Briefings in Bioinformatics, 2025, **26(3)**, bbaf209
https://doi.org/10.1093/bib/bbaf209
Problem Solving Protocol

MMsurv: a multimodal multi-instance multi-cancer survival prediction model integrating pathological images, clinical information, and sequencing data

Hailong Yang[1,2,‡], Jia Wang[3,‡], Wenyan Wang[1], Shufang Shi[2,4], Lijing Liu[5], Yuhua Yao[6], Geng Tian[2], Peizhen Wang[1,*], Jialiang Yang[2,*]

[1]School of Electrical and Information Engineering, Anhui University of Technology, No. 1530 Maxiang Road, Huashan District, Ma'anshan, Anhui 243032, China
[2]Department of Sciences, Geneis Beijing Co., Ltd., No. 31 Xinbei Road, Laiguangying, Chaoyang District, Beijing 100102, China
[3]Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Department of Thoracic Surgery II, Peking University Cancer Hospital & Institute, No. 52 Fucheng Road, Haidian District, Beijing 100142, P.R. China
[4]Department of Pathology, Beijing Friendship Hospital, Capital Medical University, No. 95 Yong'an Road, Xicheng District, Beijing 100050, China
[5]Qinghe Clinic, Beijing North Medical District, Chinese PLA General Hospital, No. 100 Youyi Road, Haidian District, Beijing 100080, China
[6]School of Mathematics and Statistics, Hainan Normal University, No. 99 Longkunnan Road, Longhua District, Haikou, Hainan 571158, China

[*]Corresponding authors. Peizhen Wang, School of Electrical and Information Engineering, Anhui University of Technology, No. 1530 Maxiang Road, Huashan District, Ma'anshan, Anhui 243032, China. E-mail: pzhwang@ahut.edu.cn; Jialiang Yang, Department of Sciences, Geneis Beijing Co., Ltd., No. 31 Xinbei Road, Laiguangying, Chaoyang District, Beijing 100102, China. E-mail: yangjl@geneis.cn

[‡]Hailong Yang and Jia Wang contributed equally to this study.

Abstract

Accurate prediction of patient survival rates in cancer treatment is essential for effective therapeutic planning. Unfortunately, current models often underutilize the extensive multimodal data available, affecting confidence in predictions. This study presents MMSurv, an interpretable multimodal deep learning model to predict survival in different types of cancer. MMSurv integrates clinical information, sequencing data, and hematoxylin and eosin-stained whole-slide images (WSIs) to forecast patient survival. Specifically, we segment tumor regions from WSIs into image tiles and employ neural networks to encode each tile into one-dimensional feature vectors. We then optimize clinical features by applying word embedding techniques, inspired by natural language processing, to the clinical data. To better utilize the complementarity of multimodal data, this study proposes a novel fusion method, multimodal fusion method based on compact bilinear pooling and transformer, which integrates bilinear pooling with Transformer architecture. The fused features are then processed through a dual-layer multi-instance learning model to remove prognosis-irrelevant image patches and predict each patient's survival risk. Furthermore, we employ cell segmentation to investigate the cellular composition within the tiles that received high attention from the model, thereby enhancing its interpretive capacity. We evaluate our approach on six cancer types from The Cancer Genome Atlas. The results demonstrate that utilizing multimodal data leads to higher predictive accuracy compared to using single-modal image data, with an average C-index increase from 0.6750 to 0.7283. Additionally, we compare our proposed baseline model with state-of-the-art methods using the C-index and five-fold cross-validation approach, revealing a significant average improvement of nearly 10% in our model's performance.

Keywords: multi-instance learning; survival prediction; word embedding; multi-cancer; C-index

Introduction

As the incidence and mortality rates steadily rise, cancer is becoming a leading cause of death globally and a significant public health concern [1]. According to the Global Cancer Report in 2012, there were 12.7 million new cancer cases and 7.6 million deaths, but they increased to 20 million new cases and 9.7 million deaths in 2022 [2, 3].

Therefore, developing a method for accurately predicting cancer patient survival rates is crucial to advance personalized treatment strategies, potentially reducing cancer mortality and improving patient quality of life [4].

Clinical information is an indispensable component of cancer research [5–7], encompassing factors such as patient age, gender, tumor type, staging, and pathological characteristics. In current clinical practice, the assessment of many cancer types still relies on clinical information for grading and staging, which assists categorizing patients into different risk groups and guiding treatment decisions [8]. Nevertheless, with the advancement of cancer research, we are increasingly realizing the high heterogeneity of cancer. It has been observed that even patients at the same stage or level of cancer may exhibit significant differences in their actual survival outcomes [9, 10].

In recent years, gene expression has shown significant potential in identifying prognostic factors and performing effectively in survival prediction [11–14]. For instance, Liang *et al.* performed an analysis on the mRNA expression profiles of hepatocellular carcinoma patients within The Cancer Genome Atlas (TCGA) database. Employing a Cox regression model augmented by the Least Abso-

Received: June 26, 2024. **Revised:** September 3, 2024. **Accepted:** April 17, 2025

lute Shrinkage and Selection Operator, they adeptly identified 26 differentially expressed genes that are significantly correlated with the overall survival rates of hepatocellular carcinoma patients [15]. Similarly, Wu *et al.* utilized bioinformatics analysis to identify 11 prognosis-related genes, which were incorporated into a prognostic model with higher diagnostic accuracy than other clinical pathological features [16]. However, the predictive accuracy of these clinical applications still requires improvement. After all, the tumor environment is a complex, rapidly evolving environment i.e. difficult to characterize by molecular analysis alone [17–19].

In addition to the prognostic methods utilizing gene expression and clinical data mentioned earlier, tumor pathology image analysis is also indispensable and has been demonstrated to provide crucial information for cancer diagnosis and prognosis [13, 20], and several works have been developed to employ popular deep learning techniques for exploiting the potentials of pathological images [21, 22]. For example, Skrede *et al.* employed deep learning in conjunction with digital scans of traditional tumor tissue slides to develop effective prognostic biomarkers for stratifying colon cancer (COAD) patients [23]. Abbet *et al.* proposed a self-supervised approach that utilizes joint learning of tissue region representations and clustering metrics to represent the interactions between complex tissues and directly predict the clinical outcomes of COAD [24].

Predictive models based solely on single types of molecular biomarkers are limited. Recently, the trend has shifted toward integrating multimodal data to enhance disease diagnosis and prediction, crucial for understanding cancer's heterogeneity and complexity. For example, Wang *et al.* successfully combined genomic data and pathological images to predict breast cancer prognosis, showing that using complementary information across modalities significantly improved the accuracy compared to using a single data type [25]. Li *et al.* developed a novel hierarchical multimodal fusion method called HFBSurv, which progressively integrates genomic and image features (IFs) using the factorized bilinear model [26]. Therefore, the use of multimodal fusion and morphological features extracted from whole-slide images (WSIs) holds potential clinical application value [14]. However, there are still some limitations in the aforementioned aspects. For example, most studies utilize only two data modalities, the encoding of clinical features is often arbitrary, and there is limited research on specific cancer types, among other issues.

Due to their black-box nature, most multimodal prognostic prediction algorithms lack interpretability, which is crucial for building trust in their decisions, especially in prognostic tasks. These models often output only risk predictions without showing how decisions are made, leading experts to question their reliance on relevant prognostic factors [27].

This study aims to develop a multimodal survival prediction method, MMsurv, for cancer prognosis, integrating genomic data, clinical information, and tissue pathology images. We addressed potential information loss in clinical data by adopting word embeddings from natural language processing, enhancing our model's ability to capture clinical nuances. Additionally, we introduced a novel multimodal fusion technique, multimodal fusion method based on compact bilinear pooling and transformer (MMF-CBPT), for effective integration of diverse data types. We also analyzed cell-type composition in highly attended model tiles to elucidate predictive capabilities. Our method, validated across six cancer types from the TCGA database, demonstrated superior performance compared to existing state-of-the-art approaches.

Materials and methods

Dataset preparation

This study obtained hematoxylin and eosin (H&E) stained formalin-fixed paraffin-embedded WSIs along with corresponding clinical and sequencing data from TCGA (https://portal.gdc.cancer.gov/). The data covered six different types of cancer, including breast cancer (BRCA), COAD, esophageal cancer (ESCA), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), and stomach adenocarcinoma (STAD).

WSIs were excluded if they had poor quality, lacked corresponding clinical and sequencing data, or did not indicate invasive carcinoma. For breast cancer, only female samples were included.

Clinical data, matched with H&E-stained pathology image data, had missing features removed. Previous studies emphasized the importance of T, N, and M staging, gender, and age in cancer prognosis [7, 28, 29]. These five features were used for all cancer types, except breast cancer, which included only female subjects. To maintain uniformity in clinical feature count, the number of lymph node examinations was added to the breast cancer model. Table 1 details the clinical data used in this study.

Moreover, previous studies have demonstrated a link between cancer genes and prognosis [30] Thus, Fragments Per Kilobase of transcript per Million mapped reads (FPKM) sequencing data for all corresponding cancer types were included in our analysis.

Overview of the MMsurv framework

We proposed a multimodal and multi-instance (MMsurv) model that integrates image, clinical features, and sequencing data for cancer prognosis prediction. The model's architecture is depicted in Fig. 1. In this framework, we initially engaged pathology experts to visually annotate regions of interest (ROIs) on the images, and then divided each ROI into tiles of size 512 × 512 pixels. After obtaining the tiles, we applied the Macenko method for color normalization of the images [31], and employed a convolutional neural network to extract pathological IFs from each tile (Fig. 1a).

Meanwhile, we applied word embedding to process clinical features and removed redundant and weakly correlated information from the sequencing data. The concatenated clinical and sequencing features were referred to as overall features (OFs) (Fig. 1b). We fed both IF and OF into the multi-instance model, which consists of two stages: multiple instance learning (MIL) model1, responsible for selecting important tiles related to prognosis after fusion with other data from the WSI, and MIL model2, which predicts the prognosis of patients (Fig. 1c). This architecture allows MMsurv to effectively exclude tile features with weak correlation to prognosis among millions of pixels, and the final scores are computed in the last linear layer.

In the next sections, we will present each core part of our MMsurv framework in detail.

Data processing

Recent advancements in deep learning have significantly improved image classification and prediction tasks, notably impacting medical imaging [32, 33]. However, WSI pose challenges due to their large size, high resolution, and the associated memory and computational efficiency issues. To mitigate these, we implemented manual annotations, preprocessing steps, and segmentation techniques to preprocess the WSI images [19].

Two experienced pathologists, each with over 20 years of experience, visually assessed WSIs and annotated tumor regions, as depicted by the dotted lines in Fig. 1a. For uniformity, each tumor region was divided into 512 × 512 pixel tiles. These tiles were then

Table 1. Clinical characteristics of different types of cancer

Clinical feature		BRCA	COAD	ESCA	LIHC	LUAD	STAD
T stage	T1	216	10	17	167	146	14
	T2	470	69	32	83	245	80
	T3	96	274	57	69	34	183
	T4	19	44	4	12	15	104
N stage	N0	390	230	50	230	274	116
	N1	266	96	45	4	83	106
	N2	90	71	8	0	46	79
	N3	53	0	5	0	2	75
	NX	2	0	2	97	5	5
M stage	M0	664	298	99	243	273	338
	M1	10	59	4	3	20	26
	MX	127	40	7	85	117	17
Age	<40 years	58	11	2	28	2	4
	40–60 years	396	113	60	136	135	122
	>60 years	347	273	48	167	273	255
Gender	Male	0	208	94	225	189	248
	Female	801	189	16	106	221	133
Lymph nodes	<5	290	—	—	—	—	—
Examined	5–15.	316	—	—	—	—	—
Number	>15	195	—	—	—	—	—

binarized to highlight the target regions, which include the tissue of interest and surrounding areas. Tiles with a target region ratio under 30% were discarded to minimize noise. To address inconsistencies resulting from tissue preparation, staining protocols, and image acquisition, we apply Macenko's color normalization technique to the tile images. It estimates the light intensity of each pixel to correct for illumination variations, ensuring consistent color, brightness, and contrast across different images. Figure 2 illustrates the images prior to and subsequent to the application of color normalization processing.

Next, tile images are processed using a pre-trained ResNet50 model, which encodes each 512 × 512 tile into a 2048-dimensional feature vector. These vectors capture detailed image representations, preparing them for further analysis and integration in the MMsurv model.

Clinical data processing

Although the clinical data provided by TCGA is highly structured, the direct application of this data in deep learning models poses significant challenges in terms of encoding methods. Traditional encoding techniques, such as one-hot encoding and label encoding, face considerable limitations when handling discrete clinical features like T stage, N stage, and M stage. One-hot encoding treats different categorical variables as entirely independent entities, overlooking the potential intrinsic relationships between them. Meanwhile, label encoding simply converts categorical variables into numerical values, which may obscure the intuitive meaning of these values, thereby impacting the interpretability of the model.

To overcome these challenges, we employed word embedding techniques, which are widely used in the field of natural language processing. This approach transforms discrete features into continuous vector representations with semantic relationships, preserving critical information from the original data while enhancing the model's ability to process these data [34]. We utilized a word vector table developed by De Vine *et al.* [35], trained on clinical records and a large MEDLINE medical journal abstract set, to measure semantic similarity between medical concepts. To eliminate the influence of feature scales, we applied the following min–max normalization to normalize the continuous values of clinical features to the range [0, 1].

$$\mathrm{x_{new}} = \frac{\mathrm{x} - \mathrm{x_{min}}}{\mathrm{x_{max}} - \mathrm{x_{min}}}, \tag{1}$$

where "max" represents the maximum value in the sample data and "min" represents the minimum value in the sample data.

Sequencing data processing

To combat the curse of dimensionality, we conducted feature selection through Cox analysis to determine the significance of each feature based on *P*-values. Only gene features with *P*-values under 0.01 were used to build a Deepsurv model [36], from which we selected the *N* genes that most improved prediction performance. These genes, deemed significantly related to prognosis, were incorporated into further modeling.

Gene signatures used across all cancer types can be found in Supplementary Table S1.

Prediction of cancer prognosis through multi-instance model

To enhance the prognostic prediction of cancer, we have undertaken optimization based on the current advanced MIL algorithm, DTFD, which was originally designed for weakly supervised classification purposes [37], as shown in Fig. 3.

Within the MIL framework, we treated each gigapixel WSI as a comprehensive collection of numerous instances, ranging from tens of thousands to even millions. These instances were derived from extracted features of tiles cropped from the patient's WSI, enabling a more nuanced analysis of the underlying characteristics. Formally, given *N* WSI images the bag $H_n = \{h_{n,1}, h_{n,2}, \ldots, h_{n,k}\}, n \in \{1, 2, \ldots, N\}$ consists of *K* instances, where "$h_{n,k}$" denotes the tile features extracted by the convolutional neural network h, for which we utilize the ResNet50 encoder.

To mitigate the potential overfitting problem caused by the limited number of available WSIs, we have implemented a *k*-means

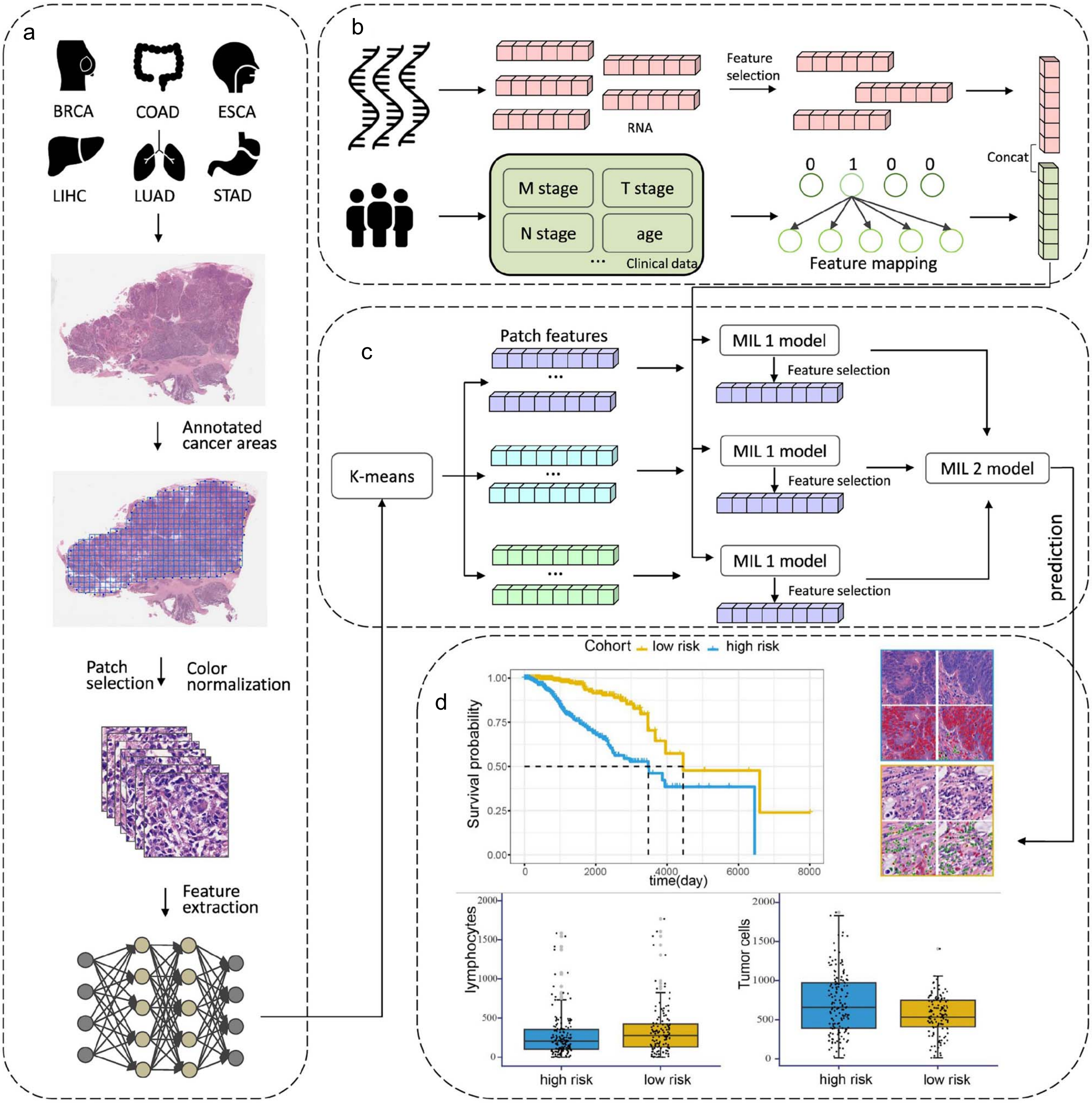

Figure 1. The overall framework of a multi-instance multi-cancer prognosis prediction model based on pathological images, clinical, and sequencing data. (a) Pathological image processing and feature extraction: Mark the region of interest on the pathological H&E stained image downloaded from TCGA, cut the region of interest into a 512 × 512 tiles and perform color normalization processing, and finally send it to ResNet50 to extract pathology image features. (b) Other modal feature processing: Use word embedding feature mapping for clinical features related to prognosis and feature screening for sequencing data to remove redundant information. (c) Model: The features of the pathological images are divided into three clusters using the k-means clustering algorithm, and each group is sent to the multiple instance learning (MIL) model1 with the corresponding processed other modality features to filter out and predict the first four tiles that are important to the task and send them to the MIL model2 gets the final predicted risk score. (d) Interpretation of the model: Kaplan–Meier analysis was performed on the model predictions to visualize patient stratification of low-risk and high-risk patients for each cancer type. We also linked model prediction performance to the cellular level through cellular quantification of tiles of high interest in both low-risk and high-risk patient cohorts.

clustering approach, whereby the instances of each patient were divided into M groups, i.e.$H_m = \left\{h_n^m | m = 1, 2, \ldots, M\right\}$. Considering that survival outcome information is accessible at the patient level rather than on individual tiles, we allocated patient-level survival information to the M groups.

MMsurv comprises two distinct multiple-instance learning models, each consisting of four components: a fully connected layer (L_p), an attention module (L_{attn}), a modality fusion module, and a prediction layer (L_{pred}). In the initial step, the features of each group are fed into the first multiple-instance learning model (MIL model1). This model employs a fully connected layer (L_p) with learnable weights ($w \in R^{512\times2048}$) and bias ($b \in R^{512}$) to map the group features ($h_n^m \in R^{\frac{K}{M}\times2048}$) to a more compact 512-dimensional features. Subsequently, the attention module (L_{attn}) assigns a relevance score to each feature, evaluating its significance in relation to patient-level prognosis. The attention

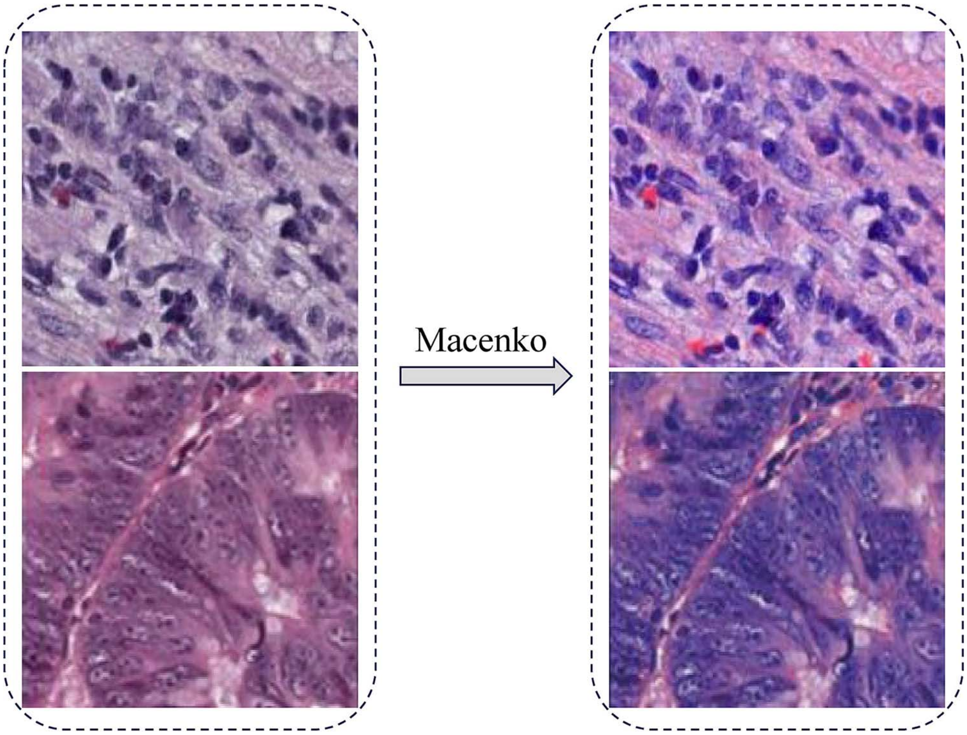

Figure 2. Comparison images before and after Macenko color normalization. The image on the left illustrates the state prior to color normalization, whereas the image on the right demonstrates the results following color normalization.

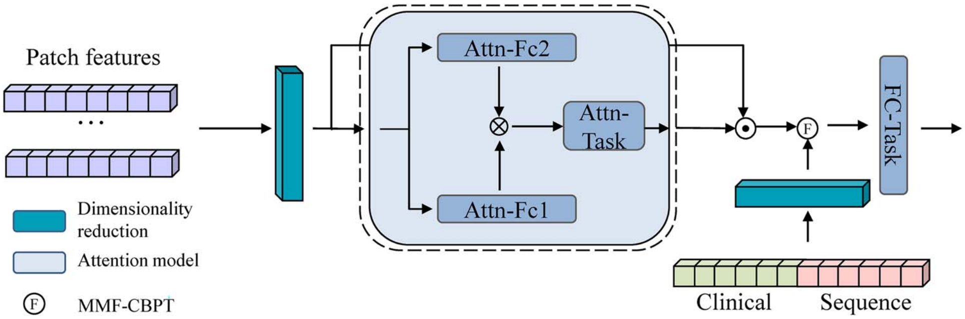

Figure 3. MIL model. The image tiles of each group are dimensionally reduced into key feature vectors from mechanism to obtain the relative importance of each tile to assign attention scores; the score weights the dimensions encoded by the pre-trained convolutional neural network and passed to the attention for peer review attention pool. And use MMF-CBPT to fuse the corresponding processed clinical and sequencing information into the image information, and finally get the prediction result through the prediction layer.

scores a_n^m are computed using the following method:

$$a_n^m = \frac{\exp\left\{w_a \times \left(\tanh V_a h_n^m \odot sigm\ U_a h_n^m\right)\right\}}{\sum_{j=1}^{\frac{K}{M}} \exp\left\{w_a \times \left(\tanh V_a h_n^j \odot sigm\ U_a h_n^j\right)\right\}}, \tag{2}$$

Subsequently, the attention scores, computed as weight coefficients, are utilized to perform an attention pooling operation, aggregating the tile features into a group feature h_{group} with a feature size of K/M:

$$h_{\text{group}} = \text{Attn}\,(A, H) = \sum_{m=1}^{\frac{K}{M}} a_n^m h_n^m, \tag{3}$$

To further enhance the predictive performance of the model, we integrated features from different modalities, including I_f and O_f. Considering the underexploited potential for data complementarity in existing multimodal fusion technologies, we propose a new multimodal fusion method MMF-CBPT. The final group-level prediction score is obtained by applying a fully connected layer (L_{pred}) to the features processed through MMF-CBPT.

The purpose of MIL model1 is to distill features and select patch features that are more relevant to prognosis. Previous studies have demonstrated the feasibility of obtaining signal strength indicating the prediction target for individual instances in attention-based MIL models through the application of Grad-CAM [38]. We apply the same calculation approach to MIL model1, obtaining the signal strength for each instance. The top d instances with the highest signal strength within each group are then forwarded to MIL model2. MIL model2 shares the same network architecture as MIL model1. Finally, the L_{pred} layer of MIL model2 generates the prediction results for each patient.

Multimodal fusion method based on compact bilinear pooling and transformer

To facilitate efficient integration and comprehensive analysis of multimodal data, this paper presents a novel fusion method, the

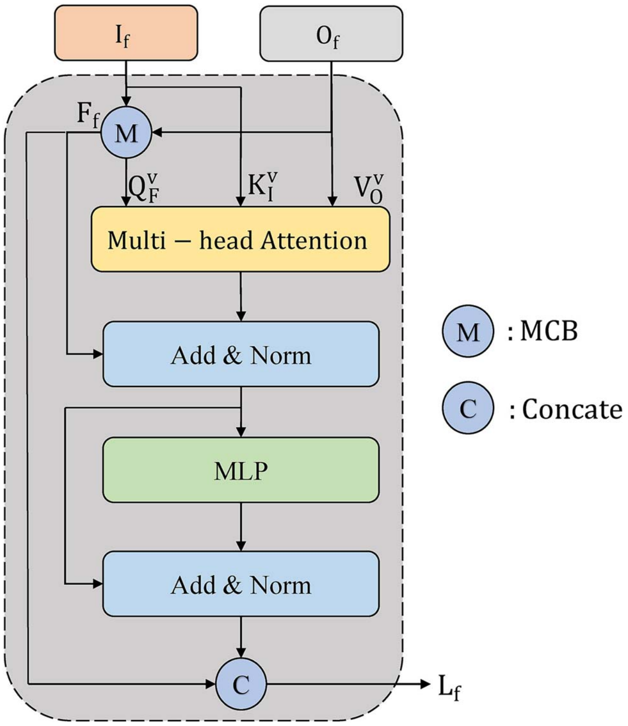

Figure 4. Illustration of the proposed MMF-CBPT architecture. Given the patient inputs I_f and O_f, the initially fused features F_f are produced through MCB pooling. These features, along with the original inputs, are then processed by a multi-head attention module to generate deeply fused features M_f. To avoid the loss of features, F_f and M_f are further concatenated to obtain the final fused feature L_f.

MMF-CBPT. This approach integrates multimodal compact bilinear (MCB) [39] and transformer technologies to synchronously process and amalgamate complex information from diverse data sources. The architecture of the MMF-CBPT model is depicted in Fig. 4.

The operation starts by taking inputs I_f and O_f into the MCB pooling module for initial fusion, producing preliminarily fused features F_f. This bilinear pooling efficiently captures complex interactions among features from different sources for a more comprehensive representation. Next, F_f, I_f, and O_f are input into the multi-head attention module of the transformer architecture, transforming them into query (Q), key (K), and value (V) matrices (Q_F^v, K_I^v, V_O^v), enabling the computation of deeply fused features M_f, which encapsulate comprehensive multimodal data insights. These features, along with F_f, are concatenated and processed through a linear layer to produce the final fused feature L_f. To enhance the model's nonlinear capabilities and prevent overfitting, a Rectified Linear Unit (ReLU) activation function and a Dropout layer follow the linear layer.

The overall computational formula is described as follows:

$$F_f = \text{MCB}\left(I_f, O_f\right), \tag{4}$$

$$M_f^* = \text{Norm}\left(\text{MHA}\left(Q_F^v, K_I^v, V_O^v\right) + Q_F^v\right), \tag{5}$$

$$M_f = \text{Norm}\left(\text{MLP}\left(M_f^*\right) + M_f^*\right), \tag{6}$$

$$Z = \text{Concate}\left(F_f, M_f\right), \tag{7}$$

$$L_f = \text{ReLU}\left(W_1 Z + b_1\right), \tag{8}$$

where W_1 and b_1 denote the weight and bias matrices of the linear layer, respectively.

Loss function and implementation details of the model

Cancer prognosis prediction is a survival data problem, and our objective is to maximize the concordance index (C-index). The computational formula for the C-index is delineated as follows:

$$\text{C} - \text{index} = \frac{1}{n}\sum\nolimits_{i\in\{1\ldots N\}} \sum_{y_j \geq y_i} I\left(h_\theta\left(x_j\right) > h_\theta\left(x_i\right)\right), \tag{9}$$

where n represents the number of comparable patient pairs, y_i represents the actual observed survival status of the ith patient, the neural network model h_θ is trained to predict the risk of survival.

Prior research has established the Cox partial likelihood loss [40] as the optimal method for achieving enhanced consistency. We employ the standard formula of the Cox loss to train MIL model1 and MIL model2.

The Cox loss function is defined as follows:

$$l\left(\theta\right) = -\sum\nolimits_{i:E_{i=1}} \left(h_\theta\left(x_i\right) - \log\left(\sum_{j:t_j \geq t_i} \exp\left(h_\theta\left(x_j\right)\right)\right)\right), \tag{10}$$

where E_i, t_i, and x_i represent the survival status, survival time, and data for each patient, respectively.

In this study, we divided the patient data instances into M groups and conducted a detailed search on the number of instances (d) sent to MIL2 in each group. Based on the results, the values were set to 3 and 4, respectively. Details of the search are provided in Tables 2 and 3.

MMsurv is designed following modern deep learning principles and implemented using the PyTorch platform. In this study, we employed the Adam optimizer to update the model parameters, setting an initial learning rate of 1e−3 and a weight decay of 1e−4. To ensure numerical stability, a small constant $\epsilon = 1\text{e}{-4}$ was included in the optimizer to prevent division by zero during gradient updates. The model was trained for a total of 100 epochs. Additionally, to mitigate overfitting, a dropout layer with a rate of 0.5 was applied after the first linear layer. During feature extraction by the backbone model, a batch size of 32 was used for each iteration, and four worker threads were employed to parallelize data loading for improved efficiency. To evaluate the model's accuracy, we employ five-fold cross-validation and ensure that WSIs from the same patient case are never simultaneously used as both training and testing data. For training, we utilized a server equipped with an Intel(R) Xeon(R) Gold 6242R @ 3.10 GHz CPU and NVIDIA Tesla A100 GPU.

Regarding the loss function, we employed the Cox partial likelihood loss function, given its widespread applicability and effectiveness in survival analysis. Additionally, the primary evaluation metric for model performance was the C-index, which is widely used to assess the predictive accuracy of survival models.

Interpretability of models and quantitative statistical analysis

To quantify our MMsurv model, we concatenated the predicted risks from each test set fold and plotted Kaplan–Meier curves based on survival time. We assessed survival distribution differences between patient groups using the log-rank test, considering

Table 2. The impact of M on model performance

Cancer site	1	2	3	5
BRCA	0.6022 (0.5720–0.6855)	0.6219 (0.5847–0.6956)	**0.6841** (0.6531–0.7532)	0.593 (0.5066–0.7157)
COAD	0.5218 (0.4557–0.6274)	0.6356 (0.5922–0.7338)	**0.6711** (0.6408–0.6898)	0.6194 (0.5306–0.6932)
ESCA	0.7192 (0.6861–0.7893)	**0.779** (0.7297–0.8299)	0.7686 (0.7471–0.8000)	0.7517 (0.6414–0.8166)
LIHC	**0.6665** (0.6224–0.7132)	0.5373 (0.4667–0.5862)	0.6633 (0.6365–0.7047)	0.6092 (0.5755–0.6433)
LUAD	0.5223 (0.4697–0.5628)	0.5867 (0.5205–0.6692)	0.6319 (0.6027–0.6781)	**0.6743** (0.6308–0.7147)
STAD	0.5203 (0.4340–0.5727)	0.5918 (0.5516–0.6581)	**0.6309** (0.5865–0.6678)	0.5993 (0.5191–0.6635)
Overall	0.5921	0.6254	**0.675**	0.6412

Bold values indicate the best performance achieved under the corresponding setting.

Table 3. The impact of d on model performance

Cancer site	2	4	6	8
BRCA	0.698 (0.624–0.7551)	**0.6841** (0.6531–0.7532)	0.5607 (0.5185–0.6599)	0.6631 (0.6147–0.7537)
COAD	0.5134 (0.4406–0.6219)	**0.6711** (0.6408–0.6898)	0.6654 (0.5469–0.7886)	0.6709 (0.6183–0.7585)
ESCA	0.6549 (0.6113–0.7187)	**0.7686** (0.7471–0.8000)	0.7166 (0.6481–0.7908)	0.6267 (0.5672–0.6944)
LIHC	0.5335 (0.4920–0.5855)	**0.6633** (0.6365–0.7047)	0.5679 (0.5071–0.6313)	0.6302 (0.5445–0.7235)
LUAD	0.5319 (0.4628–0.6439)	0.6319 (0.6027–0.6781)	**0.6487** (0.6212–0.7500)	0.5093 (0.4126–0.5890)
STAD	0.5611 (0.4700–0.6620)	**0.6309** (0.5865–0.6678)	0.5643 (0.4556–0.6930)	0.5367 (0.5010–0.5869)
Overall	0.5821	**0.675**	0.6206	0.6062

Bold values indicate the best performance achieved under the corresponding setting.

differences significant at P-values <0.05. Additionally, we analyzed cell information from highly attended tiles in WSIs with varying survival distributions to enhance model interpretability (Fig. 1d). For this, we used the HoverNet model, a pre-trained tool for nucleus segmentation and classification, to conduct detailed analyses of thousands of nuclei, providing precise insights into cell composition within pathological images [41].

Results

Comparison results of different backbone models

As a feature extractor for pathological images, the backbone model plays a crucial role in classification. To select an appropriate backbone model, the performances of ResNet18, ResNet50, ResNet101, and Xception in predicting patient outcomes from pathological images were compared under a single image data input scenario. Detailed results are summarized in Table 4. Among various cancers, ResNet50 had the highest C-index at 0.6750, surpassing other models (ResNet18: 0.6071, ResNet101: 0.5909, Xception: 0.6468); although it was outperformed by ResNet18 in LIHC and by ResNet101 and Xception in LUAD. Considering all factors, ResNet50 was chosen as the backbone model.

Comparison of results from different multimodal fusion methods

In this study, we conducted a comprehensive evaluation of the effectiveness of various multimodal data fusion methods in predicting cancer survival. For this purpose, we compared the MMF-CBPT method with other existing multimodal fusion approaches across multiple types of cancer. The specific fusion methods evaluated were:

1. MCB: Employs compact bilinear pooling to capture complex feature interactions across different modalities, enhancing understanding of multimodal data relationships. It approximates the bilinear model using random projection and fast Fourier transform, significantly lowering computational requirements.
2. BalanceMLA [42]: This framework utilizes real-time gradient modulation to dynamically monitor and adjust the optimization strategies for each modality based on their contributions. It also introduces dynamically varying Gaussian noise to prevent potential declines.

The experimental findings, detailed in Table 5, indicate that the multimodal fusion method MMF-CBPT, introduced in this study, surpasses unimodal methods, MCB, and BalanceMLA in achieving higher C-index values across various cancer datasets. Specifically, MMF-CBPT achieved a C-index of 0.7283 across all datasets, representing enhancements of 7.9%, 9.1%, and 2.2% over unimodal methods, MCB, and BalanceMLA, respectively. These findings corroborate the capability of MMF-CBPT to effectively harness the complementary information present in multimodal data, thereby enhancing the accuracy of survival predictions.

Table 4. Comparison results of different backbone models

Cancer site	ResNet18	ResNet50	ResNet101	Xception
BRCA	0.6182 (0.5473–0.7517)	**0.6841** (0.6531–0.7532)	0.6749 (0.5615–0.7781)	0.6256 (0.5986–0.7211)
COAD	0.5929 (0.4987–0.6303)	**0.6711** (0.6408–0.6898)	0.6276 (0.5818–0.6560)	0.6226 (0.5463–0.6958)
ESCA	0.6874 (0.6330–0.7705)	**0.7686** (0.7471–0.8000)	0.5461 (0.5229–0.5923)	0.7503 (0.6729–0.8406)
LIHC	**0.6641** (0.6599–0.6893)	0.6633 (0.6365–0.7047)	0.5084 (0.4297–0.6135)	0.6472 (0.6242–0.6743)
LUAD	0.5326 (0.4454–0.6438)	0.6319 (0.6027–0.6781)	0.6351 (0.5990–0.7212)	**0.6678** (0.5518–0.7639)
STAD	0.6076 (0.5469–0.7325)	**0.6309** (0.5865–0.6678)	0.5535 (0.4901–0.6658)	0.5673 (0.5127–0.6334)
Overall	0.6071	**0.675**	0.5909	0.6468

Bold values indicate the best performance achieved under the corresponding setting.

Table 5. Comparison of multimodal fusion methods

Cancer site	MMsurv Base	MMsurv MCB	MMsurv BalanceMLA	MMsurv MMF-CBPT
BRCA	0.6841 (0.6531–0.7532)	0.7184 (0.6684–0.7313)	0.7033 (0.6307–0.7313)	**0.7643** (0.6373–0.8279)
COAD	0.6711 (0.6408–0.6898)	0.6341 (0.5822–0.6790)	0.7702 (0.7326–0.8197)	**0.782** (0.6972–0.8610)
ESCA	0.7686 (0.7471–0.8000)	0.7503 (0.6918–0.7961)	0.7704 (0.6943–0.8235)	**0.7803** (0.7241–0.8372)
LIHC	0.6633 (0.6365–0.7047)	0.6415 (0.6285–0.6662)	0.6586 (0.5684–0.7404)	**0.6864** (0.6633–0.7084)
LUAD	0.6319 (0.6027–0.6781)	0.6135 (0.5724–0.6459)	**0.6937** (0.6185–0.7679)	0.6927 (0.6608–0.7293)
STAD	0.6309 (0.5865–0.6678)	0.649 (0.5783–0.6911)	**0.6789** (0.5708–0.7462)	0.6641 (0.6167–0.6950)
Overall	0.675	0.6678	0.7125	**0.7283**

Bold values indicate the best performance achieved under the corresponding setting.

The MMsurv models demonstrate superior performance relative to previous state-of-the-art techniques

To rigorously assess the performance of the MMsurv model, we employed five-fold cross-validation while maintaining a consistent partitioning strategy throughout all experiments. Feature extraction was performed using a ResNet50 model pre-trained on ImageNet, and the same hyperparameter settings were applied during training to ensure a fair comparison with existing state-of-the-art methods in the pathology domain. To ensure the accuracy of our comparisons, we re-implemented previous methods using their publicly available open-source code, thereby minimizing any potential bias.

The baseline MMsurv model's performance is illustrated in Table 6 and Fig. 5a, where MMsurv outperforms other existing methods across six different cancer types, achieving an overall C-index of 0.6750. This value is notably higher than those of MCAT [14] (C-index: 0.5746) and DeepAttnMISL [43] (C-index: 0.5948). Notably, the MMsurv baseline model demonstrated substantial performance improvements in BRCA and ESCA, with increases of ∼14% and 11% over MCAT and DeepAttnMISL, respectively.

To further demonstrate the advantages of the MMsurv model, we also compared it with two multimodal models, MOTCat and SURVPATH, as shown in Table 7. MMsurv excelled in four out of six cancer types, achieving an overall C-index of 0.7283, surpassing both MOTCat [44] (C-index: 0.6969) and SURVPATH [45] (C-index: 0.7067).

In conclusion, these findings indicate that the MMsurv model effectively harnesses multimodal data to predict patient prognosis with greater precision, underscoring its significant potential in the field of cancer prognosis.

The fusion of multimodal data can improve the model prediction effect

We conducted ablation experiments, detailed in Table 8 and Fig. 5b, comparing the MMsurv baseline model to three other models: Image_Clinical (combines H&E images with clinical data), Image_Embedding (combines H&E images with clinically mapped word embeddings), and Image_Embedding_Sequence (combines H&E images, clinically mapped word embeddings, and sequencing data).

It can be seen that in a separate view of the two feature sets fused with the image, clinical features seem to play a crucial role in survival prediction, as their overall C-index (0.6273) surpasses that of single sequence features (0.6007) and achieves higher C-indices in 5 out of 6 cancer types.

When combining clinical and IFs, the resulting C-indices for COAD, LIHC, and STAD cancers are lower, with an overall C-index of 0.6734, below that of IFs alone. This suggests that directly inputting clinical data may prevent the model from learning effective semantic relationships and introduce incorrect associations. However, mapping clinical information via word embedding enhances IFs, improving the overall C-index to 0.7105, an increase of about 3.5%.

Table 6. Comparison of C-index between the MMsurv baseline model and other models

Cancer site	MCAT	DeepAttnMSIL	MMsurv
BRCA	0.5600 (0.4890–0.6150)	0.6021 (0.5629–0.6568)	**0.6841** (0.6531–0.7532)
COAD	0.5460 (0.4660–0.6170)	0.6059 (0.5239–0.6812)	**0.6711** (0.6408–0.6898)
ESCA	0.6195 (0.5200–0.7612)	0.6297 (0.5057–0.7558)	**0.7686** (0.7471–0.8000)
LIHC	0.6180 (0.5630–0.6840)	0.5778 (0.5277–0.6477)	**0.6633** (0.6365–0.7047)
LUAD	0.5480 (0.4890–0.5970)	0.5695 (0.5278–0.6079)	**0.6319** (0.6027–0.6781)
STAD	0.5560 (0.4940–0.5980)	0.5837 (0.5242–0.6542)	**0.6309** (0.5865–0.6678)
Overall	0.5746	0.5948	**0.675**

Bold values indicate the best performance achieved under the corresponding setting.

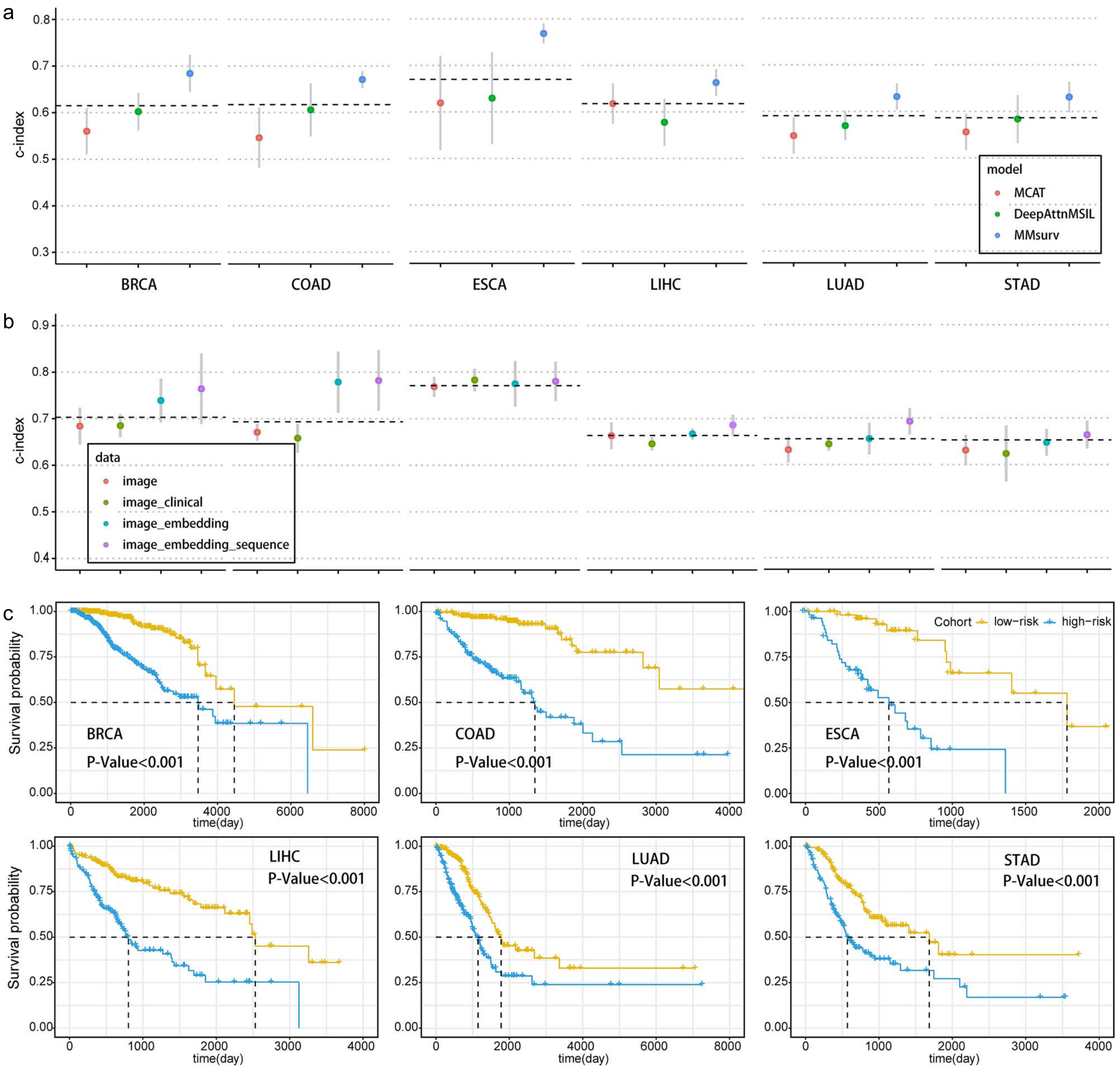

Figure 5. The result display of the model. (a) Comparison of the performance of the MMsurv baseline model and other methods in the C-index. (b) Comparison of the C-index of the MMsurv model in different modal data inputs. (c) MMsurv performance evaluation based on Kaplan–Meier curve.

Our study integrated sequencing data into the MMsurv model, enhancing predictive accuracy for all tumor types with a 1.8% increase in the C-index. We validated the multimodal model by categorizing patient risk scores into low-risk and high-risk groups using the median as a threshold. Kaplan–Meier curves visualized the stratification quality, and the log-rank test assessed the statistical significance. As depicted in Fig. 5c, the survival curves and *P*-values confirmed the alignment between predicted and actual survival times for these groups. The survival curves for the low and high-risk groups exhibit a clear separation, and

Table 7. Comparison of the C-index between the MMsurv model and other multimodal models

Cancer site	MOTCat	SURVPATH	MMsurv
BRCA	0.6870	0.7374	**0.7643**
	(0.6327–0.7886)	(0.6931–0.7787)	(0.6373–0.8279)
COAD	0.7218	0.7569	**0.7820**
	(0.6679–0.8106)	(0.7112–0.8124)	(0.6972–0.8610)
ESCA	0.7638	**0.7924**	0.7803
	(0.6997–0.8150)	(0.7033–0.8932)	(0.7241–0.8372)
LIHC	0.6722	0.6413	**0.6864**
	(0.6127–0.7132)	(0.6215–0.6860)	(0.6633–0.7084)
LUAD	**0.6973**	0.6619	0.6927
	(0.6297–0.7486)	(0.6014–0.7120)	(0.6608–0.7293)
STAD	0.6395	0.6506	**0.6641**
	(0.6086–0.6849)	(0.6182–0.6834)	(0.6167–0.6950)
Overall	0.6969	0.7067	**0.7283**

Bold values indicate the best performance achieved under the corresponding setting.

Table 8. Comparison results of different data fusions

Caner site	MMsurv				Deepsurv	
	Image	Image	Image	Image Embedding sequence	Clinical	Sequence
		Clinical	Embedding			
BRCA	0.6841	0.6852	0.7391	**0.7643**	0.6136	0.6171
	(0.6531–0.7532)	(0.6456–0.7038)	(0.6821–0.8079)	(0.6373–0.8279)	(0.5119–0.7637)	(0.5550–0.7021)
COAD	0.6711	0.6583	0.7787	**0.782**	0.703	0.5791
	(0.6408–0.6898)	(0.6300–0.7043)	(06897–0.8501)	(0.6972–0.8610)	(0.6200–0.7995)	(0.5355–0.6325)
ESCA	0.7686	**0.783**	0.7748	0.7803	0.6459	0.686
	(0.7471–0.8000)	(0.7582–0.8143)	(0.7100–0.8256)	(0.7241–0.8372)	(0.5106–0.7915)	(0.5733–0.8846)
LIHC	0.6633	0.646	0.6672	**0.6864**	0.6505	0.5915
	(0.6365–0.7047)	(0.6304–0.6652)	(0.6567–0.6745)	(0.6633–0.7084)	(0.5501–0.7479)	(0.5694–0.658)
LUAD	0.6319	0.6442	0.6557	**0.6927**	0.5784	0.5634
	(0.6027–0.6781)	(0.6213–0.6579)	(0.6150–0.7003)	(0.6608–0.7293)	(0.5406–0.6195)	(0.5397–0.6322)
STAD	0.6309	0.6235	0.6476	**0.6641**	0.5724	0.5669
	(0.5865–0.6678)	(0.5398–0.6815)	(0.6176–0.6788)	(0.6167–0.6950)	(0.5025–06258)	(0.5166–0.6561)
Overall	0.675	0.6734	0.7105	**0.7283**	0.6273	0.6007

Bold values indicate the best performance achieved under the corresponding setting.

the log-rank test yields *P*-values significantly below 0.001 for all cancer types (**BRCA** $P = 1.83 \times 10^{-9}$, **COAD** $P = 3.16 \times 10^{-11}$, **ESCA** $P = 4.53 \times 10^{-6}$, **LIHC** $P = 1.24 \times 10^{-8}$, **LUAD** $P = 4.78 \times 10^{-6}$, **STAD** $P = 1.47 \times 10^{-5}$). The differentiation between high-risk and low-risk populations assists in guiding personalized treatments and supports clinical decision-making.

Relationship between MMsurv regions of high interest and cell types

To demonstrate the superiority of our MMsurv model, we visualized cell-level prediction results and highlighted important regions. We used the HoverNet algorithm for cell nucleus segmentation and classification in histopathological images, followed by statistical analysis with box plots to compare cell types between high-risk and low-risk groups. Figure 6 shows semantic segmentation maps versus original images of high-attention tissue regions for both risk groups across various cancer types, along with a comparison of the quantitative cell type distribution in these regions.

Due to the lower prevalence of epithelial, stromal, and necrotic cells in the tiles, we focused on analyzing tumor cells and lymphocytes, which are crucial in tumor morphology and immune infiltration in pathological images. We found that low-risk patients consistently had more lymphocytes in highly scrutinized areas, while high-risk patients had more tumor cells in these regions across various cancer types. Additionally, in 5 out of 6 cancer types, statistically significant differences were observed in the highly scrutinized regions concerning lymphocytes (**BRCA** $P = 1.5 \times 10^{-3}$, **COAD** $P = 9.2 \times 10^{-3}$, **LIHC** $P = 3.3 \times 10^{-2}$, **LUAD** $P = 6.2 \times 10^{-5}$, **STAD** $P = 1.7 \times 10^{-3}$) and tumor cells (**BRCA** $P = 3.6 \times 10^{-3}$, **COAD** $P = 8.7 \times 10^{-3}$, **LIHC** $P = 1.6 \times 10^{-3}$, **LUAD** $P = 1.4 \times 10^{-2}$, **STAD** $P = 9.3 \times 10^{-3}$), as illustrated in Fig. 6a, b, and d–f. These findings demonstrate the interpretability of the MMsurv model, enabling the localization of the most representative regions within the WSI.

Discussion

Predictive prognosis is crucial for guiding cancer treatment decisions, yet current methods often fail to effectively utilize the vast multimodal data available, resulting in limited accuracy [11, 46]. We introduced MMsurv, a multi-instance, multimodal deep learning framework that integrates histopathological images, clinical data, and sequencing data using a novel fusion technique, MMF-CBPT. Validated with 2364 WSIs across six cancer types, MMsurv uses a multi-instance model to selectively focus on key tiles that boost predictive performance, showcasing its potential in diverse tumor types and resource-constrained settings.

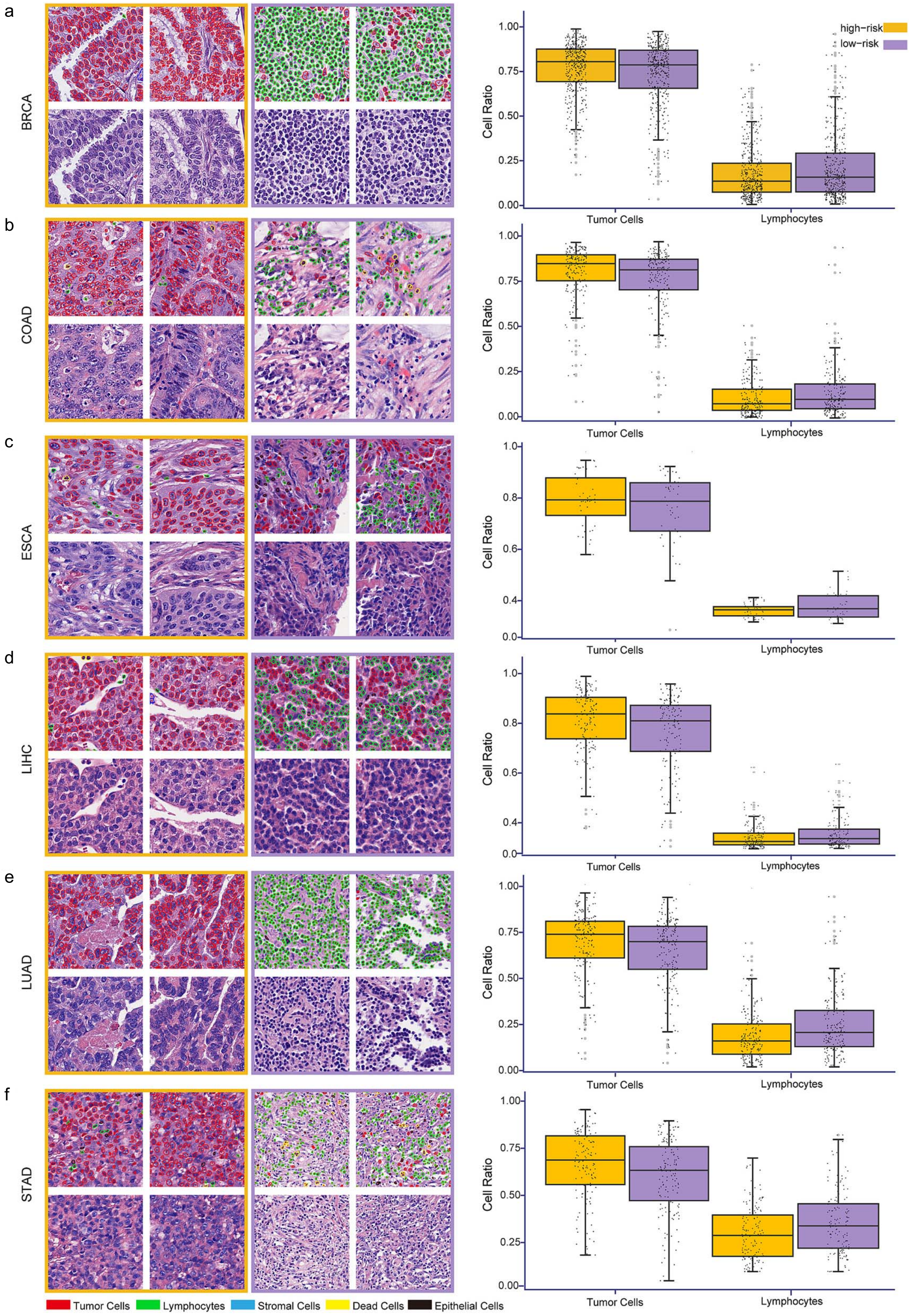

Figure 6. Visualization of high attention tiles. For the 12 high-attention tiles screened for each cancer patient, the cell segmentation is performed, and the cell segmentation map (upper left) and the original image (lower left) are obtained. At the same time, the tumor cells and lymphocytes in the high-low risk patient group are quantitatively compared (right picture). (a) BRCA, (b) COAD, (c) ESCA, (d) LIHC, (e) LUAD, and (f) STAD.

The rapid progress in survival prediction models highlights the benefits of multimodal data fusion, which integrates complementary information from various modalities to enhance predictive accuracy [24, 47]. Building on this, our study introduces a novel fusion approach, MMF-CBPT, that combines bilinear pooling with Transformer technologies to effectively merge pathological image data, clinical data, and sequencing data. This method has significantly outperformed existing models, achieving accuracy improvements of 9.1% over MCB and 2.2% over BalanceMLA.

Traditional methods that input clinical data directly into models often miss variable correlations, leading to information loss. This was addressed by using word embeddings to process clinical data, which improved MMsurv's performance across various cancers. Further enhancements were achieved by incorporating sequencing data, showing our model's ability to effectively integrate complementary information across modalities and enhance data utilization.

Due to the vast number of parameters in deep learning models, they have often been criticized as black boxes with limited interpretability [48]. By employing the HoverNet model, we quantitatively explored the relevance of the model's highly attended tiles to the specific predictive task and linked these regions to cell-level annotations, providing biological insights to our model. Our results demonstrate high-risk patients have more highly attended tiles and less lymphocytes while the opposite is true for low-risk patients.

Although MMsurv has shown enhanced prognostic performance across various cancer types, there are several avenues for expansion and improvement. Limited availability of sequencing, clinical, and histopathological data highlights the need for more cancer samples to improve prognostic capabilities. Our study, based on TCGA patients, necessitates validation using external independent datasets to confirm its generalizability. Future work should also explore advanced computational methods to better integrate multimodal data and optimize MMsurv's performance. Additionally, expanding the scope to include more cancer types is crucial for a comprehensive evaluation of the model.

Key Points

- We present MMsurv, a novel multimodal multi-instance framework that integrates clinical information, sequencing data, and hematoxylin and eosin-stained whole-slide images (WSIs) for predicting survival in patients with multi-cancer. Our experimental results consistently demonstrate the superior performance of MMsurv compared to other state-of-the-art approaches across six different cancer types.
- We proposed a multimodal fusion method, multimodal fusion method based on compact bilinear pooling and transformer, which achieves precise integration of pathological images, clinical data, and sequencing data, and demonstrates superior performance across multiple cancer types.
- We innovatively employ word embedding techniques on clinical data to enhance the model's ability to learn the underlying associations among clinical variables and improve its predictive performance.
- In order to enhance the interpretation of MMsurv, we employ cell segmentation to systematically analyze the cellular composition within the WSI tiles that receive high attention from the model.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Funding

This research was supported by the National Natural Science Foundation of China (Grant No. 62162025), the Hainan Provincial Natural Science Foundation of China (Grant No. 122RC653), and the Natural Science Foundation of Hubei Province (Grant No. 2022CFB283) awarded to Jialiang Yang.

Data availability

Due to privacy restrictions, the annotated data supporting the findings of this study are not publicly available but can be obtained from the corresponding author upon reasonable request.

References

1. Bray F, Laversanne M, Weiderpass E. *et al.* The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer* 2021;**127**:3029–30. https://doi.org/10.1002/cncr.33587
2. Torre LA, Bray F, Siegel RL. *et al.* Global cancer statistics, 2012. *CA Cancer J Clin* 2015;**65**:87–108. https://doi.org/10.3322/caac.21262
3. Bray F, Laversanne M, Sung H. *et al.* Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2024;**74**:229–63. https://doi.org/10.3322/caac.21834
4. Huang S, Yang J, Fong S. *et al.* Artificial intelligence in cancer diagnosis and prognosis: opportunities and challenges. *Cancer Lett* 2020;**471**:61–71. https://doi.org/10.1016/j.canlet.2019.12.007
5. Kim J, Braun D, Dhingra TG. *et al.* Clinical factors associated with gastric cancer in individuals with Lynch syndrome. *Clin Gastroenterol Hepatol* 2020;**18**:830–837.e1. https://doi.org/10.1016/j.cgh.2019.07.012
6. Goswami S, Peipert BJ, Mongelli MN. *et al.* Clinical factors associated with worse quality-of-life scores in United States thyroid cancer survivors. *Surgery* 2019;**166**:69–74. https://doi.org/10.1016/j.surg.2019.01.034
7. Xu C, Xiong B. Prognostic nomograms for patients with primary sarcomatoid carcinoma of the urinary bladder: based on the SEER database. *Urol J* 2024;**21**:87–97. https://doi.org/10.22037/uj.v20i.7595
8. Amin MB, Greene FL, Edge SB. *et al.* The eighth edition AJCC cancer staging manual: continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA Cancer J Clin* 2017;**67**:93–9. https://doi.org/10.3322/caac.21388

9. Beck AH. Open access to large scale datasets is needed to translate knowledge of cancer heterogeneity into better patient outcomes. *PLoS Med* 2015;**12**:e1001794. https://doi.org/10.1371/journal.pmed.1001794
10. Fisher R, Pusztai L, Swanton C. Cancer heterogeneity: implications for targeted therapeutics. *Br J Cancer* 2013;**108**:479–85. https://doi.org/10.1038/bjc.2012.581
11. Gulati S, Martinez P, Joshi T. *et al.* Systematic evaluation of the prognostic impact and intratumour heterogeneity of clear cell renal cell carcinoma biomarkers. *Eur Urol* 2014;**66**:936–48. https://doi.org/10.1016/j.eururo.2014.06.053
12. Maroto P, Rini B. Molecular biomarkers in advanced renal cell carcinoma. *Clin Cancer Res* 2014;**20**:2060–71. https://doi.org/10.1158/1078-0432.CCR-13-1351
13. Shao W, Han Z, Cheng J. *et al.* Integrative analysis of pathological images and multi-dimensional genomic data for early-stage cancer prognosis. *IEEE Trans Med Imaging* 2020;**39**:99–110. https://doi.org/10.1109/TMI.2019.2920608
14. Chen RJ, Lu MY, Williamson DF., *et al.* : Pan-cancer integrative histology-genomic analysis via multimodal deep learning**.** *Cancer Cell* 2022, **40:**865–878. e866.
15. Liang JY, Wang DS, Lin HC. *et al.* A novel ferroptosis-related gene signature for overall survival prediction in patients with hepatocellular carcinoma. *Int J Biol Sci* 2020;**16**:2430–41. https://doi.org/10.7150/ijbs.45050
16. Wu Z, Wang L, Wen Z. *et al.* Integrated analysis identifies oxidative stress genes associated with progression and prognosis in gastric cancer. *Sci Rep* 2021;**11**:3292.
17. Lovly CM, Salama AK, Salgia R. Tumor heterogeneity and therapeutic resistance. *Am Soc Clin Oncol Educ Book* 2016;**35**:e585–93. https://doi.org/10.1200/EDBK_158808
18. Alizadeh AA, Aranda V, Bardelli A. *et al.* Toward understanding and exploiting tumor heterogeneity. *Nat Med* 2015;**21**:846–53. https://doi.org/10.1038/nm.3915
19. Cheerla A, Gevaert O. Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics* 2019;**35**:i446–54. https://doi.org/10.1093/bioinformatics/btz342
20. Srinidhi CL, Ciga O, Martel AL. Deep neural network models for computational histopathology: a survey. *Med Image Anal* 2021;**67**:101813. https://doi.org/10.1016/j.media.2020.101813
21. Yang J, Ju J, Guo L. *et al.* Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning. *Comput Struct Biotechnol J* 2022;**20**:333–42. https://doi.org/10.1016/j.csbj.2021.12.028
22. Huang K, Lin B, Liu J. *et al.* Predicting colorectal cancer tumor mutational burden from histopathological images and clinical information using multi-modal deep learning. *Bioinformatics* 2022;**38**:5108–15. https://doi.org/10.1093/bioinformatics/btac641
23. Skrede OJ, De Raedt S, Kleppe A. *et al.* Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *Lancet* 2020;**395**:350–60. https://doi.org/10.1016/S0140-6736(19)32998-8
24. Abbet C, Zlobec I, Bozorgtabar B. *et al.* Divide-and-rule: self-supervised learning for survival analysis in colorectal cancer. In: Martel AL, Abolmaesumi P, Stoyanov D. *et al.* (eds), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020: 23rd International Conference, October 4–8, 2020, Lima, Peru, Proceedings, Part V. Lecture Notes in Computer Science*, vol **12265**. Cham, Switzerland: Springer; 2020. pp. 480–89.
25. Wang Z, Li R, Wang M. *et al.* GPDBN: deep bilinear network integrating both genomic data and pathological images for breast cancer prognosis prediction. *Bioinformatics* 2021;**37**:2963–70. https://doi.org/10.1093/bioinformatics/btab185
26. Li R, Wu X, Li A. *et al.* HFBSurv: hierarchical multimodal fusion with factorized bilinear models for cancer survival prediction. *Bioinformatics* 2022;**38**:2587–94. https://doi.org/10.1093/bioinformatics/btac113
27. Cadario R, Longoni C, Morewedge CK. Understanding, explai,ning, and utilizing medical artificial intelligence. *Nat Hum Behav* 2021;**5**:1636–42. https://doi.org/10.1038/s41562-021-01146-0
28. Wang BY, Huang JY, Chen HC. *et al.* The comparison between adenocarcinoma and squamous cell carcinoma in lung cancer patients. *J Cancer Res Clin Oncol* 2020;**146**:43–52. https://doi.org/10.1007/s00432-019-03079-8
29. Wang J, Li S, Liu Y. *et al.* Metastatic patterns and survival outcomes in patients with stage IV colon cancer: a population-based analysis. *Cancer Med* 2020;**9**:361–73. https://doi.org/10.1002/cam4.2673
30. Nagy A, Munkacsy G, Gyorffy B. Pancancer survival analysis of cancer hallmark genes. *Sci Rep* 2021;**11**:6047.
31. Macenko M, Niethammer M, Marron JS. *et al.* A method for normalizing histology slides for quantitative analysis. In: Prince JL, Pham DL, Myers KJ. (eds). *Proceedings of the 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI 2009); June 28–July 1, 2009; Boston, MA, USA*. Piscataway, NJ, USA: IEEE; 2009. pp. 1107–10.
32. Wang H, Xing F, Su H. *et al.* Novel image markers for non-small cell lung cancer classification and survival prediction. *BMC Bioinformatics* 2014;**15**:310.
33. Yao Y, Lv Y, Tong L. *et al.* ICSDA: a multi-modal deep learning model to predict breast cancer recurrence and metastasis risk by integrating pathological, clinical and gene expression data. *Brief Bioinform*. 2022;**23**:bbac448. https://doi.org/10.1093/bib/bbac448
34. Bengio Y, Senecal JS. Adaptive importance sampling to accelerate training of a neural probabilistic language model. *IEEE Trans Neural Netw* 2008;**19**:713–22. https://doi.org/10.1109/TNN.2007.912312
35. De Vine L, Zuccon G, Koopman B. *et al.* Medical semantic similarity with a neural language model. In: Snoek CGM, Sebastiani F. (eds). *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (CIKM 2014); November 3–7, 2014; Shanghai, China*. New York, NY, USA: Association for Computing Machinery (ACM); 2014. pp. 1819–22.
36. Katzman JL, Shaham U, Cloninger A. *et al.* DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol* 2018;**18**:24. https://doi.org/10.1186/s12874-018-0482-1
37. Zhang H, Meng Y, Zhao Y. *et al.* Dtfd-mil: double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: Koltun V, Torralba A. (eds). *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022); June 19–24, 2022; New Orleans, LA, USA*. Piscataway, NJ, USA: IEEE Computer Society; 2022. pp. 18802–12.
38. Selvaraju RR, Cogswell M, Das A. *et al.* Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Koltun V, Daniilidis K. (eds). *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017); October 22–29, 2017; Venice, Italy*. Piscataway, NJ, USA: IEEE Computer Society; 2017. pp. 618–26.
39. Fukui A, Park DH, Yang D. *et al.* Multimodal compact bilinear pooling for visual question answering and visual grounding. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016); November 2–6, 2016; Austin,*

Texas, USA. Stroudsburg, PA, USA: Association for Computational Linguistics; 2016. pp. 457–68.

40. Ching T, Zhu X, Garmire LX. Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Comput Biol* 2018;**14**:e1006076. https://doi.org/10.1371/journal.pcbi.1006076
41. Graham S, Vu QD, Raza SEA. *et al.* Hover-Net: simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med Image Anal* 2019;**58**:101563. https://doi.org/10.1016/j.media.2019.101563
42. Peng X, Wei Y, Deng A. *et al.* Balanced multimodal learning via on-the-fly gradient modulation. In: Koltun V, Torralba A. (eds). *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022); June 19–24, 2022; New Orleans, LA, USA*. Piscataway, NJ, USA: IEEE Computer Society; 2022. pp. 8238–47.
43. Yao J, Zhu X, Jonnagaddala J. *et al.* Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Med Image Anal* 2020;**65**:101789. https://doi.org/10.1016/j.media.2020.101789
44. Xu Y, Chen H. Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In: Koltun V, Daniilidis K. (eds). *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2023); October 2–6, 2023; Paris, France*. Piscataway, NJ, USA: IEEE Computer Society; 2023. pp. 21241–51.
45. Jaume G, Vaidya A, Chen RJ. *et al.* Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In: Koltun V, Torralba A. (eds). *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024); June 17–21, 2024; Seattle, WA, USA*. Piscataway, NJ, USA: IEEE Computer Society; 2024. pp. 11579–90.
46. Jiang S, Suriawinata AA, Hassanpour S. MHAttnSurv: multi-head attention for survival prediction using whole-slide pathology images. *Comput Biol Med* 2023;**158**:106883. https://doi.org/10.1016/j.compbiomed.2023.106883
47. Fu X, Patrick E, Yang JYH. *et al.* Deep multimodal graph-based network for survival prediction from highly multiplexed images and patient variables. *Comput Biol Med* 2023;**154**:106576. https://doi.org/10.1016/j.compbiomed.2023.106576
48. Xu J, Xue K, Zhang K. Current status and future trends of clinical diagnoses via image-based deep learning. *Theranostics* 2019;**9**:7556–65. https://doi.org/10.7150/thno.38065

Briefings in Bioinformatics, 2025, **26(3)**, bbaf209
https://doi.org/10.1093/bib/bbaf209
Problem Solving Protocol