OXFORD

Briefings in Bioinformatics, 2025, **26(2)**, bbaf121

https://doi.org/10.1093/bib/bbaf121

Problem Solving Protocol

Deep learning-driven survival prediction in pan-cancer studies by integrating multimodal histology-genomic data

Yongfei Hu[1,‡], Xinyu Li[2,‡], Ying Yi[1], Yan Huang[3,*], Guangyu Wang[4,*], Dong Wang[1,2,*]

[1]Dermatology Hospital, Southern Medical University, No. 2, Lujing Road, Yuexiu District, Guangzhou 510091, China
[2]Department of Bioinformatics, School of Basic Medical Sciences, Guangdong Province Key Laboratory of Molecular Tumor Pathology, Southern Medical University, 1023 Shatai South Road, Baiyun District, Guangzhou 510515, China
[3]Cancer Research Institute, School of Basic Medical Sciences, Southern Medical University, 1023 Shatai South Road, Baiyun District, Guangzhou 510515, China
[4]Department of Gastrointestinal Medical Oncology, Harbin Medical University Cancer Hospital, No. 150 Haping Road, Nangang District, Harbin 150000, China
*Corresponding authors. Yan Huang, Cancer Research Institute, School of Basic Medical Sciences, Southern Medical University, 1023 Shatai South Road, Baiyun District, Guangzhou 510515, China. E-mail: huangyan24@smu.edu.cn; Guangyu Wang, Department of Gastrointestinal Medical Oncology, Harbin Medical University Cancer Hospital, No. 150 Haping Road, Nangang District, Harbin 150000, China. E-mail: guangyuwang@hrbmu.edu.cn; Dong Wang, Dermatology Hospital, Southern Medical University, 1023 Shatai South Road, Baiyun District, Guangzhou 510091, China. E-mail: wangdong79@smu.edu.cn
‡Yongfei Hu and Xinyu Li contributed equally to this work.

Abstract

Accurate cancer prognosis is essential for personalized clinical management, guiding treatment strategies and predicting patient survival. Conventional methods, which depend on the subjective evaluation of histopathological features, exhibit significant inter-observer variability and limited predictive power. To overcome these limitations, we developed cross-attention transformer-based multimodal fusion network (CATfusion), a deep learning framework that integrates multimodal histology-genomic data for comprehensive cancer survival prediction. By employing self-supervised learning strategy with TabAE for feature extraction and utilizing cross-attention mechanisms to fuse diverse data types, including mRNA-seq, miRNA-seq, copy number variation, DNA methylation variation, mutation data, and histopathological images. By successfully integrating this multi-tiered patient information, CATfusion has become an advanced survival prediction model to utilize the most diverse data types across various cancer types. CATfusion's architecture, which includes a bidirectional multimodal attention mechanism and self-attention block, is adept at synchronizing the learning and integration of representations from various modalities. CATfusion achieves superior predictive performance over traditional and unimodal models, as demonstrated by enhanced C-index and survival area under the curve scores. The model's high accuracy in stratifying patients into distinct risk groups is a boon for personalized medicine, enabling tailored treatment plans. Moreover, CATfusion's interpretability, enabled by attention-based visualization, offers insights into the biological underpinnings of cancer prognosis, underscoring its potential as a transformative tool in oncology.

Keywords: cancer survival prediction; deep learning; multimodal data fusion; self-supervised learning

Introduction

Cancer prognosis is a critical component in the clinical management of patients, guiding treatment strategies and offering insights into survival outcomes [1]. In the standard clinical practice for numerous cancers, doctors manually examine histological aspects of tumors, including their invasiveness, cellular abnormalities, tissue death, and cell division rates [2–4]. These evaluations help in grading and staging cancer, which classifies patients into risk categories to inform treatment choices [5, 6]. For example, the TNM system evaluates the primary tumor's characteristics like size, progression, and abnormality to assign a stage [7]. This stage influences decisions on treatment plans, surgery eligibility, radiation therapy levels, and other therapeutic approaches. However, it has been shown that the subjective evaluation of pathological features can be quite inconsistent, with significant differences in assessments between different observers and even within the same observer over time. As a result, patients classified in the same grade or stage may experience notably diverse treatment outcomes [8]. During the past years, to fully exploit the hidden information of histopathological images, a number of computational approaches have been developed to retrieve a wealth of features from these images [9–12]. These features that quantitatively capture the cellular characteristics such as size, form, arrangement, and texture were used to predict the patient survival [13]. Alternatively, other methodologies leverage genomic data to forecast cancer survival rates. This is because cancer is intimately linked to genetic mutations and irregular gene expression that disrupt standard cell functions and biological mechanisms [14]. Therefore, delving into genomic data is highly pertinent for predicting cancer survival [15–17].

Those traditional prognostic models, relying primarily on clinical parameters and unimodal data sources, have made significant contributions to oncology but are increasingly recognized as

Received: October 10, 2024. **Revised:** February 10, 2025. **Accepted:** February 28, 2025

limited in their ability to capture the complexity and heterogeneity of cancer [18, 19]. As medical data continue to accumulate, the evolution of precision medicine demands a more nuanced understanding, necessitating the integration of diverse data types, including genomics, histopathology, and clinical data [20, 21]. The integration of macroscopic information from images and microscopic information from molecular profiles to predict the outcomes of cancer patients has become a mainstream focus of current research [22, 23]. Recent advances in computational pathology and machine learning have paved the way for a new generation of prognostic models [24, 25]. These models, exemplified by Pathomic Fusion [26], CAMR [27], and GPDBN [28], leverage deep learning techniques to integrate multimodal data, offering a more comprehensive and objective assessment of cancer survival. The integration of heterogeneous data sources, such as copy number variations, mRNA expression, and histopathological images, allows these models to capture the intricate interplay between molecular and morphological cancer characteristics. However, the fusion techniques employed in these methods, such as the Kronecker product and low-rank multimodal fusion, still struggle to effectively align the diverse and heterogeneous multimodal data. This can often result in markedly different representations within the feature space, known as the modality gap problem. As a result, the presence of modality gaps impedes the comprehensive integration of multimodal information, significantly constraining further advancements in the predictive performance of cancer survival. In biological terms, these gaps in data representation can hinder the seamless merging of information from various biological data types, thereby limiting the potential for more accurate cancer prognosis.

To address this problem, our proposed model, cross-attention transformer-based multimodal fusion network (CATfusion), for pan-cancer survival prediction by integrating multimodal histology-genomic data, embodies a triple advantage. Firstly, the use of a self-supervised learning framework, TabAE, for feature extraction from genomic data, addresses the challenge of maintaining data integrity while reducing dimensionality. Secondly, the integration of the greatest variety of data types, including mRNA-seq, miRNA-seq, copy number variation, DNA methylation variation, mutation data, and histopathological slides, allows for a comprehensive analysis that captures the multifaceted nature of cancer. Thirdly, CATfusion's architecture, which includes a bidirectional multimodal attention mechanism and self-attention block, is adept at synchronizing the learning and integration of representations from various modalities. The model's enhanced outcomes, characterized by elevated C-index and survival area under the curve (AUC) scores surpassing those of single-modal and conventional models, highlight its promise for practical clinical application. Its proficiency in accurately categorizing patients into low- and high-risk cohorts is a boon for individualized medicine, enabling the customization of therapeutic strategies in accordance with each patient's unique risk profile.

Materials and methods

Dataset description

Genomic (copy number variation, mutation), epigenomic (DNA methylation variation), and transcriptomic (mRNA-seq, miRNA-seq) data for 32 cancer types in The Cancer Genome Atlas (TCGA) [29] were downloaded from the Firehose of the Broad Institute (http://gdac.broadinstitute.org/, January 2023 version). Diagnostic whole slide images (WSIs) and their corresponding clinical data were also obtained from TCGA and are publicly accessible through the NIH Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/). Detailed statistics on the number of samples for each modality can be found in Supplementary Table 1. Due to the limited sample size in miRNA-seq data, TCGA-GBM was excluded from the multimodal fusion analysis.

In this study, the preprocessing of pathological images adheres to the methodology outlined in the Self-supervised Image Search for Histology (SISH) publication [30], aiming to dissect extensive WSIs into manageable segments, termed "mosaics." The detailed description of the preprocessing of genomics data and pathological images are delineated in Supplementary File 1.

Representation learning for genomic features

In this study, we introduce TabAE, a novel autoencoder-decoder architecture that capitalizes on the inherent information redundancy within structured datasets to facilitate feature extraction. Illustrated in Fig. 1b, we exemplify the process using gene expression profile data from a single sample, initially characterized by a [1, 10 240] dimensionality. By employing a 1024-length sliding window, the 1D expression vector is reshaped into a 2D format [10, 1024]. Subsequently, a random masking strategy with a 40% probability is implemented, selectively omitting certain column vectors, as depicted by the shaded regions post "Random Mask" application. The remaining vectors are subsequently subjected to an encoding process via an encoder architecture. This encoder is constructed from four layers of transformer blocks, which integrate self-attention mechanisms to extract features from the vectors. Following this, the feature vectors that were initially masked with random values are reinserted at their respective positions, reconstructing a feature vector identical in size to the original input. The feature vector then undergoes a decoding phase through a decoder, which is assembled from two layers of transformer blocks with self-attention mechanisms, facilitating the reconstruction of the vector. The objective loss function is the mean squared error between the output and the input vectors at the masked positions, which is defined as follows:

$$Loss = 1/n \sum_{i=1}^{n} (x_i - y_i)^2 \quad (1)$$

Here, y and x respectively denote the output vector and the input vector at the corresponding masked positions.

Ultimately, the genomic data of all samples are passed through the encoder of their respective trained TabAE to obtain the feature vectors.

In order to validate the capacity of feature vectors from diverse cancer samples to delineate the intrinsic features of each cancer type, we employ a classic random forest classification model. This model is trained on feature vectors derived from one type of omics data. Following the training phase, the model's classification accuracy is meticulously tested on samples from each cancer type to ensure its ability to accurately distinguish between them. In addition, we evaluated the impact of different gene embedding methods on the classification performance of feature maps obtained using RNA-seq TabAE. In addition, we evaluated the impact of different gene embedding methods [31–33] on the classification performance of feature maps obtained using RNA-seq TabAE (Supplementary Table 8). Overall, the performance of the three methods was largely comparable. However, in certain challenging cancer types, such as TCGA-COAD, our gene embedding method demonstrated a slight advantage.

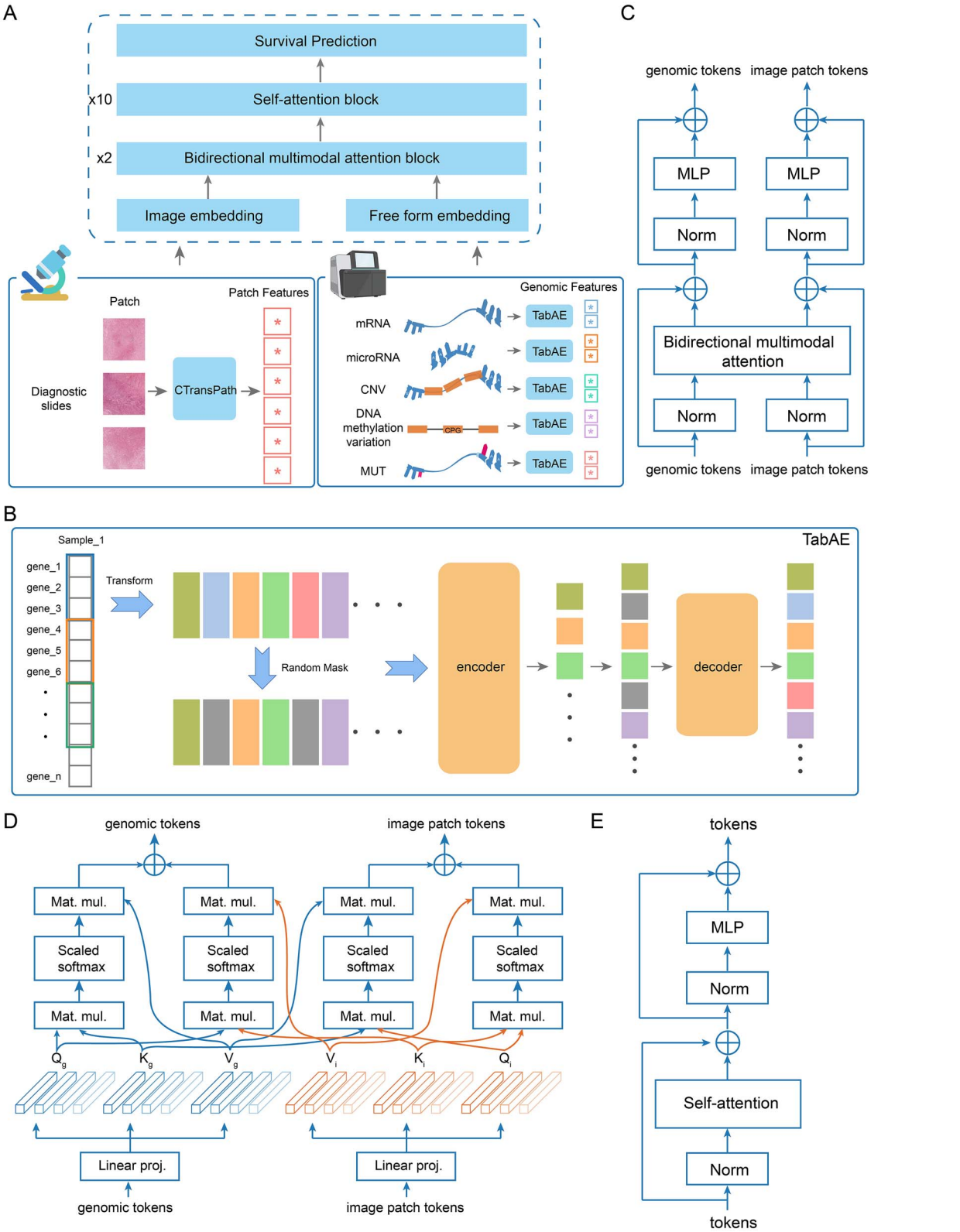

Figure 1. Pan-cancer integrative histology-genomic fusion analysis platform for survival estimation (CATfusion) workflow. (a) Illustration of the proposed CATfusion framework. (b) The process of extracting features from genomic data. Illustration of the proposed TabAE framework. (c) The intricate architecture of a bidirectional multimodal attention block, which encompasses two-layer normalization layers (norm), one bidirectional multimodal attention layer and one MLP. (d) In the bidirectional multimodal attention layer, meticulous attention mechanisms are at play, synchronizing the learning and integration of representations drawn from multiple modalities. (e) Detailed architecture of a self-attention block.

Representation learning for histologic image

In parallel with the feature characterization of genomic data, the fragmented mosaics derived from pathological images necessitate feature extraction to capture their underlying pathological signatures. For this purpose, various state-of-the-art feature extraction tools, renowned for their efficacy in the field, have been utilized to delineate the features inherent in these mosaics. Illustrated in Supplementary Fig. 5A, in an effort to identify the most informative features, this study has employed seven distinct feature extractors for pathological images: Prov-GigaPath [34], Hibou [35], Kaiko [36], Phikon v2 [37], BiomedCLIP [38], PLIP [39], and CTransPath [40]. BiomedCLIP and PLIP are deep learning architectures that have been pretrained on datasets comprising image–text pairs, capitalizing on the synergistic relationship between visual and textual data. Conversely, Prov-GigaPath, Hibou, Kaiko, Phikon v2, and CTransPath employs a self-supervised learning approach to enhance the model's ability to discern relevant features.

To verify whether histopathological slides and omics data are correlated and whether they can complement each other, we conducted cross-modal analyses using RNA-seq and methods like RNAPath [41] and HE2RNA [42] to infer gene expression profiles from histopathological slides. The inferred profiles showed a moderate correlation (around 0.45) with actual RNA-seq data across cancer types (Supplementary Fig. 5E). Similar correlations were observed with methylation and miRNA expression.

Multimodalities fusion (CATfusion)

In this research, we propose a novel, multimodal fusion approach leveraging cross-attention mechanisms (CATfusion), designed to prognosticate the survival risks of oncology patients by amalgamating data from pathological slides and diverse genomic modalities. As illustrated in Fig. 1a, CATfusion ingests features extracted from pathological images and a spectrum of genomic datasets, with respective dimensions tailored for each data type. To harmonize the data dimensions, an embedding layer is employed to equate the feature space's second dimension to 768 units. The pathological imagery undergoes a linear transformation, is prefixed with a classification-specific "cls" token, and is endowed with positional embeddings to delineate the provenance of features. A dropout layer is integrated to mitigate the risk of overfitting. Parallel processing is applied to genomic datasets through analogous embedding layers to adjust their semantic representation. Post-embedding, the integrated features from pathology and genomics are directed into a bidirectional attention fusion module, as outlined in Supplementary Fig. 1A, where self-attention mechanisms are adeptly transitioned into cross-attention to foster intermodality data exchange, as visualized in Fig. 1c. The ensuing feature amalgamation is then subjected to a series of 10 self-attention block transformer modules (Fig. 1e), enhancing the model's capacity to discern and assimilate intrinsic data characteristics. The culmination of this process is the conveyance of the enriched features to a dedicated survival prediction layer, which performs a survival analysis to ascertain the survival prospects of cancer patients.

In the bidirectional multimodal attention fusion module, the structure of the cross-attention fusion operation is shown in Fig. 1d. Given the multiomics feature matrix $\mathbf{G}$ and the pathological image feature matrix $\mathbf{I}$, they are first transformed through converters to obtain their respective query matrices (Q_g, Q_i), key matrices (K_g, K_i), and value matrices (V_g, V_i). Then, the matrices $\mathbf{K}$ and $\mathbf{V}$ are exchanged for each modality and input into the corresponding multihead attention, calculating the image attention matrix under multiomics conditions and the multiomics attention matrix under image conditions. The input for the multiomics feature matrix G^{t+1} and the pathological image feature matrix I^{t+1} at the $\mathbf{t}+1$ layer can be calculated using the following formula:

$$G^{t+1} = softmax\left(\frac{Q_g^{t\,T} K_g^t}{\sqrt{d_{k_g}}}\right) V_g^t + softmax\left(\frac{Q_g^{t\,T} K_i^t}{\sqrt{d_{k_i}}}\right) V_i^t \tag{2}$$

$$I^{t+1} = softmax\left(\frac{Q_i^{t\,T} K_i^t}{\sqrt{d_{k_i}}}\right) V_i^t + softmax\left(\frac{Q_i^{t\,T} K_g^t}{\sqrt{d_{k_g}}}\right) V_g^t \tag{3}$$

Here, d_{k_g}, d_{k_i} represent the dimensions of the key matrices, respectively.

Survival prediction

The output of the network was a single node, which estimates the risk function. The weights of the network are trained with the time-to-event (death) outcomes to optimize the Cox likelihood function. The Cox partial likelihood was defined as follows:

$$\mathrm{L}(\theta) = \prod_{i=1}^{n}\left[\frac{\exp\left(\mathrm{h}_\theta\left(\mathbf{x}_i\right)\right)}{\sum_{j\in \mathbf{f}(\mathbf{x}_i)} \exp\left(\mathrm{h}_\theta\left(\mathbf{x}_j\right)\right)}\right]^{\delta_i} \tag{4}$$

When we minimized negative log partial likelihood, the loss function was

$$\mathrm{I}(\theta) = -\sum_{i=1}^{n} \delta_i \left(\mathrm{h}_\theta\left(\mathbf{x}_i\right) - \log \sum_{j\in \mathbf{f}(\mathbf{x}_i)} \exp\left(\mathrm{h}_\theta\left(\mathbf{x}_j\right)\right)\right) \tag{5}$$

δ_i was an indicator of whether the survival time was censored ($\delta_i = 0$) or observed ($\delta_i = 1$). θ was the weight of the network. $\mathbf{f}(\mathbf{x}_i)$ denoted the set of individuals who were at risk for failure time of individual i. Training details are delineated in Supplementary File 1.

The process of achieving model interpretability

In our study, we have adopted the model interpretation procedures as detailed in the PORPOISE [43]. In brief, for a given WSI, to perform visual interpretation of the relative importance of different tissue regions toward the patient-level prognostic prediction, we first compute attention scores for 100 mosaics (without overlap) from all tissue regions in the slide. We refer to the attention score distribution across all mosaics from all WSIs from the patient case as the reference distribution. For fine-grained attention heatmaps, attention scores for each WSI are recomputed by increasing the tiling overlap to up to 90%. For visualization, the attention scores are converted to percentile scores between 0.0 (low attention) to 1.0 (high attention) using the initial reference distribution and spatially registered onto the corresponding WSIs (scores from overlapping mosaics are averaged). The resulting heatmap, referred to as local WSI interpretability, is transformed to RGB values using a colormap and overlayed onto the original slide image with a transparency value of 0.5.

Achieving corresponding cell labels in high attention regions of slide image

For sets of WSIs belonging to different patient cohorts, we performed global WSI interpretability by quantifying and characterizing the morphological patterns in the highest-attended image

patches from each WSI. Since WSIs have differing image dimensions, we extracted a proportional amount of high attention image patches (1%) to the total image dimension. On average, approximately 135 512 × 512 image patches used as high attention regions in each slide. These attention patches are analyzed using a HoverNet model pretrained for simultaneous cell instance segmentation and classification [44]. Cells are classified as either tumor cells (red), inflammatory cells (green), stromal cells (blue), necrosis cells (yellow), or non-neoplastic epithelial cells (orange). For each of these cell types, we analyzed the cell type frequency across all counted cells in the highest-attended image patches in a given patient, then analyzed the cell fraction distribution across all patients in low- and high-risk patients, defined as patients below and above the 25% and 75% predicted risk percentiles, respectively.

Results

Effectiveness of TabAE architecture for representation learning in genomic data

To ensure an unbiased extraction of intrinsic sample characteristics from each type of genomic data, we deployed a self-supervised learning framework, yielding the construction of the TabAE model as illustrated in Fig. 1b. This model was trained rigorously across a spectrum of TCGA genomic datasets encompassing mRNA-seq, miRNA-seq, copy number variation, DNA methylation variation, and mutation data, as detailed in Supplementary Fig. 1B. In our training regimen, the TCGA-TGCT dataset served as a validation cohort, while the remaining samples were apportioned into training and test subsets with a ratio of 4:1, culminating in a five-fold cross-validation process. The trajectory of loss reduction throughout training is delineated in Supplementary Fig. 2. We observed a tendency for overfitting during the TabAE model training when utilizing miRNA-seq data, attributable to its relatively few input features. In contrast, other omics datasets did not exhibit significant overfitting phenomena. Post the training of the TabAE models for each genomic category, we extracted the feature maps from the encoding layer of the TabAE, which were subsequently utilized as the genomic features for CATfusion input.

To determine whether these feature maps could capture the unique characteristics of various cancer types, we used the Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction algorithm to visualize the sample distribution across different cancer types in a 2D plane, as shown in Fig. 2a and Supplementary Fig. 3A. The features derived from mRNA-seq, miRNA-seq, and DNA methylation variation effectively distinguished samples across different cancer types. Additionally, the validation dataset, TCGA-TGCT, clearly segregated from other cancer types. Conversely, the mutation data exhibited the least effectiveness in discrimination (Supplementary Fig. 3A), corroborating existing research that posits a higher degree of stochasticity in mutation information. Parallel findings were observed in experiments utilizing a random forest classification model (see Methods, Fig. 2b, Supplementary Fig. 3B and C, and Supplementary Fig. 4A and B).

Choose the optimal whole-slide image feature extractor

In parallel with the genomic data feature analysis, the extraction of features from the fragmented pathology images, simply referred to as mosaics, is imperative for revealing the intrinsic pathological signatures. For this purpose, we employed seven published and well-trained self-supervised learning algorithms—Prov-GigaPath, Hibou, Kaiko, Phikon v2, BiomedCLIP, PLIP, and CTransPath—to serve as feature extractors for these pathological images. To evaluate how well the mosaic features capture pathological heterogeneity across various cancer types, we used an attention-based multiple instance learning (AttMIL) classification model. The model was trained on the extracted mosaic features to classify cancer samples into their specific cancer types.

The AttMIL model, when trained with features derived from CTransPath, demonstrated superior performance, attaining an accuracy of 84% on the test dataset with the minimal incidence of overfitting, as depicted in Supplementary Fig. 5B–D and Supplementary Table 2. This outcome underscores the robustness of the CTransPath-extracted features in discerning the pathological nuances of different cancer types (Supplementary Fig. 6). Based on these findings, we selected the image mosaics features extracted by CTransPath to serve as the image component input for the CATfusion model. This integration of features enables a holistic analysis that encapsulates the multifaceted characteristics of the cancer phenotypes.

Multimodalities fusion (CATfusion) for survival prediction

To tackle the complexities of creating integrated image-omics biomarkers for predicting cancer outcomes, we propose a deep learning-based multimodal fusion (CATfusion) algorithm that uses both histopathological slides and molecular profile features (mRNA-seq, miRNA-seq, copy number variation, DNA methylation variation, mutation) to assess and elucidate the comparative risk associated with cancer mortality (Fig. 1a). In an initial evaluation, we subjected our CATfusion model to a five-fold cross-validation utilizing paired whole-slide imaging and genomic data across 31 cancer types. We also performed a comparative analysis with single-modality deep learning models: one employing an attention-based multiple-instance learning (AttMIL-WSI) framework [45] on whole-slide images mosaic features exclusively, and another relying on genomic data features solely (AttMIL-Genomic). In evaluating the comparative effectiveness of these models, we leveraged the cross-validated concordance index [46] (C-index) to quantify predictive performance, employed Kaplan–Meier curves [47] for visual assessment of patient risk stratification, and utilized the log-rank test for statistical validation of the stratification's ability to differentiate between low- and high-risk patient groups, specifically at the 50th percentile threshold of the predictive risk scores. Furthermore, alongside the C-index, we present the dynamic AUC, known as the survival AUC [48]. This metric mitigates the overestimation inherent in model performance calculations due to censored data, offering a more accurate reflection of predictive capabilities.

Encompassing a spectrum of 31 cancer types, CATfusion achieved an overall C-index of 0.668, outperforming AttMIL-WSI and AttMIL-Genomic, which garnered C-index of 0.642 and 0.650, respectively (Fig. 3b; Supplementary Table 3). In terms of survival AUC, CATfusion exhibited a parallel enhancement with an overall score of 0.628, surpassing the 0.612 and 0.601 achieved by AttMIL-WSI and AttMIL-Genomic (Fig. 3c; Supplementary Table 3). When evaluating the models' performance on individual cancer types, CATfusion dominated, attaining the top C-index for 23 out of 31 types, and for 26 types, it showed statistically significant effectiveness in distinguishing between low- and high-risk patient groups (Fig. 3 a and b; Supplementary Fig. 7; Supplementary Table 3). In contrast to AttMIL-Genomic, which

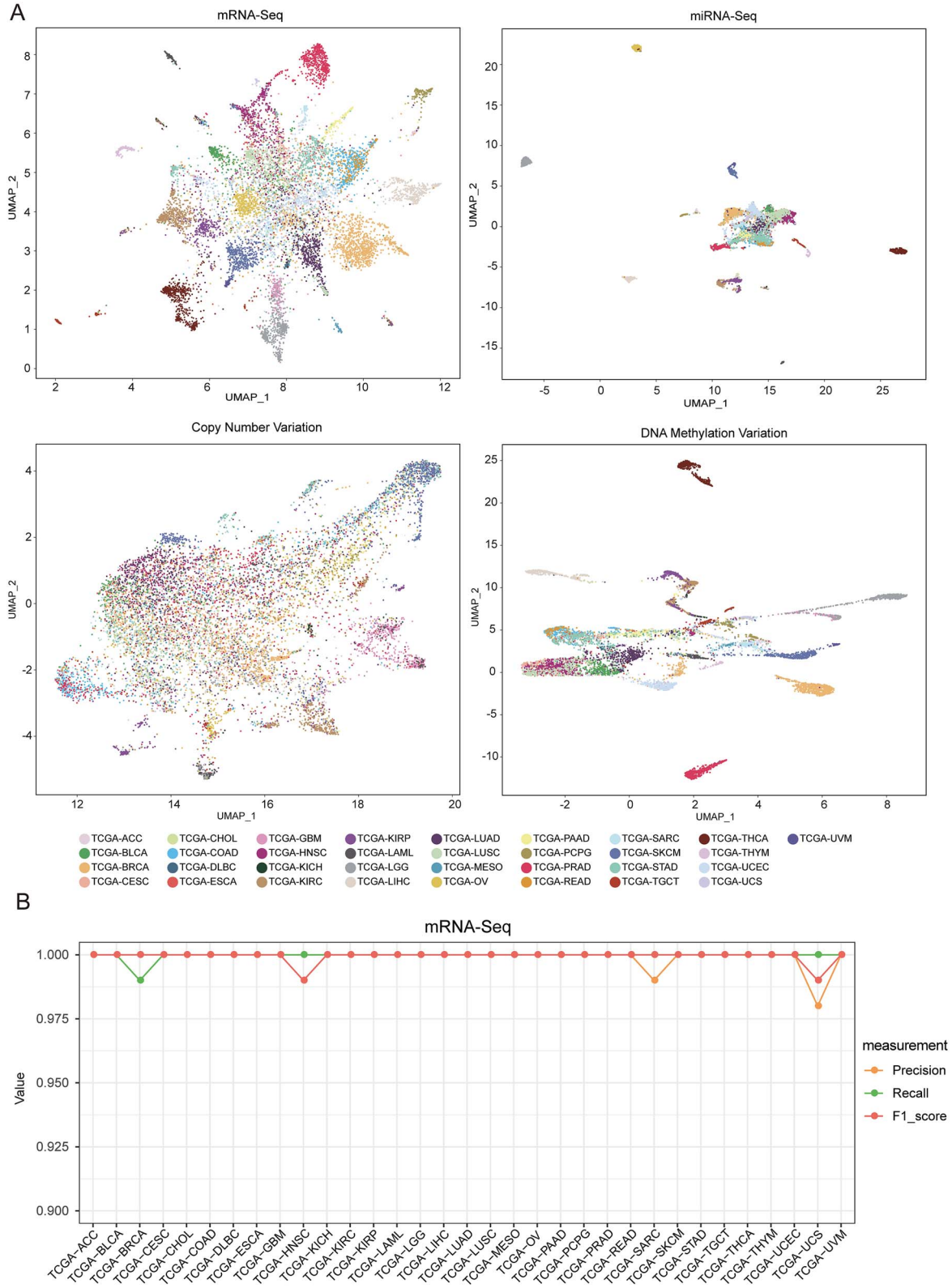

Figure 2. Evaluate the features extracted from different genomic data through unsupervised methods. (a) The scatter plot delineates the outcomes of dimensionality reduction via UMAP, showcasing the features derived from a spectrum of genomic datasets, including mRNA-seq, miRNA-seq, copy number variation, and DNA methylation variation. (b) The classic random forest classification model is utilized to fit the features extracted from the mRNA-seq data by TabAE. A line chart illustrates the performance of the trained random forest classification model across various cancer type datasets.

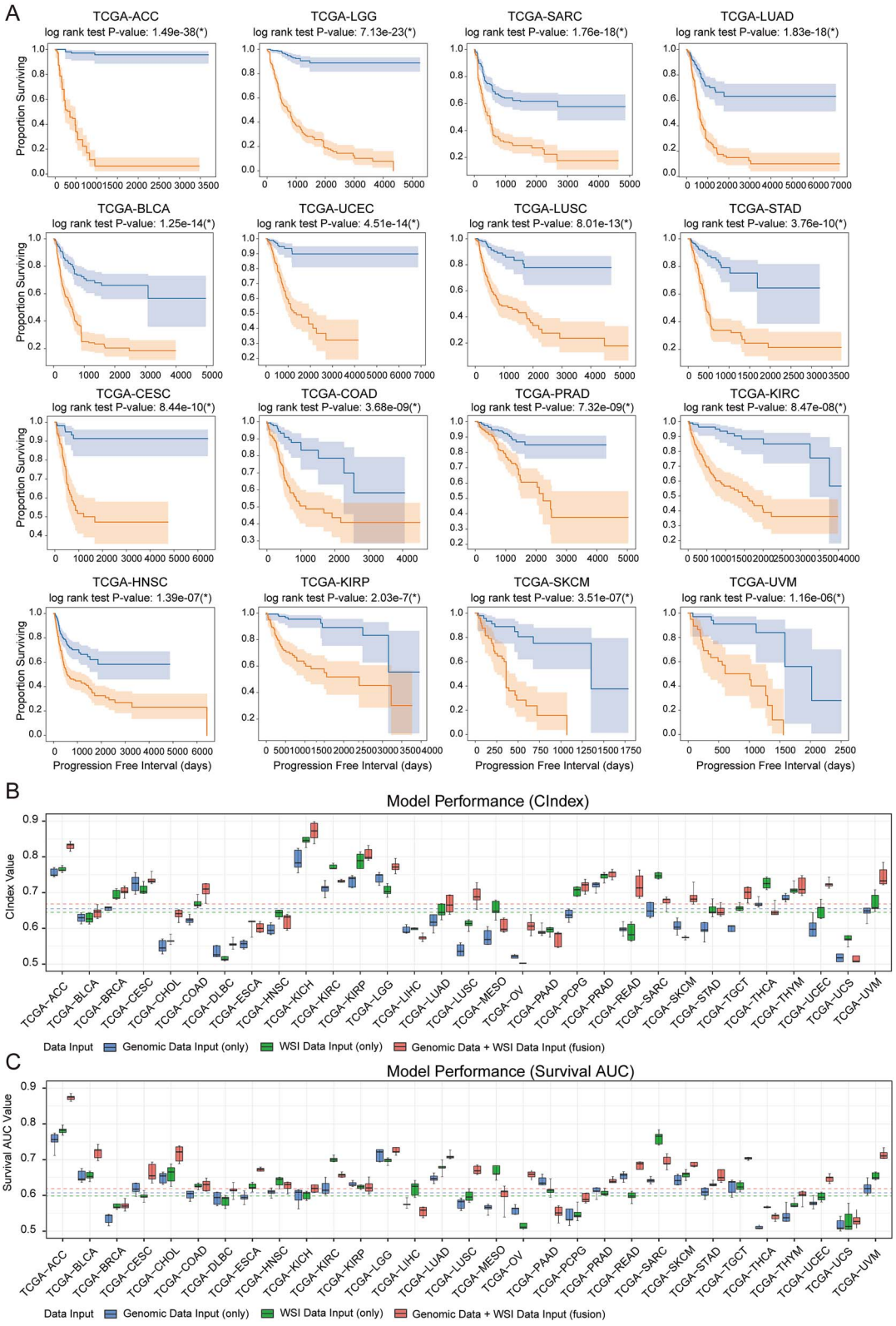

Figure 3. Model performances of CATfusion for survival prediction. (a) Kaplan–Meier analysis of patient stratification of low- and high-risk patients via CATfusion across 16 cancer types. Low and high risks are defined by the median 50% percentile of hazard predictions via CATfusion. Log rank test was used to test for statistical significance in survival distributions between low- and high-risk patients ($*P < .05$). (b) The C-index performance of survival models with different model inputs (genomic data, WSI data, genomic data and WSI data) is evaluated for each cancer type using a five-fold cross-validation method. Horizontal line for each model shows average C-index performance across all cancer types. Boxplots correspond to C-indices of five replicates on the aggregated risk predictions. (c) The survival AUC performance of survival models with different model inputs (genomic data, WSI data, genomic data and WSI data) is evaluated for each cancer type using a five-fold cross-validation method. Horizontal line for each model shows average survival AUC performance across all cancer types. Boxplots correspond to C-indices of five replicates on the aggregated risk predictions.

Table 1. Comparison of multimodal fusion trained with various fusion operators.

	C-index			
Datasets	**Concat**	**Kronecker dot**	**Low-rank**	**CATfusion**
TCGA-ACC	0.607 ± 0.036	0.671 ± 0.016	0.624 ± 0.021	**0.83 ± 0.022**
TCGA-BLCA	0.624 ± 0.027	0.631 ± 0.017	0.618 ± 0.026	**0.647 ± 0.028**
TCGA-BRCA	0.576 ± 0.034	0.607 ± 0.029	0.651 ± 0.022	**0.724 ± 0.027**
TCGA-CESC	0.611 ± 0.031	0.681 ± 0.034	0.663 ± 0.030	**0.742 ± 0.026**
TCGA-CHOL	0.607 ± 0.026	0.626 ± 0.038	**0.641 ± 0.013**	0.631 ± 0.027
TCGA-COAD	0.612 ± 0.043	0.644 ± 0.027	0.712 ± 0.029	**0.722 ± 0.021**
TCGA-DLBC	0.536 ± 0.041	0.551 ± 0.031	**0.589 ± 0.030**	0.565 ± 0.029
TCGA-ESCA	0.521 ± 0.037	0.572 ± 0.023	0.581 ± 0.036	**0.605 ± 0.014**
TCGA-HNSC	0.563 ± 0.041	0.582 ± 0.030	**0.640 ± 0.025**	0.62 ± 0.028
TCGA-KICH	0.627 ± 0.016	0.716 ± 0.026	0.683 ± 0.016	**0.87 ± 0.034**
TCGA-KIRC	0.618 ± 0.047	0.634 ± 0.042	0.676 ± 0.026	**0.725 ± 0.007**
TCGA-KIRP	0.794 ± 0.026	0.749 ± 0.034	0.751 ± 0.030	**0.804 ± 0.014**
TCGA-LGG	0.762 ± 0.037	0.711 ± 0.016	0.754 ± 0.029	**0.849 ± 0.036**
TCGA-LIHC	**0.618 ± 0.041**	0.567 ± 0.029	0.553 ± 0.017	0.577 ± 0.025
TCGA-LUAD	0.611 ± 0.027	0.621 ± 0.026	0.631 ± 0.025	**0.667 ± 0.017**
TCGA-LUSC	0.506 ± 0.043	0.624 ± 0.029	**0.653 ± 0.010**	0.651 ± 0.021
TCGA-MESO	0.537 ± 0.044	**0.617 ± 0.017**	0.573 ± 0.026	0.609 ± 0.024
TCGA-OV	0.561 ± 0.037	**0.591 ± 0.024**	0.537 ± 0.015	0.552 ± 0.018
TCGA-PAAD	**0.631 ± 0.034**	0.604 ± 0.011	0.583 ± 0.024	0.569 ± 0.023
TCGA-PCPG	0.657 ± 0.026	0.649 ± 0.020	0.651 ± 0.027	**0.716 ± 0.029**
TCGA-PRAD	0.573 ± 0.030	0.583 ± 0.016	0.657 ± 0.031	**0.748 ± 0.018**
TCGA-READ	0.624 ± 0.037	0.663 ± 0.029	0.673 ± 0.026	**0.724 ± 0.034**
TCGA-SARC	0.627 ± 0.023	**0.676 ± 0.016**	0.637 ± 0.012	0.67 ± 0.023
TCGA-SKCM	0.631 ± 0.046	0.651 ± 0.042	0.660 ± 0.027	**0.692 ± 0.039**
TCGA-STAD	0.546 ± 0.019	**0.673 ± 0.018**	0.606 ± 0.024	0.655 ± 0.014
TCGA-TGCT	0.618 ± 0.023	0.637 ± 0.024	0.628 ± 0.015	**0.694 ± 0.035**
TCGA-THCA	**0.610 ± 0.036**	0.549 ± 0.010	0.566 ± 0.025	0.563 ± 0.025
TCGA-THYM	0.609 ± 0.017	0.678 ± 0.016	**0.732 ± 0.027**	0.716 ± 0.033
TCGA-UCEC	0.652 ± 0.026	0.682 ± 0.026	0.649 ± 0.026	**0.729 ± 0.023**
TCGA-UCS	0.517 ± 0.011	0.537 ± 0.021	**0.563 ± 0.030**	0.528 ± 0.039
TCGA-UVM	0.651 ± 0.040	0.648 ± 0.027	0.683 ± 0.020	**0.748 ± 0.022**

Note: The bolded numbers indicate the maximum values under the corresponding indicators. The number preceding the plus-minus sign (±) represents the mean value, and the number following it represents the standard deviation.

solely relies on genomic data, CATfusion maintained a steady performance in both C-index and survival AUC across 28 types of cancer. Though AttMIL-Genomic had a comparable performance on some cancer types, we observed both substantial improvement in fusion model performance and patient stratification for breast invasive carcinoma (TCGA-BRCA), colon adenocarcinoma (TCGA-COAD), kidney renal papillary cell carcinoma (TCGA-KIRP), and uterine corpus endometrial carcinoma (TCGA-UCEC). Compared with AttMIL-WSI, CATfusion also demonstrated consistent performance in C-index across 23 cancer types. We noticed both substantial improvement in model performance for rectum adenocarcinoma (TCGA-READ) and brain lower grade glioma (TCGA-LGG).

Overall, however, model performances were found to improve following multimodal integration for almost all cancer types (Fig. 2b), the comparative analysis of different aggregation strategies on the TCGA datasets confirmed this result (Supplementary Table 4). In examining unimodal models that were close to CATfusion performance, AttMIL-Genomic showed significance in stratifying pancreatic adenocarcinoma (TCGA-PAAD), and AttMIL-WSI showed significance in stratifying sarcoma (TCGA-SARC) and mesothelioma (TCGA-MESO).

Among all single cancer types included in our study, TCGA-LGG had the largest performance increase with multimodal training, reaching a C-index performance of 0.849 [95% confidence interval (CI): 0.813–0.885, $P = 7.126 \times 10^{-47}$, log rank test], compared with 0.729 (95% CI: 0.713–0.745, $P = 2.45 \times 10^{-26}$, log rank test) using AttMIL-Genomic and 0.708 (95% CI: 0.696–0.72, $P = 1.40 \times 10^{-24}$, log rank test) using AttMIL-WSI (Supplementary Table 3). Following the correction of potential optimistic bias with high censorship via survival AUC evaluation, we observed similar model performances with CATfusion reaching an AUC of 0.727(SD: 0.054) compared with 0.709 (SD: 0.023) in AttMIL-Genomic and 0.694 (SD: 0.006) in AttMIL-WSI.

To focus on more in-depth analysis of the results for lung cancer, we expanded our data collection to include relevant datasets from the CPTAC database (Supplementary Table 5), retrained the CATfusion model on lung cancer datasets that are present in both the CPTAC and TCGA cohorts (TCGA-LUAD, TCGA-LUSC, CPTAC-LUAD, CPTAC-LUSC) and subsequently tested the model. The results are summarized in Supplementary Table 6. Overall, CATfusion achieved a high C-index value, indicating strong predictive performance.

Evaluation of CATfusion

In addition to conducting ablation studies in comparing unimodal and multimodal models, we also assessed Cox proportional hazard models using age, gender, and tumor grade covariates as baselines, which were still outperformed by CATfusion (Supplementary Table 7). We also performed comprehensive comparative analysis of multimodal fusion techniques, employing diverse fusion operators (Table 1). The study juxtaposes four distinct fusion strategies: the conventional concatenation approach, the Kronecker dot product approach, low-rank

Table 2. A comparison of survival prediction performance is made between other multimodal models and our model on the TCGA datasets for lower grade glioma (TCGA-LGG), breast cancer (TCGA-BRCA), and lung squamous cell carcinoma (TCGA-LUSC).

Methods		C-index		
		TCGA-LGG	TCGA-BRCA	TCGA-LUSC
Traditional	RSF	0.790 ± 0.022	0.626 ± 0.040	0.542 ± 0.029
	Lasso–Cox	0.785 ± 0.014	0.611 ± 0.055	0.567 ± 0.027
Deep learning	PORPOISE	0.821 ± 0.023	0.714 ± 0.014	**0.659 ± 0.020**
	Pathomic Fusion	0.817 ± 0.027	0.717 ± 0.013	0.618 ± 0.025
	CAMR	0.803 ± 0.020	0.722 ± 0.021	0.627 ± 0.011
	GPDBN	0.818 ± 0.028	0.706 ± 0.032	0.630 ± 0.023
	CATfusion (our)	**0.849 ± 0.036**	**0.724 ± 0.027**	0.651 ± 0.021

Note: The bolded numbers indicate the maximum values under the corresponding indicators. The number preceding the plus-minus sign (±) represents the mean value, and the number following it represents the standard deviation.

multimodal fusion, and the CATfusion. Results reveals that the CATfusion demonstrates a consistent and statistically significant enhancement in predictive accuracy across the majority of the evaluated cancer types (18/31), as evidenced by the highest C-index values. Notably, the CATfusion achieves a remarkable C-index of 0.87 ± 0.034 for the TCGA-KIRC dataset, underscoring its superiority in integrating multimodal data for robust cancer outcome prediction.

We have conducted a rigorous comparative analysis of survival prediction performance, juxtaposing our proposed CATfusion model against a spectrum of traditional and deep learning methodologies on TCGA datasets for lower-grade glioma (TCGA-LGG), breast cancer (TCGA-BRCA), and lung squamous cell carcinoma (TCGA-LUSC). The traditional approaches, including RSF [49], and Lasso–Cox [50], were benchmarked alongside deep learning algorithms such as PORPOISE [43], Pathomic Fusion [26], CAMR [27], and GPDBN [28]. To ensure equitable assessment, all comparative models were evaluated using identical input features, as delineated in Table 2. The comparative analysis reveals that while traditional models achieve moderate success, the deep learning methods generally surpass them in predictive accuracy. Notably, Pathomic Fusion demonstrates a commendable C-index value on the TCGA-LGG dataset. However, our CATfusion model emerges as the preeminent performer, securing a C-index of 0.849 for TCGA-LGG, 0.724 for TCGA-BRCA, and 0.651 for TCGA-LUSC. These results not only eclipse the performance of the second-best methods by a significant margin—2.0%, 4.3%, and 2.0%, respectively—but also underscore the superiority of CATfusion in amalgamating multimodal data to enhance the prognostication of patient outcomes.

Model interpretability and visualization

For interpretation and further validation of our models, we applied attention- and gradient-based interpretability to our trained CATfusion model in order to explain how WSIs features are respectively used to predict prognosis. For each slide, we used a custom visualization tool that overlays attention weights computed from CATfusion onto the diagnostic slide, which is displayed as a high-resolution attention heatmap that shows relative prognostic relevance of image regions used to predict risk.

We assessed high-attention regions of WSIs in the top 25% (high-risk group) and bottom 25% (low-risk group) of predicted patient risks for each cancer type, which reflect favorable and poor cancer prognosis, respectively (Fig. 4a; Supplementary Fig. 8A). In addition to visual inspection from two pathologists, we simultaneously segmented and classified cell-type identities across high-attention regions in our WSIs (Fig. 4b; Supplementary Fig. 8B). Taking TCGA-LUAD (lung adenocarcinoma) and TCGA-SKCM (skin cutaneous melanoma) as examples, we generally observed that high-attention regions in low-risk patients corresponded with greater immune cell presence and lower tumor grade than that of high-risk patients, demonstrating statistically significant differences in lymphocyte cell fractions in high-attention regions (Fig. 4c; Supplementary Fig. 8C). In the selected example of lung adenocarcinoma (Fig. 4), the original histological section does not allow for the discernment of regions with the naked eye. However, our model is able to highlight high-attention areas near the alveoli (indicated by circles), which correspond to the dilation of capillaries and infiltration of inflammatory cells in cancerous tissues. These subtle pathological features are often not easily detected by pathologists through visual inspection alone. Furthermore, we also observed that high-attention regions in high-risk patients corresponded with increased tumor cell presence and tumor invasion in TCGA-LUAD and TCGA-SKCM, demonstrating statistically significance differences in tumor cell fractions.

Discussion

We introduce a novel deep learning-driven approach, CATfusion, for pan-cancer survival prediction by integrating multimodal histology-genomic data. Our methodology introduces several innovative elements. Firstly, the use of a self-supervised learning framework, TabAE, for feature extraction from genomic data, addresses the challenge of maintaining data integrity while reducing dimensionality. Secondly, the integration of the greatest variety of data types, including WSI, mRNA-seq, miRNA-seq, copy number variation, DNA methylation variation, and mutation data, allows for a comprehensive analysis that captures the multifaceted nature of cancer. Thirdly, the application of cross-attention mechanisms in CATfusion enables effective fusion of heterogeneous data, leading to enhanced predictive capabilities. Returning to the research motivation, human solid tumors, whether in normal tissue or in the context of tumorigenesis, are not driven by a single cell type but rather by complex multicellular communities. These communities form the smallest functional units within tumors, and differences in their composition are responsible for individual variability and prognostic outcomes. Our approach is motivated by the need to understand the multicellular nature of solid tumors and the availability of a large and diverse dataset like TCGA, which enables us to uncover clinically relevant patterns across multiple cancer types.

Despite the promising results, our study has potential limitations. The generalizability of our model may be constrained by the

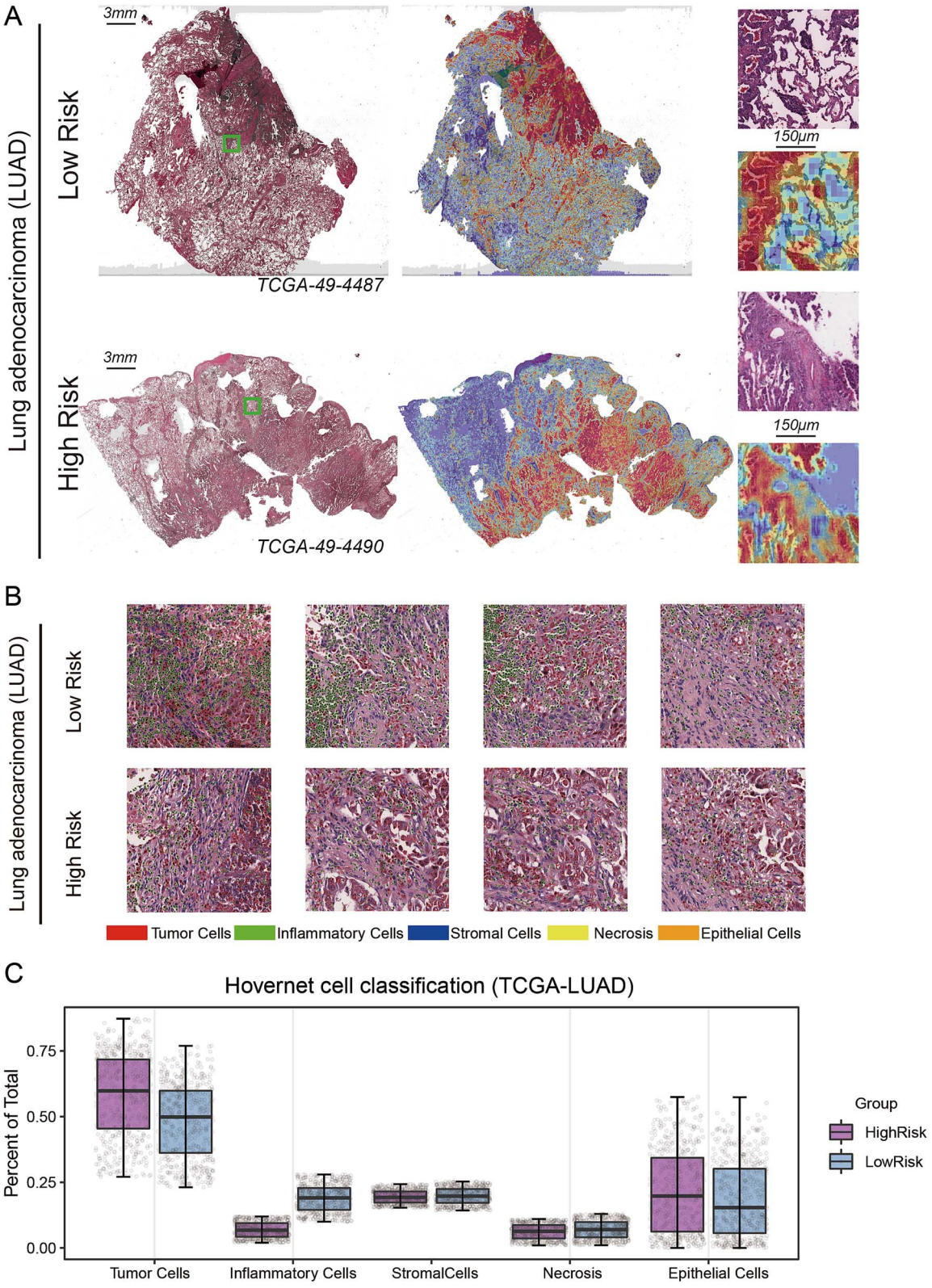

Figure 4. Model explanation and global interpretability analyses of CATfusion on LUAD. (a) WSIs, associated attention heatmaps, regions of interest (ROIs), ROI heatmaps, and selected high attention patches from example low-risk (top) and high-risk (bottom) cases. In LUAD, high attention for low-risk cases tends to focus on regions with dense inflammatory infiltrate, predominantly composed of lymphocytes, while in high-risk cases, high attention focuses on regions of central necrosis within tumor nests. (b) Exemplar high-attention patches from low-risk (top) and high-risk (bottom) cases with corresponding cell labels. (c) Quantification of cell types in high-attention patches for TCGA-LUAD, with statistical significance for increased inflammatory cells in low-risk patients. Boxes indicate quartile values and whiskers extend to data points within 1.5× the interquartile range.

composition of the TCGA dataset, which may not fully represent the diversity of cancer patient populations worldwide. Additionally, the reliance on self-supervised learning for feature extraction may also introduce biases based on the training data, which could affect the model's performance. Looking forward, several avenues for future research open up. Expanding the dataset to include more diverse patient populations and integrating clinical variables could enhance the model's applicability.

In conclusion, our study presents CATfusion, a robust model for pan-cancer survival prediction. It integrates multimodal data and deep learning to improve accuracy and interpretability, showing potential for clinical application. Future work should address limitations and build on strengths.

Key Points

- Our CATfusion employs a self-supervised learning framework, TabAE, for feature extraction from genomic data, addresses the challenge of maintaining data integrity while reducing dimensionality.
- The integration of the greatest variety of data types, including mRNA-seq, miRNA-seq, copy number variation, DNA methylation variation, mutation data, and histopathological slides, allows for a comprehensive analysis that captures the multifaceted nature of cancer.
- Our CATfusion's architecture, which includes a bidirectional multimodal attention mechanism and self-attention block, is adept at synchronizing the learning and integration of representations from various modalities.

Author contributions

D.W. and Y.H. conceived the study. D.W., Y.H., and G.W. designed and supervised the study. D.W. and Y.F.H. performed the experiments. Y.F.H. performed bioinformatic analyses. D.W. and Y.F.H. wrote the manuscript with input from all of the authors. D.W., G.W., Y.Y., J.J.K., and Y.F.H. revised the manuscript.

Supplementary Data

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

Conflict of interest: None declared.

Funding

This work was supported by grants from the National Key Research and Development Project of China (2022YFA0806303), National Natural Science Foundation of China (82370106), Guangdong Basic and Applied Basic Research Foundation (2024A1515011769), Heilongjiang Provincial Natural Science Foundation of China (PL2024H172), and Haiyan Fund's General Projects (JJMS2023-01).

Data availability

The data underlying this article are available in TCGA, at http://gdac.broadinstitute.org/.

Code availability

The code to reproduce the results is available at https://github.com/Wanglabsmu/CATfusion.

References

1. 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.
2. 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.
3. Wang D, Liu B, Zhang Z. Accelerating the understanding of cancer biology through the lens of genomics. *Cell* 2023;**186**:1755–71. https://doi.org/10.1016/j.cell.2023.02.015.
4. Yao N, Zhang N, Wang J. *et al.* Experiences with cancer survey in China. *Cancer* 2019;**125**:3068–78. https://doi.org/10.1002/cncr.32164.
5. van Dooijeweert C, van Diest PJ, Ellis IO. Grading of invasive breast carcinoma: the way forward. *Virchows Arch* 2022;**480**:33–43. https://doi.org/10.1007/s00428-021-03141-2.
6. Seethala RR. Histologic grading and prognostic biomarkers in salivary gland carcinomas. *Adv Anat Pathol* 2011;**18**:29–45. https://doi.org/10.1097/PAP.0b013e318202645a.
7. Deepa P, Gunavathi C. A systematic review on machine learning and deep learning techniques in cancer survival prediction. *Prog Biophys Mol Biol* 2022;**174**:62–71. https://doi.org/10.1016/j.pbiomolbio.2022.07.004.
8. Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? *Nat Rev Cancer* 2012;**12**:323–34. https://doi.org/10.1038/nrc3261.
9. Di D, Zhang J, Lei F. *et al.* Big-hypergraph factorization neural network for survival prediction from whole slide image. *IEEE Trans Image Process* 2022;**31**:1149–60. https://doi.org/10.1109/TIP.2021.3139229.
10. 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.
11. Parvaiz A, Nasir ES, Fraz MM. From pixels to prognosis: a survey on AI-driven cancer patient survival prediction using digital histology images. *J Imaging Inform Med* 2024;**37**:1728–51. https://doi.org/10.1007/s10278-024-01049-2.
12. Saillard C, Schmauch B, Laifa O. *et al.* Predicting survival after hepatocellular carcinoma resection using deep learning on histological slides. *Hepatology* 2020;**72**:2000–13. https://doi.org/10.1002/hep.31207.
13. Wang Z, Ma J, Gao Q. *et al.* Dual-stream multi-dependency graph neural network enables precise cancer survival analysis. *Med Image Anal* 2024;**97**:103252. https://doi.org/10.1016/j.media.2024.103252.
14. Subramanian I, Verma S, Kumar S. *et al.* Multi-omics data integration, interpretation, and its application. *Bioinform Biol Insights* 2020;**14**:1177932219899051. https://doi.org/10.1177/1177932219899051.
15. Shao W, Wang T, Huang Z. *et al.* Weakly supervised deep ordinal Cox model for survival prediction from whole-slide pathological images. *IEEE Trans Med Imaging* 2021;**40**:3739–47. https://doi.org/10.1109/TMI.2021.3097319.
16. Gao J, Lyu T, Xiong F. *et al.* Predicting the survival of cancer patients with multimodal graph neural network. *IEEE/ACM Trans Comput Biol Bioinform* 2022;**19**:699–709. https://doi.org/10.1109/TCBB.2021.3083566.
17. Sun D, Wang M, Li A. A multimodal deep neural network for human breast cancer prognosis prediction by integrating

multi-dimensional data. *IEEE/ACM Trans Comput Biol Bioinform* 2018;**16**:841–50. https://doi.org/10.1109/TCBB.2018.2806438.
18. Tran KA, Kondrashova O, Bradley A. *et al.* Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Med* 2021;**13**:152. https://doi.org/10.1186/s13073-021-00968-x.
19. Chaudhary K, Poirion OB, Lu L. *et al.* Deep learning-based multi-omics integration robustly predicts survival in liver cancer. *Clin Cancer Res* 2018;**24**:1248–59. https://doi.org/10.1158/1078-0432.CCR-17-0853.
20. Stahlschmidt SR, Ulfenborg B, Synnergren J. Multimodal deep learning for biomedical data fusion: A review. *Brief Bioinform* 2022;**23**:bbab569. https://doi.org/10.1093/bib/bbab569.
21. Wei T, Yuan X, Gao R. *et al.* Survival prediction of stomach cancer using expression data and deep learning models with histopathological images. *Cancer Sci* 2023;**114**:690–701. https://doi.org/10.1111/cas.15592.
22. Sun Z, Lin M, Zhu Q. *et al.* A scoping review on multimodal deep learning in biomedical images and texts. *J Biomed Inform* 2023;**146**:104482. https://doi.org/10.1016/j.jbi.2023.104482.
23. Hao D, Li Q, Feng QX. *et al.* SurvivalCNN: a deep learning-based method for gastric cancer survival prediction using radiological imaging data and clinicopathological variables. *Artif Intell Med* 2022;**134**:102424. https://doi.org/10.1016/j.artmed.2022.102424.
24. 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.
25. Yin Z, Chen T, Shu Y. *et al.* A gallbladder cancer survival prediction model based on multimodal fusion analysis. *Dig Dis Sci* 2023;**68**:1762–76. https://doi.org/10.1007/s10620-022-07782-4.
26. Chen RJ, Lu MY, Wang J. *et al.* Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Trans Med Imaging* 2022;**41**:757–70. https://doi.org/10.1109/TMI.2020.3021387.
27. Wu X, Shi Y, Wang M. *et al.* CAMR: cross-aligned multimodal representation learning for cancer survival prediction. *Bioinformatics* 2023 Jan 13;**39**:btad025. https://doi.org/10.1093/bioinformatics/btad025.
28. 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.
29. Blum A, Wang P, Zenklusen JC. SnapShot: TCGA-analyzed Tumors. *Cell* 2018;**173**:530. https://doi.org/10.1016/j.cell.2018.03.059.
30. Chen C, Lu MY, Williamson DFK. *et al.* Fast and scalable search of whole-slide images via self-supervised deep learning. *Nat Biomed Eng* 2022;**6**:1420–34. https://doi.org/10.1038/s41551-022-00929-8.
31. Szalata A, Hrovatin K, Becker S. *et al.* Transformers in single-cell omics: a review and new perspectives. *Nat Methods* 2024;**21**:1430–43. https://doi.org/10.1038/s41592-024-02353-z.
32. Theodoris CV, Xiao L, Chopra A. *et al.* Transfer learning enables predictions in network biology. *Nature* 2023;**618**:616–24. https://doi.org/10.1038/s41586-023-06139-9.
33. Sharma A, Vans E, Shigemizu D. *et al.* DeepInsight: a methodology to transform a non-image data to an image for convolution neural network architecture. *Sci Rep* 2019;**9**:11399. https://doi.org/10.1038/s41598-019-47765-6.
34. Xu H, Usuyama N, Bagga J. *et al.* A whole-slide foundation model for digital pathology from real-world data. *Nature* 2024;**630**:181–8. https://doi.org/10.1038/s41586-024-07441-w.
35. Nechaev D, Pchelnikov A, Ivanova E. Hibou: a family of foundational vision transformers for pathology. *arXiv* 2024;**5**:2406.05074.
36. Ai K, Aben N, de Jong ED. *et al.* Towards large-scale training of pathology foundation models. *arXiv* 2024;**2**:2404.15217.
37. Filiot A, Jacob P, Mac Kain A. *et al.* Phikon-v2, a large and public feature extractor for biomarker prediction. *arXiv* 2024;**3**:2409.09173.
38. Zhang S, Xu Y, Usuyama N. *et al.* BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv* 2023;**4**:2303.00915.
39. Huang Z, Bianchi F, Yuksekgonul M. *et al.* A visual-language foundation model for pathology image analysis using medical twitter. *Nat Med* 2023;**29**:2307–16. https://doi.org/10.1038/s41591-023-02504-3.
40. Wang X, Yang S, Zhang J. *et al.* Transformer-based unsupervised contrastive learning for histopathological image classification. *Med Image Anal* 2022;**81**:102559. https://doi.org/10.1016/j.media.2022.102559.
41. Cisternino F, Ometto S, Chatterjee S. *et al.* Self-supervised learning for characterising histomorphological diversity and spatial RNA expression prediction across 23 human tissue types. *Nat Commun* 2024;**15**:5906. https://doi.org/10.1038/s41467-024-50317-w.
42. Schmauch B, Romagnoni A, Pronier E. *et al.* A deep learning model to predict RNA-seq expression of tumours from whole slide images. *Nat Commun* 2020;**11**:3877. https://doi.org/10.1038/s41467-020-17678-4.
43. Chen RJ, Lu MY, Williamson DFK. *et al.* Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell* 2022;**40**:e866. https://doi.org/10.1016/j.ccell.2022.07.004.
44. 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.
45. Ghaffari Laleh N, Muti HS, Loeffler CML. *et al.* Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. *Med Image Anal* 2022;**79**:102474. https://doi.org/10.1016/j.media.2022.102474.
46. Mayr A, Schmid M. Boosting the concordance index for survival data—a unified framework to derive and evaluate biomarker combinations. *PLoS One* 2014;**9**:e84483. https://doi.org/10.1371/journal.pone.0084483.
47. Andrade C. Survival analysis, Kaplan-Meier curves, and Cox regression: basic concepts. *Indian J Psychol Med* 2023;**45**:434–5. https://doi.org/10.1177/02537176231176986.
48. Zhang J, Ning J, Li R. Evaluating dynamic discrimination performance of risk prediction models for survival outcomes. *Stat Biosci* 2023;**15**:353–71. https://doi.org/10.1007/s12561-023-09362-0.
49. Ishwaran H, Gerds TA, Kogalur UB. *et al.* Random survival forests for competing risks. *Biostatistics* 2014;**15**:757–73. https://doi.org/10.1093/biostatistics/kxu010.
50. Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med* 1997;**16**:385–95. https://doi.org/10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3.

Briefings in Bioinformatics, 2025, **26(2)**, bbaf121
https://doi.org/10.1093/bib/bbaf121
Problem Solving Protocol