clinical practice and is available for all patients. Leveraging clinical information as

the baseline modality, we evaluated our framework's performance across various modality subsets. Initially focusing on clinical information alone, our framework achieved an AUC of 0.646 on GastricRes, a C-index of 0.660 on GastricSur, and a C-index of 0.627 on TCGA-STAD. Upon integrating additional modalities, our framework exhibited a gradual enhancement in performance. Especially when involving WSIs, the performance of our framework significantly improved (AUC of0.768 on GastricRes, C-index of 0.672 on GastricSur, and C-index of 0.647 on TCGA-STAD), demonstrating the importance of WSIs in GC analysis. When integrating all modalities, our framework achieved an AUC of 0.785 (+13.9%) on GastricRes, a C-index of 0.674 (+1.4%) on GastricSur, and a C-index of 0.647 (+2.0%) on TCGA-STAD, showcasing remarkable improvements over the baseline. These results underscore the effectiveness of our framework in integrating heterogeneous multimodal data for enhanced prediction performance in GC analysis.

It is worth noting that applying knowledge distillation can lead to significant enhancements in the performance of clinical records, with an AUC of 0.766 (+12.0%) on GastricRes, a C-index of 0.705 (+4.5%) on GastricSur, and a C-index of 0.629 (+0.2%) on TCGA-STAD. That is, knowledge distillation can strengthen the robustness of our framework in handling incomplete multimodal data. Furthermore, the performance gains achieved by knowledge distillation were further enhanced when considering all available modalities, achieving an AUC of 0.803 (+1.8%) on GastricRes, a C-index of 0.715 (+4.1%) on GastricSur, and a C-index of 0.661 (+1.4%) on TCGA-STAD. This observation emphasizes the effectiveness of knowledge distillation in enhancing the performance of multimodal learning frameworks. However, the benefits derived from knowledge distillation are not universally applicable in all scenarios. For instance, the CT scans in GastricSur and clinical information in TCGA-STAD alone may lack the discriminative information required for accurate survival prediction, rendering knowledge distillation ineffective in improving the performance. In such cases, the inclusion of other informative modalities becomes imperative to bolster predictive accuracy.

iMD4GC reveals contribution of individual clinical record

Clinical information encompasses a wide array of information, including demographics, serological markers, and pathological assessments. Prior investigations have highlighted the potential of leveraging these records to anticipate responses to NACT treatment and to forecast the overall survival of GC patients. However, the specific contribution of individual clinical record to the model predictions remains uncertain. To gain insights into the significance of different clinical records, we employ Integrated Gradients to calculate the contribution of each clinical record in the model decision-making process,[42] as illustrated in Figures 4A-B. Positive contributions indicate variables that positively influence the model's predictions, while negative contributions signify factors that exert a negative impact. Conversely, zero contributions suggest that certain variables have a minimal effect on the model's decision-making process. Notably, factors such as lesion location and tumor stage exhibit minimal impact on the decision-making process of the model, suggesting that these variables may not significantly contribute to predicting NACT response or survival prognosis. Similarly, demographic factors like gender, age, and BMI demonstrate limited influence on the model's predictions. In contrast, variables related to serological testing and pathological examinations emerge as key determinants, underscoring their importance in forecasting patient outcomes. These findings underscore the effectiveness of our framework in identifying key clinical records for predicting NACT response and overall survival of GC patients.

iMD4GC enables identification of discriminative pathological patterns

In the field of computational pathology, the analysis of WSIs presents a formidable challenge due to their large size. To surmount this obstacle, multiple instance learning (MIL) has been introduced to process gigapixel WSIs.[43,44] This methodology involves segmenting each WSI into a series of smaller patches, which are then input into the model for predictive analysis. However, not all patches have equal importance in model decision-making. It is necessary to pinpoint the most informative patches that significantly contribute to the model's predictions, as this can unveil crucial insights into the underlying biological mechanisms of GC. Within our iMD4GC framework, we utilize attention scores to identify the discriminative patterns and reveal the relative importance of each pathological patch in the model predictions, as depicted in Figures 4C-D. Patches with high attention scores are considered more valuable for the model predictions, while those with low attention scores carry less significance. Notably, our proposed framework demonstrates the capability to identify morphology specific to different prediction tasks, even in the absence of pixel-level and patch-level annotations. This indicates that our model can learn meaningful features directly from WSIs, enabling accurate prediction without the need for detailed annotations. By leveraging attention scores, iMD4GC provides insights into pathological patterns that contribute significantly to predictions, enhancing our understanding of the underlying factors influencing NACT response and survival prognosis. Furthermore, we utilized Hover-Net to segment the cellular morphology within the high-attention regions, enabling a detailed comparison of cellular structures between different groups.[45] The segmentation results are illustrated in Figure 4E. These findings offer valuable insights into the cellular characteristics that influence the model predictions, shedding light on the biological mechanisms underlying GC progression and patient outcomes.

iMD4GC enables localization of tumor region in radiology images

In prior investigations cantered on predicting NACT response and conducting survival analyses for GC, a predominant reliance on radiomics features derived from CTs was noted, necessitating meticulous manual delineation of the tumor region at the pixel level by radiologists. However, this process is extremely time-consuming and labor-intensive. In our framework, we circumvent the need for pixel-level annotations by directly inputting radiology images into the model for prediction. Nevertheless, our framework still provides a coarse localization of the tumor region. Drawing inspiration from interpretability methodologies employed in pathological analysis, we have incorporated attention scores to unveil the varying importance of each slice within the CTs for model predictions, thereby providing axial localization of the tumor region. Subsequently, we utilized Grad-CAM to highlight the relative importance of pixels in each slice for the model predictions, enabling sagittal and coronal localization of the tumor region.[46] There is a localization example depicted in Figure 5A. Upon comparison with the ground truth annotations delineated by radiologists, the localization results exhibit a favorable level of accuracy, they can provide valuable information to aid radiologists in locating the tumor and delineating its region. The consistency observed between the localization results and the ground truth further validates the effectiveness of our proposed framework in predicting NACT response and survival prognosis.

iMD4GC enables potential biomarker discovery in transcriptomics profiles

We utilized RNA-seq as the transcriptomics profile, which enables the measurement of gene expression levels. This information offers valuable insights into the functional activity of genes and their involvement in various biological processes. However, transcriptomics profiles are characterized by their exceptionally high dimensionality, necessitating an analysis of the relative importance of each gene in the model predictions. To tackle this complexity, we leveraged attention scores to unveil the significance of individual genes in our predictive models. In Figure 5B, we present the distribution of attention scores for the top 100 genes based on their contribution values. Additionally, we provide a list of the top 20 genes that demonstrate potential associations with survival prediction in GC. These findings highlight the significance of specific genes in influencing the model predictions and offer potential targets for further investigation and understanding of GC prognosis. Among the highlighted genes, HOXA9, a transcription factor renowned for its involvement in embryonic development and cellular differentiation, emerges as a notable candidate linked to GC progression.[47-49] Its overexpression hints at a plausible association with adverse survival outcomes among GC patients. RBMX, an RNA-binding protein, showcases altered expression patterns in GC,[50-52] potentially fueling tumorigenesis and influencing patient survival trajectories. CALM2, a calcium-binding protein pivotal in intracellular signaling, has been implicated in various cancers, including GC,[53,54] suggesting that dysregulation of CALM2 could impact survival outcomes in this patient cohort. Furthermore, SUMO2, a protein modifier crucial in post-translational modifications, has been correlated with GC onset and progression.[55]

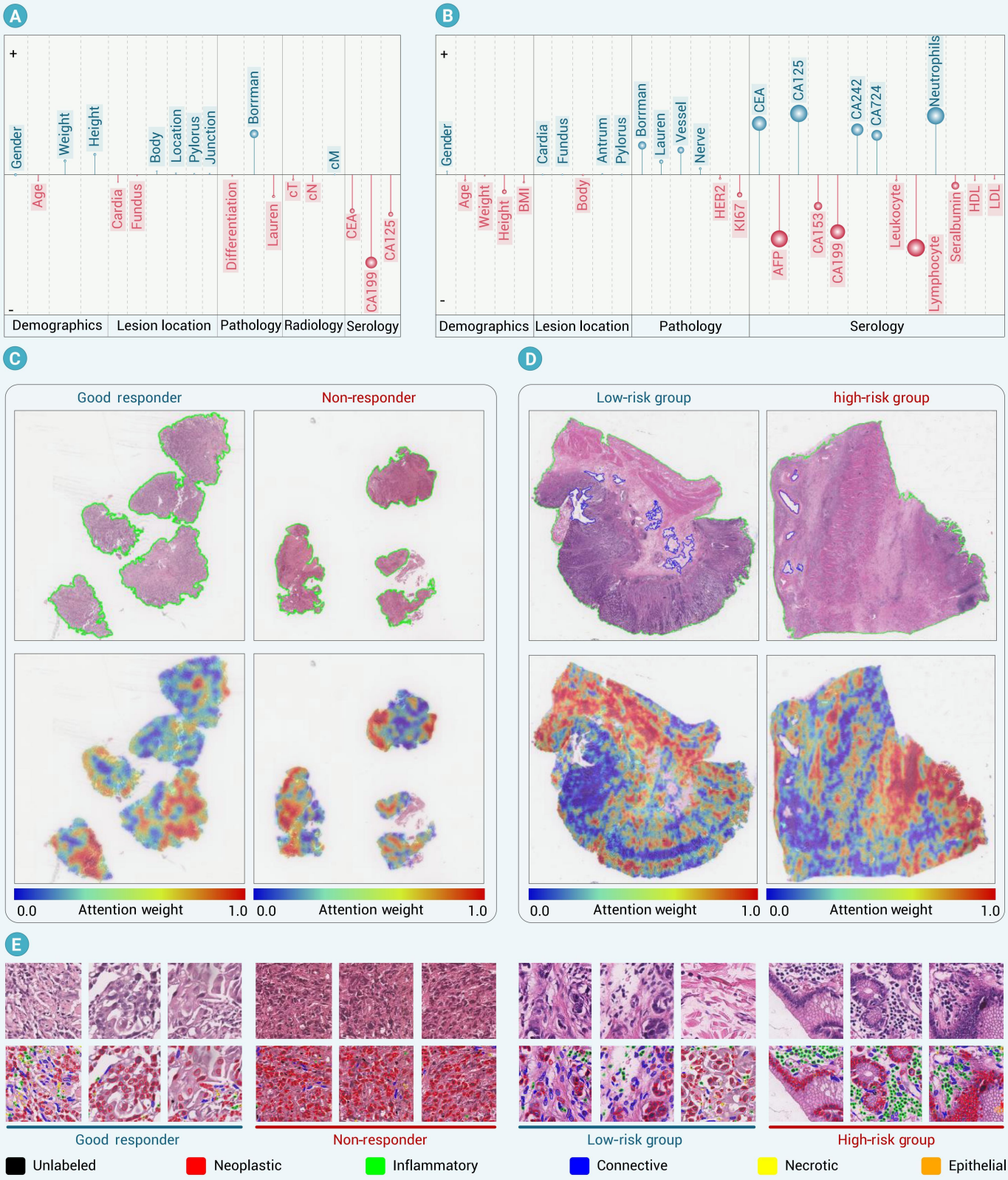
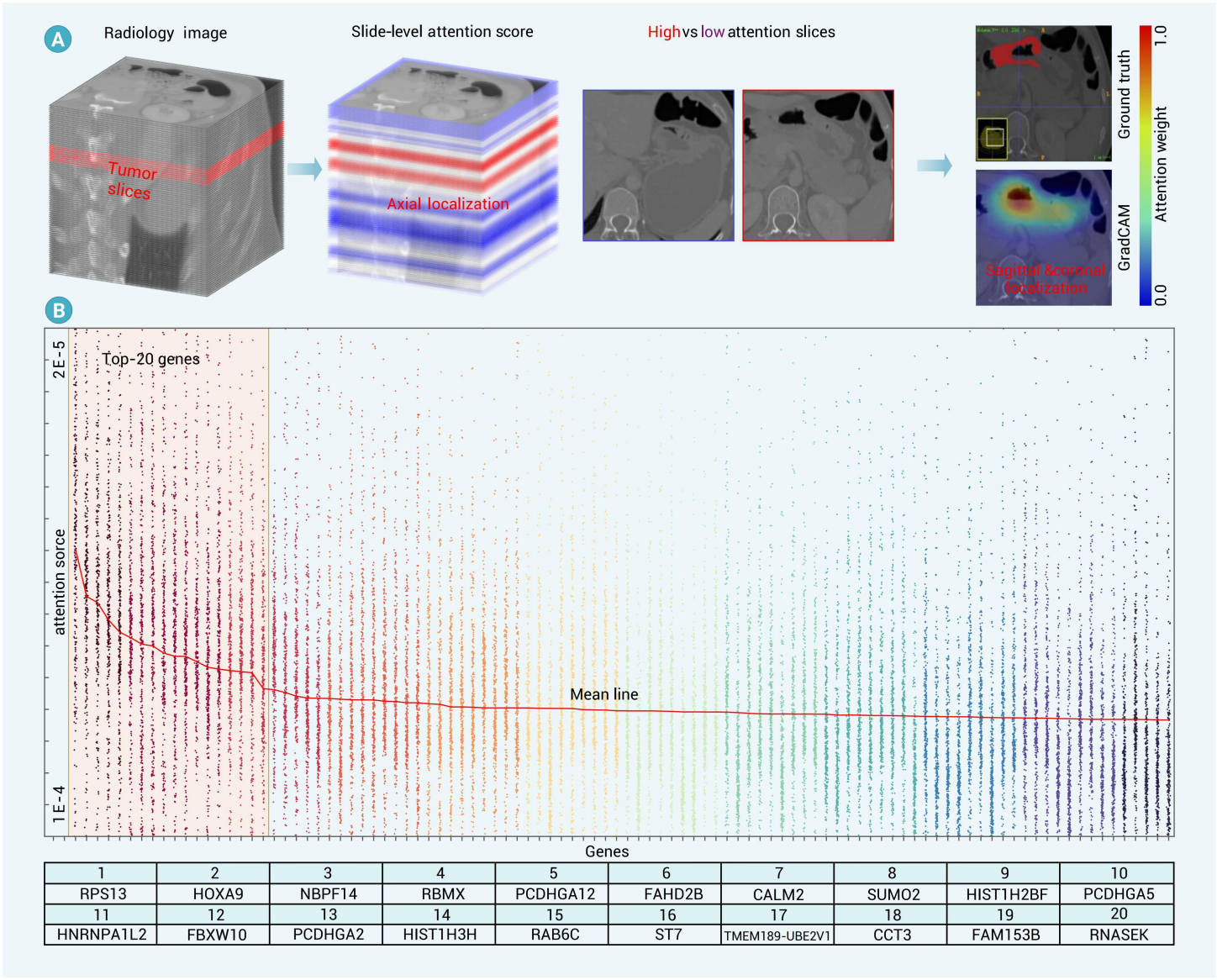


Figure 4. Interpretability analysis of clinical records and pathological images (A-B) Contribution of individual clinical record for NACT response prediction and survival analysis. The contribution values are calculated by Integrated Gradients. (C-D) Visualization comparison between good responders and non-responders. The regions with high attention scores are deemed more valuable for model prediction, while those with low attention scores carry less significance. (E) Cellular morphology comparison between non-responders and good responders, and low-risk and high-risk groups. Cell types within high-attention regions are annotated.

Further exploration is imperative to unravel the functions, interactions, and prognostic implications of the remaining genes identified in our analysis concerning survival prediction in GC. These insights not only deepen our understanding of the molecular landscape of GC but also hold promise for identifying novel therapeutic targets and refining prognostic strategies in the battle against this complex disease.



1	2	3	4	5	6	7	8	9	10
RPS13	HOXA9	NBPF14	RBMX	PCDHGA12	FAHD2B	CALM2	SUMO2	HIST1H2BF	PCDHGA5
11	12	13	14	15	16	17	18	19	20
HNRNPA1L2	FBXW10	PCDHGA2	HIST1H3H	RAB6C	ST7	TMEM189-UBE2V1	CCT3	FAM153B	RNASEK

Figure 5. Interpretability analysis of radiology images and transcriptomics profiles for GC analysis (A) An example of localization results for radiology images. The attention scores in iMD4GC provides the axial localization of the tumor region first. And then, utilizing Grad-CAM shows the sagittal and coronal localization of the tumor region. (B) The attention distribution of the top-100 genes for survival prediction. Each point represents the attention score of one sample. The table shows the top-20 genes with the highest contribution values.

iMD4GC achieves robustness and generalizability across centers

To assess the robustness and generalizability of our proposed iMD4GC framework across different medical centers, we conducted cross-center validation experiments on both the GastricRes and GastricSur datasets. Specifically, for treatment response prediction, there are four medical centers contributing to the GastricRes dataset. Among them, three centers (Shanxi Cancer Hospital, Sichuan Cancer Hospital, and the Sixth Affiliated Hospital of Sun Yat-sen University)) were merged to form the internal cohort and randomly split into training (70%), validation (10%), and test sets (20%), while the remaining center (Yunnan Cancer Hospital) served as the external validation set. Similarly, for survival prediction, the GastricSur dataset comprises two medical centers. One center (Yunnan Cancer Hospital) was employed for model training, while the other center was designated as the external validation set. The experimental results are shown in Figures 6A-D. For NACT response prediction, our framework achieved an AUC of 0.818 (95% CI 0.724-0.897) on the internal test set and an AUC of 0.801 (95% CI 0.747-0.912) on the external validation set. For survival prediction, our framework attained a C-index of 0.724 (95% CI 0.638-0.804) on the internal test set and a C-index of 0.706 (95% CI 0.656-0.753) on the external validation set. These results demonstrate that our proposed framework maintains robust performance across different medical centers.

When comparing with other baseline methods, we employed Wilcoxon signed-rank test to assess the statistical significance of performance differences. The results indicate that the performance differences between our proposed iMD4GC model and the baseline models are statistically significant ($p < 0.05$). By default, we utilize CTransPath to extract features from pathological images. With recent advancements in pathology foundation models, its powerful feature extraction capabilities can further enhance the performance of our framework. To validate this, we replaced CTransPath with the state-of-the-art pathology foundation model, UNI,[56] and conducted comparative experiments on both the GastricRes and GastricSur datasets. For NACT response prediction, our framework achieved an AUC of 0.835 (95% CI 0.755-0.915) on the internal test set and an AUC of 0.821 (95% CI 0.756-0.876) on the external validation set. For survival prediction, our framework attained a C-index of 0.738 (95% CI 0.655-0.825) on the internal test set and a C-index of 0.717 (95% CI 0.669-0.763) on the external validation set. These experimental results demonstrated that the integration of advanced foundation models like UNI led to improved performance in both NACT response prediction and survival analysis tasks, further underscoring the adaptability and effectiveness of our iMD4GC framework in leveraging cutting-edge technologies for enhanced clinical outcomes.

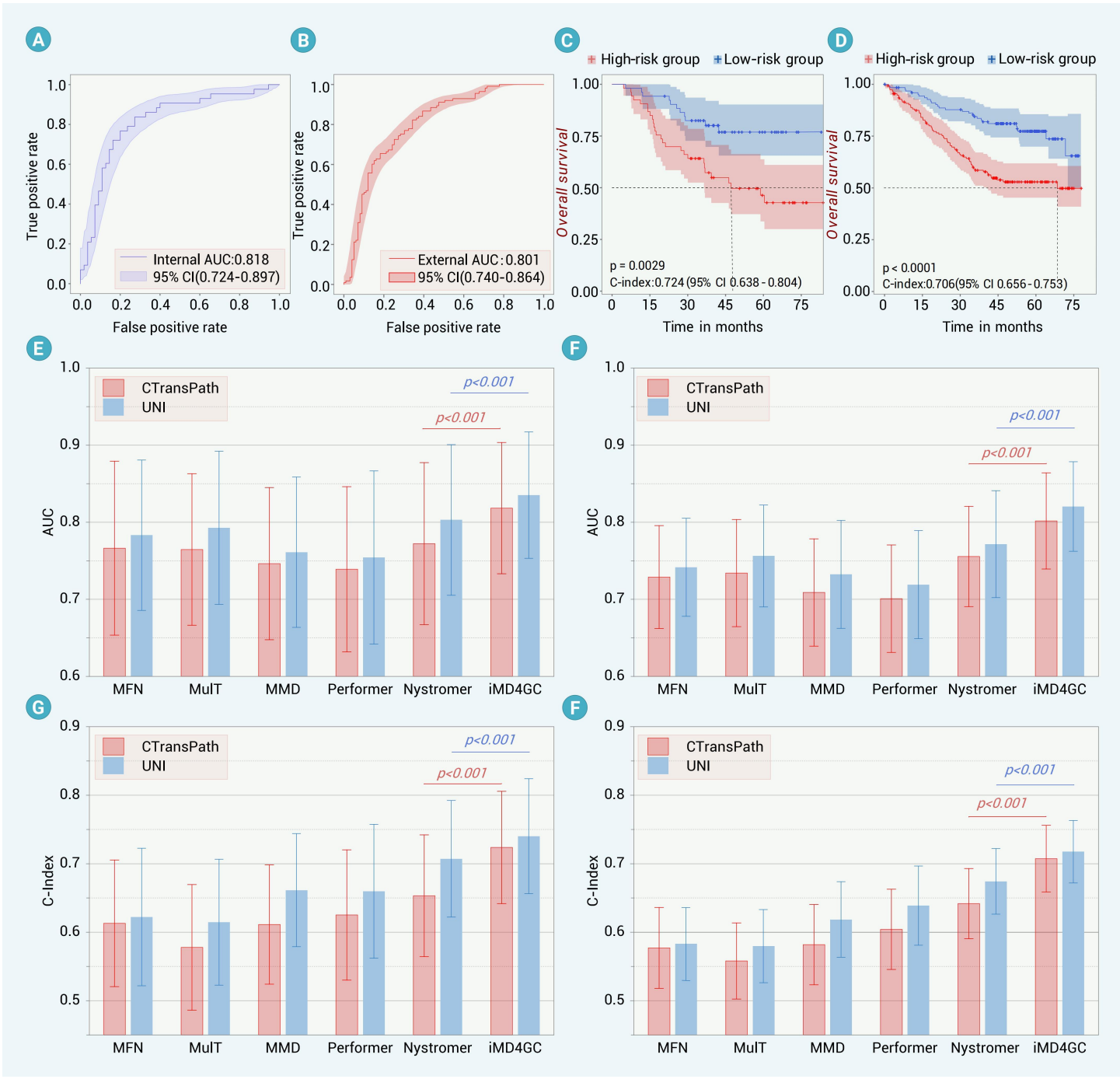


Figure 6. Cross-center validation results (A-B) AUC-ROC curves for NACT response prediction on internal and external validation sets, respectively. (C-D) Kaplan-Meier curves for overall survival predictions on internal and external validation sets, respectively. (E-F) Performance comparison for NACT response prediction among different models on internal and external validation sets, respectively. (G-H) Performance comparison for overall survival prediction among different models on internal and external validation sets, respectively. The Wilcoxon signed-rank test was used to assess statistical significance between best and second-best models.

DISCUSSION

In medical AI, the application of deep learning models has demonstrated exceptional performance in various clinical tasks. However, this lack of interpretability creates barriers to the widespread adoption of such models in critical healthcare settings. The inherent interpretability of iMD4GC represents a significant breakthrough in addressing these concerns, providing transparent analysis into the model's decision-making process. Through a comprehensive interpretability analysis, we revealed the discriminative patterns of each modality contributing to the model predictions. This analysis was conducted in collaboration with an interdisciplinary team of experts, including clinicians, radiologists, pathologists, and biologists, ensuring the validity and clinical relevance of our findings. The findings gained from this interpretability analysis hold the potential to discover new biomarkers or therapeutic targets, thereby benefiting GC treatment and management.

This study focused on four primary modalities: clinical information, pathology images, radiology images, and transcriptomics profiles. Nevertheless, it is imperative to acknowledge the existence of additional modalities generated during the medical journey of patients, including endoscopy images and clinical reports. Integrating these additional modalities has the potential to yield valuable insights and enhance the predictive capabilities of our models. The flexible and scalable nature of iMD4GC enables the seamless extension to include these additional modalities through tailored preprocessing and tokenization strategies. In our future research, we intend to incorporate these modalities into iMD4GC and make more precise predictions for GC study. Besides, the datasets collected in this study have potential imbalance in terms of demographic factors, which may introduce prediction biases in the

model's learning process. Similarly, the scale of datasets is relatively limited, which may constrain the model's predictive performance. To address these limitations, it is crucial to undertake efforts to mitigate demographic biases and enhance dataset diversity. Alongside increasing the dataset's scale, the incorporation of foundation models is a promising avenue for future exploration. Foundation models possess powerful feature extraction and multimodal information fusion abilities, which can significantly enhance the model's performance and generalizability.

Recently, potential heterogeneity, possible incompleteness, and lack of interpretability have become the most significant challenges in AI research when applied multimodal fusion framework to clinical practice. Beyond its application in GC analysis, the proposed iMD4GC framework has vast potential for broader clinical applications, spanning diverse cancer types and non-cancerous diseases. The flexible scalability of iMD4GC streamlines the integration of varied data sources, facilitating precise oncology through AI and multimodal data fusion. Leveraging the expanding wealth of available data, researchers and clinicians can gain a comprehensive understanding of diseases, enhancing patient care and outcome.

CONCLUSION

As one of the most prevalent malignancies globally, GC continues to impose significant burdens on public healthcare systems. To enhance the efficiency and accuracy of GC analysis, AI has emerged as a promising tool, demonstrating remarkable advancements across various clinical applications. However, the intrinsic heterogeneity of multimodal data presents formidable challenges to seamless integration, leading existing studies to predominantly focus on either unimodal data applications or a restricted number of modalities. Moreover, the incompleteness of multimodal data is a common issue due to various constraints, such as data corruption, privacy concerns, and technical limitations, further hindering the direct application of existing multimodal learning methods in clinical practice. To address these challenges, we proposed a novel multimodal learning framework for GC study, iMD4GC, which is robust to heterogeneous multimodal data and effective in handling incomplete multimodal data. Extensive experiments on three collected datasets demonstrated that iMD4GC achieved promising performance in predictions of NACT treatment response and survival analysis, outperforming other compared methods by a substantial margin. The interpretability analysis demonstrated the clinical relevance of the learned representations, providing valuable insights into the discriminative patterns of each modality contributing to the model predictions. Overall, iMD4GC represents a significant advancement in multimodal learning for GC analysis, with potential applications across various clinical domains.

REFERENCES

1. Bray F., Laversanne M., Sung H., et al. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA-Cancer J. Clin.* **74**:229–263. DOI:10.3322/caac.21834
2. Zurlo I.V., Schino M., Strippoli A., et al. (2022). Predictive value of NLR, TILs (CD4+/CD8+) and PD-L1 expression for prognosis and response to preoperative chemotherapy in gastric cancer. *Cancer Immunol. Immunother.* **71**:45–55. DOI:10.1007/s00262-021-02960-1
3. Liu Z., Wang Y., Shan F., et al. (2021). Treatment switch in poor responders with locally advanced gastric cancer after neoadjuvant chemotherapy. *Ann. Surg. Oncol.* **28**:8892–8907. DOI:10.1245/s10434-021-10087-x
4. Yoshikawa T., Sasako M., Yamamoto S., et al. (2009). Phase II study of neoadjuvant chemotherapy and extended surgery for locally advanced gastric cancer. *Br. J. Surg.* **96**:1015–1022. DOI:10.1002/bjs.6665
5. Lowy A.M., Mansfield P.F., Leach S.D., et al. (1999). Response to neoadjuvant chemotherapy best predicts survival after curative resection of gastric cancer. *Ann. Surg.* **229**:303–308. DOI:10.1097/00000658-199903000-00001
6. Fu J., Tang L., Li Z.Y., et al. (2020). Diffusion kurtosis imaging in the prediction of poor responses of locally advanced gastric cancer to neoadjuvant chemotherapy. *Eur. J. Radiol.* **128**:108974. DOI:10.1016/j.ejrad.2020.108974
7. Zhang J., Cui Y., Wei K., et al. (2022). Deep learning predicts resistance to neoadjuvant chemotherapy for locally advanced gastric cancer: A multicenter study. *Gastric Cancer* **25**:1050–1059. DOI:10.1007/s10120-022-01328-3
8. Wang S., Zhu Y., Yu L., et al. (2019). RMDL: Recalibrated multi-instance deep learning for whole slide gastric image classification. *Med. Image Anal.* **58**:101549. DOI:10.1016/j.media.2019.101549
9. Yu H., Zhang X., Song L., et al. (2021). Large-scale gastric cancer screening and localization using multi-task deep neural network. *Neurocomputing* **448**:290–300. DOI:10.1016/j.neucom.2021.03.006
10. Meng L., Dong D., Chen X., et al. (2020). 2D and 3D CT radiomic features performance comparison in characterization of gastric cancer: A multi-center study. *IEEE J. Biomed. Health Inform.* **25**:755–763. DOI:10.1109/jbhi.2020.3002805
11. Yuan M., Xia Y., Chen X., et al. (2023). Cluster-induced mask transformers for effective opportunistic gastric cancer screening on non-contrast CT scans. *Med. Image Comput. Comput.-Assist. Interv.* **14223**:146-156. DOI:10.1007/978-3-031-439049_15
12. Du G., Zeng Y., Chen D., et al. (2023). Application of radiomics in precision prediction of diagnosis and treatment of gastric cancer. *Jpn. J. Radiol.* **41**:245–257. DOI:10.1007/s11604-022-01352-4
13. Jiang Y., Zhou K., Sun Z., et al. (2023). Non-invasive tumor microenvironment evaluation and treatment response prediction in gastric cancer using deep learning radiomics. *Cell Rep. Med.* **4**:101146. DOI:10.1016/j.xcrm.2023.101146
14. Chen T., Li X., Mao Q., et al. (2022). An artificial intelligence method to assess the tumor microenvironment with treatment outcomes for gastric cancer patients after gastrectomy. *J. Transl. Med.* **20**:100. DOI:10.1186/s12967-022-03298-7
15. Li X., Zhang Y., Zhang Y., et al. (2010). Survival prediction of gastric cancer by a seven-microRNA signature. *Gut* **59**:579–585. DOI:10.1136/gut.2008.175497
16. Oh S.E., Seo S.W., Choi M.G., et al. (2018). Prediction of overall survival and novel classification of patients with gastric cancer using the survival recurrent network. *Ann. Surg. Oncol.* **25**:1153–1159. DOI:10.1016/j.ejso.2018.10.287
17. Chen D., Fu M., Chi L., et al. (2022) Prognostic and predictive value of a pathomics signature in gastric cancer. *Nat. Commun.* **13**:6903. DOI:10.1038/s41467-022-34703-w
18. Xu Q., Sun Z., Li X., et al. (2021). Advanced gastric cancer: CT radiomics prediction and early detection of downstaging with neoadjuvant chemotherapy. *Eur. Radiol.* **31**:8765–8774. DOI:10.1007/s00330-021-07962-2
19. Sun J., Wang X., Zhang Z., et al. (2021). The sensitivity prediction of neoadjuvant chemotherapy for gastric cancer. *Front. Oncol.* **11**:641304. DOI:10.3389/fonc.2021.641304
20. Cui Y., Zhang J., Li Z., et al. (2022). A CT-based deep learning radiomics nomogram for predicting the response to neoadjuvant chemotherapy in patients with locally advanced gastric cancer: A multicenter cohort study. *EClinicalMedicine* **46**:101348. DOI:10.1016/j.eclinm.2022.101348
21. Gu J., Tong T., He C., et al. (2022). Deep learning radiomics of ultrasonography can predict response to neoadjuvant chemotherapy in breast cancer at an early stage of treatment: a prospective study. *Eur. Radiol.* **32**:2099–2109. DOI:10.1007/s00330-021-08293-y
22. Wang X., Zhang Y., Yang M., et al. (2024). Dynamic ultrasound-based modeling predictive of response to neoadjuvant chemotherapy in patients with early breast cancer. *Sci. Rep.* **14**:31644. DOI:10.1038/s41598-024-80409-y
23. Hottat N.A., Badr D.A., Lecomte S., et al. (2023). Assessment of diffusion-weighted MRI in predicting response to neoadjuvant chemotherapy in breast cancer patients. *Sci. Rep.* **13**:614. DOI:10.1038/s41598-023-27787-x
24. Ogier du Terrail J., Leopold A., Joly C., et al. (2023). Federated learning for predicting histological response to neoadjuvant chemotherapy in triple-negative breast cancer. *Nat. Med.* **29**:135–146. DOI:10.1038/s41591-022-02155-w
25. Zhang L., Dong D., Zhang W., et al. (2020). A deep learning risk prediction model for overall survival in patients with gastric cancer: A multicenter study. *Radiother. Oncol.* **150**:73–80. DOI:10.2139/ssrn.3514655
26. Zhong H., Wang T., Hou M., et al. (2024). Deep learning radiomics nomogram based on enhanced CT to predict the response of metastatic lymph nodes to neoadjuvant chemotherapy in locally advanced gastric cancer. *Ann. Surg. Oncol.* **31**:421–432. DOI:10.1245/s10434-023-14508-x
27. Li J., Yin H., Wang Y., et al. (2023). Multiparametric MRI-based radiomics nomogram for early prediction of pathological response to neoadjuvant chemotherapy in locally advanced gastric cancer. *Eur. Radiol.* **33**:2746–2756. DOI:10.1007/s00330-022-09219-y
28. Zhang Y., Chen Q., Yang Z., et al. (2019). BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Sci. Data* **6**:52. DOI:10.1038/s41597-019-0055-0
29. Wasserthal J., Breit H.C., Meyer M.T., et al. (2023). TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images. *Radiology: Artif. Intell.* **5**:e230024. DOI:10.1148/ryai.230024
30. He K., Zhang X., Ren S., et al. (2016). Deep residual learning for image recognition. *CVPR* **2016**:770-778. DOI:10.1109/cvpr.2016.90
31. Vaswani A., Shazeer N., Parmar N., et al. (2017). Attention is all you need. *Adv. Neural Inf. Process. Syst. arXiv* preprint. DOI:10.48550/arXiv.1706.03762.
32. Wang X., Yang S., Zhang J., et al. (2022). Transformer-based unsupervised contrastive learning for histopathological image classification. *Med. Image Anal.* **81**:102559. DOI:10.1016/j.media.2022.102559
33. Du J., Jia P., Dai Y., et al. (2019). Gene2vec: Distributed representation of genes based on co-expression. *BMC Genomics* **20**:7–15. DOI:10.1186/s12864-018-5370-x
34. Tarvainen A. andValpola H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Proc. 31st Int. Conf. Neural Inf. Process. Syst.* pp:1195-1204
35. Zadeh A., Liang P.P., Mazumder N., et al. (2018). Memory fusion network for multi-

view sequential learning. *Proc. AAAI Conf. Artif. Intell* **2018**:5634-5641. DOI:10.1609/aaai.v32i1.12021
36. Tsai Y.H.H., Bai S., Liang P.P., et al. (2019). Multimodal transformer for unaligned multimodal language sequences. *Proc. 57th Annu. Meet. Assoc. Comput. Linguist* **2019**:6558-6569. DOI:10.18653/v1/p19-1656
37. Cui C., Liu H., Liu Q., et al. (2022) Survival prediction of brain cancer with incomplete radiology, pathology, genomic, and demographic data. *Med. Image Comput. Comput. Assist. Interv* **2022**:626-635. DOI:10.1007/978-3-031-16443-9_60
38. Choromanski K.M., Likhosherstov V., Dohan D., et al. (2020). Rethinking attention with performers. *International Conference on Learning Representations. arXiv.* 2009.14794 DOI:10.48550/arXiv.2009.14794
39. Xiong Y., Zeng Z., Chakraborty R., et al. (2021). Nyströmformer: A nyström-based algorithm for approximating self-attention. *Proc. AAAI Conf. Artif. Intell* **2021**:14138-14148. DOI:10.1609/aaai.v35i16.17664
40. Hung H. and Chiang C.T. (2010). Estimation methods for time - dependent AUC models with survival data. *Can. J. Stat.* **38**:8–26. DOI:10.1002/cjs.10046
41. Lambert J. and Chevret S. (2016). Summary measure of discrimination in survival models based on cumulative/dynamic time-dependent ROC curves. *Stat. Methods Med. Res.* **25**:2088–2102. DOI:10.1177/0962280213515571
42. Sundararajan M., Taly A., and Yan Q. (2017). Axiomatic attribution for deep networks. *Int. Conf. Mach. Learn* **70**:3319-3328. DOI:10.48550/arXiv.1703.01365
43. Gadermayr M. and Tschuchnig M. (2024). Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations & future potential. *Comput. Med. Imaging Graph.* **112**:102337–102337. DOI:10.1016/j.compmedimag.2024.102337
44. Tang W., Huang S., Zhang X., et al. (2023). Multiple instance learning framework with masked hard instance mining for whole slide image classification. *Proc. IEEE/CVF Int. Conf. Comput. Vis* **2023**:4078-4087. DOI:10.1109/ICCV51070.2023.00377
45. Graham S., Vu Q.D., Raza S.E.A., et al. (2019). Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med. Image Anal.* **58**:101563. DOI:10.1016/j.media.2019.101563
46. Selvaraju R.R., Cogswell M., Das A., et al. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proc. IEEE Int. Conf. Comput. Vis.* **2017**:618-626. DOI:10.1109/iccv.2017.74
47. Yang Y.C., Wang S.W., Wu I.C., et al. (2009). A tumorigenic homeobox (HOX) gene expressing human gastric cell line derived from putative gastric stem cell. *Eur. J. Gastroenterol. Hepatol.* **21**:1016–1023. DOI:10.1097/MEG.0b013e32831d7b9c
48. Ma Y.Y., Zhang Y., Mou X.Z., et al. (2017). High level of homeobox A9 and PBX homeobox 3 expression in gastric cancer correlates with poor prognosis. *Oncol. Lett.* **14**:5883–5889. DOI:10.3892/ol.2017.6937
49. Shenoy U.S., Adiga D., Alhedyan F., et al. (2023). HOXA9 transcription factor is a double-edged sword: From development to cancer progression. *Cancer Metastasis Rev.* **43**:709–728. DOI:10.1007/s10555-023-10159-2
50. Ge S., Xia X., Ding C., et al. (2018). A proteomic landscape of diffuse-type gastric cancer. *Nat. Commun.* **9**:1012. DOI:10.1007/s10120-021-01226-0
51. Yan Q., Zeng P., Zhou X., et al. (2021). RBMX suppresses tumorigenicity and progression of bladder cancer by interacting with the hnRNP A1 protein to regulate PKM alternative splicing. *Oncogene* **40**:2635–2650. DOI:10.1038/s41388-021-01666-z
52. Wang X., Xu K., Liao X., et al. (2022). Construction of a survival nomogram for gastric cancer based on the cancer genome atlas of m6A-related genes. *Front. Genet.* **13**:936658. DOI:10.3389/fgene.2022.936658
53. Mu G., Zhu Y., Dong Z., et al. (2021). Calmodulin 2 facilitates angiogenesis and metastasis of gastric cancer via STAT3/HIF-1A/VEGF-A mediated macrophage polarization. *Front. Oncol.* **11**:727306. DOI:10.3389/fonc.2021.727306
54. Sun J., Cao Q., Lin S., et al. (2023). Knockdown of CALM2 increases the sensitivity to afatinib in HER2-amplified gastric cancer cells by regulating the Akt/FoxO3a/Puma axis. *Toxicol. Vitro* **87**:105531. DOI:10.1016/j.tiv.2022.105531
55. Hu Y., Chen C., Tong X., et al. (2021). NSUN2 modified by SUMO-2/3 promotes gastric cancer progression and regulates mRNA m5C methylation. *Cell Death Dis.* **12**:842. DOI:10.1038/s41419-021-04127-3
56. Chen R.J., Ding T., Lu M.Y., et al. (2024). Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**:850–862. DOI:10.1038/s41591-024-02857-3

FUNDING AND ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (No. 62202403, 82001986, and 82360345), Hong Kong Innovation and Technology Commission (Project No. PRP/034/22FX and ITCPD/17-9), Shenzhen Science and Technology Innovation Committee Funding (Project No. SGDX20210823103201011), the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. R6003-22 and C4024-22GF). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

F.Z. and Y.X. were responsible for the study design, model implementation, data analysis, and manuscript writing. Y.C., S.Z., Y.Z., and W.H. contributed to data collection and manuscript revision. J.W., X.W., R.C., L.L., and C.H. provided professional guidance, clinical insights, and data interpretation. Z.L. and H.C. are the corresponding authors of this study, and they initiated the project, organized the collaboration between the authors, and supervised the study. All authors contributed to the manuscript and approved the final version.

DECLARATION OF INTERESTS

Hao Chen is an Academic Editor of The Innovation Informatics and was blinded from reviewing or making final decisions on this manuscript. Peer review was handled independently of this member and their research group. The other authors declare no conflicts of interest.

ETHICAL STATEMENT AND PATIENT CONSENT

This study was approved by the Human and Artefacts Research Ethics Committee (HREP-2022-0268).

DATA AND CODE AVAILABILITY

There are three datasets involved in this study, including the GastricRes dataset for NACT response prediction, the GastricSur dataset for survival analysis, and the TCGA-STAD dataset for survival analysis. The first two datasets are not publicly released due to restrictions by privacy concern, but they are available from the corresponding author on reasonable request. The TCGA-STAD dataset is publicly available at https://portal.gdc.cancer.gov. All code was implemented in Python using PyTorch as the primary deep learning package. All code and scripts to reproduce the experiments of this study are available at https://github.com/FT-ZHOU-ZZZ/iMD4GC.