Original Paper

A Bilayer Feature Fusion Framework for Pan-Cancer Survival Prediction Based on Multihead Attention and Adaptive Differential Privacy: Model Development and Validation Study

Yun Chen[1], BE; Zhifang Deng[1], BE; Lili Wang[1], ME; Huanhuan Wang[1], PhD; Xiang Wu[1,2,3], PhD

[1]School of Medical Information and Engineering, Xuzhou Medical University, Xuzhou, China

[2]Yunlong Lake Laboratory of Deep Underground Science and Engineering, Xuzhou, China

[3]The Affiliated Hospital of Xuzhou Medical University, Xuzhou, China

Corresponding Author:
Xiang Wu, PhD
School of Medical Information and Engineering
Xuzhou Medical University
No. 209 Tongshan Road
Xuzhou, 221004
China
Phone: 86 18752167676
Email: wuxiang@xzhmu.edu.cn

Abstract

Background: In the field of precision medicine, pan-cancer survival prediction is crucial for individualized oncology diagnosis and treatment. Although multimodal data fusion techniques have significantly improved prediction accuracy, existing studies generally overlook the sensitivity of medical data and the need for privacy protection.

Objective: To address the aforementioned problem, this study aims to propose a bilayer feature fusion framework based on the multihead attention mechanism and adaptive differential privacy, which balances precise feature extraction and sensitive data protection.

Methods: Specifically, the multihead attention mechanism was integrated into bilayer feature extraction and fusion. Layer-wise relevance analysis was used to calculate the correlation between features and outcomes, and Laplacian noise was adaptively added based on the calculation results to achieve collaborative optimization of precise feature extraction and sensitive data protection. Additionally, the concordance index (C-index) and 5-fold cross-validation were used to compare the proposed method with state-of-the-art approaches.

Results: The proposed model achieved superior performance in both pan-cancer and single-cancer survival prediction, validated via the C-index and 5-fold cross-validation. In pan-cancer scenarios, the trimodal combination of clinical data, messenger RNA data, and microRNA expression data achieved the highest C-index of 0.799, outperforming multiple existing multimodal survival prediction approaches. After adaptive Laplacian noise injection for privacy protection, the model's accuracy decreased by only 0.01-0.03 while satisfying -differential privacy. For single-cancer prediction, the proposed method achieved higher C-index values in 18 out of 20 cancer types compared with representative deep learning–based survival models, with 7 types showing significant improvements (C-index difference >0.1) and more stable performance distributions. Furthermore, pan-cancer–trained models generally outperformed single-cancer–trained counterparts in most cancer types, highlighting the value of shared predictive features across cancers.

Conclusions: This study provides a solution that balances prediction accuracy and privacy security for pan-cancer survival prediction, laying the foundation for the efficient use of medical data under privacy protection. Future work may further integrate pathological images and proteomics data to expand the model's applications in cancer subtype classification and biomarker discovery.

(JMIR Med Inform 2026;14:e83743) doi: 10.2196/83743

KEYWORDS

adaptive differential privacy; multihead attention; pan-cancer survival prediction

Introduction

In the era of precision medicine, pan-cancer survival prediction refers to the task of modeling and predicting patient survival outcomes across multiple cancer types simultaneously by leveraging shared and cancer-specific prognostic patterns. As a core component of individualized oncology diagnosis and treatment, pan-cancer survival prediction has been widely validated for its clinical value. Accurately estimating survival risk for patients with different cancer types not only enables clinicians to optimize treatment strategies and personalize follow-up schedules, but also facilitates risk stratification and more efficient allocation of medical resources [1].

Early survival prediction efforts were predominantly anchored in unimodal data paradigms, relying on either genomic features or clinical variables in isolation. Statistical models laid the foundational groundwork for this field: the random survival forest (RSF) [2], for instance, addressed key challenges in survival analysis (eg, right-censored data and missing values) through log-rank–based tree splitting and adaptive imputation, establishing a robust nonparametric framework. Meanwhile, the classic Cox proportional hazards model (CPH) [3] became a clinical staple for its interpretability via linear combinations of clinical indicators, though its inability to capture complex nonlinear biological mechanisms limited its predictive power in heterogeneous cancer cohorts. The advent of deep learning catalyzed a paradigm shift in unimodal modeling: approaches like DeepSurv [4] integrated neural networks with Cox regression to automatically learn nonlinear covariate–treatment relationships, outperforming traditional models on genomic datasets. Subsequent studies further validated deep learning's superiority in capturing complex variable interactions; for example, in clinical data for oral squamous cell carcinoma [5] and histopathological images for colorectal cancer [6]. Weakly supervised learning also emerged as a promising direction for leveraging unannotated whole-slide images across multiple cancer types [7], while transfer learning facilitated cross-dataset generalization in disease-specific prognostic tasks [8]. Despite these advances, unimodal models inherently suffer from incomplete information capture: genomic data alone overlooks the spatial heterogeneity of tumor microenvironments (eg, angiogenesis patterns in pathology slides), while clinical variables fail to reveal molecular-level prognostic markers, creating an inherent ceiling for predictive performance.

In response to these limitations, multimodal data fusion has emerged as a transformative research focus in pan-cancer survival prediction, driven by the recognition that complementary data modalities (omics, imaging, and clinical) can synergistically enhance prognostic accuracy. Early multimodal frameworks integrated clinical, transcriptomic (messenger RNA [mRNA]/microRNA [miRNA]), and pathological data to achieve cross-cancer prediction [9], while recent innovations have incorporated advanced techniques such as unsupervised learning and attention mechanisms to dynamically weight modality contributions [10]. A key trend in this space is the development of specialized fusion architectures: the Multimodal Affinity Fusion Network [11] and MultiCoFusion [12]; for example, use attention mechanisms and graph convolutional networks to model intermodal dependencies and gene-gene interactions, respectively. Furthermore, bilinear fusion approaches, such as the attention-based multimodal bilinear fusion [13] and hierarchical factorized models [14], have addressed the curse of dimensionality to enable deeper cross-modal interactions. Notably, methods like Dynamic-DeepHit [15] have pioneered the use of recurrent neural network–based attention to model longitudinal dependencies for dynamic risk prediction, while cross-modal translation and alignment (CMTA) [16] explicitly learn and aligns intra- and cross-modal representations between pathology and genomics via encoder-decoder structures. End-to-end multimodal models such as MultiSurv [17], which integrate up to 6 modalities (whole-slide images, clinical data, and multiomics), have demonstrated state-of-the-art performance (concordance index [C-index]=0.779) across 33 cancer types, validating the value of comprehensive data integration. Other frameworks have incorporated multitask learning [18] or alternating training [19] to jointly optimize survival prediction and auxiliary tasks (eg, cancer grading), further improving model robustness. Recent work by Flack et al [20] introduced a robust multimodal survival prediction framework that effectively handles data heterogeneity and missing modalities through a regularization-based fusion strategy, showing competitive performance across multiple cancer types. Despite these advancements, critical challenges persist: inefficient cross-modal information alignment, inadequate intramodal feature refinement, and unresolved issues of intermodal heterogeneity and noise propagation remain major barriers to achieving consistent performance across diverse cancer types at the pan-cancer scale. More critically, the majority of these state-of-the-art multimodal survival prediction models, including MultiSurv [17], HFBSurv [14], attention-based multimodal bilinear fusion [13], and MBFusion [18], operate under the implicit assumption of centralized, nonsensitive data. They fundamentally lack integrated mechanisms to protect the highly sensitive genetic and clinical information inherent in medical datasets, creating a significant gap between methodological advancement and real-world clinical applicability where privacy is paramount.

Although multimodal data fusion has significantly improved pan-cancer survival prediction accuracy, existing research generally ignores the sensitivity of medical data and the need for privacy protection. Medical data contain sensitive information such as patients' genomic data, pathological images, and clinical records, whose leakage may lead to ethical risks such as identity recognition and insurance discrimination. For instance, empirical studies have demonstrated that even aggregated genomic data remain vulnerable to linkage attacks using auxiliary public databases, potentially inferring an individual's identity with high confidence [21]. Additionally, during the training and sharing of deep learning models, gradients or intermediate features may inadvertently leak sensitive patterns of the original training data [22]. While Chen et al [23] proposed an optimized logistic regression model (batch gradient descent logistic regression and balanced differential privacy logistic regression based on hybrid feature selection (Pearson correlation test + random forest out-of-bag algorithm) and differential privacy (DP) protection to address the problems

of insufficient prediction accuracy and privacy leakage in existing machine learning models for breast cancer prediction, and Chai et al [24] proposed the decentralized federated learning framework AdFed, combining regularization methods to train models without sharing raw data while achieving feature selection and privacy protection. To protect DNA data, Wu et al [25] proposed the DP-Motif Finding algorithm based on -DP, which uses closed frequent patterns to reduce redundant motifs, allocates privacy budget by constructing a perturbed extension tree, and uses best linear unbiased estimation postprocessing to optimize the noisy support. Wu et al [26] also proposed an adaptive federated learning scheme integrating DP, which adjusts the gradient descent process through an adaptive learning rate algorithm to avoid model overfitting and fluctuations, while introducing a DP mechanism to resist various background knowledge attacks, providing quantifiable privacy protection for the federated learning process. Wang et al [27] proposed a blockchain-based access control framework for genome-wide association studies in federated learning to protect the security of genetic data—this framework implements automated quality control to ensure training data quality, designs a blockchain-based authentication mechanism to filter malicious attackers, and adopts a periodic aggregation method combined with DP to accelerate cloud model training and resist multiple attacks [28]. Wang et al [29] proposed a method based on the ant colony optimization algorithm to detect gene interactions for genome-wide association studies—an intelligent privacy-preserving scheme. Current studies still suffer from problems such as single datasets and heavy reliance on data anonymization or centralized storage, failing to embed dynamic privacy protection mechanisms.

To solve the above problems, this study proposes a bilayer feature fusion framework based on the multihead attention (MHA) mechanism and adaptive DP, achieving collaborative optimization of precise feature extraction and sensitive data protection through technological innovation. Specifically, for feature fusion, traditional methods typically rely on fixed weights or simple concatenation to handle heterogeneous data. Simple concatenation, however, merely "physically stacks" feature vectors from different modalities, posing notable limitations. Multisource data (clinical, mRNA, miRNA, and gene copy number variation [CNV]) vary drastically in feature dimensions, semantics, and distributions; forcing them into a single-vector space postconcatenation overlooks inherent intermodal correlations; for example, links between specific mRNA expression and clinical survival time, or regulatory relationships between miRNA and CNV. Additionally, concatenation assumes "all features are equally important," failing to differentiate their contributions to survival prediction. This often buries key signals (eg, driver gene mutation sites) under redundancy, and may trigger the curse of dimensionality via rapid feature expansion, increasing computational complexity and overfitting risk. Given the aforementioned limitations of simple concatenation, our first contribution is the design of a structured bilayer feature extraction module. The first layer performs modality-specific feature extraction on clinical, mRNA, miRNA, and CNV data through fully connected (FC) networks and embedding layers. The second layer innovatively uses an MHA mechanism not merely as a fusion tool, but as a structured cross-modal interaction model that explicitly quantifies feature-level contributions. Unlike single-head attention, which models only single correlations, MHA captures differentiated correlations between features from different subspaces through parallelized independent attention heads and focuses on key survival-related features through dynamic weight allocation. Our second and primary contribution lies in the novel integration of adaptive DP directly into the multimodal feature extraction pipeline. At the privacy protection level, this study proposes adaptive DP based on layer-wise relevance analysis, balancing prediction accuracy and privacy security through a 2-step strategy. First, the layer-wise relevance propagation algorithm is used to quantify the association strength between neurons at each layer and survival prediction outcomes, identifying highly sensitive features (eg, driver gene mutation sites) in mRNA, miRNA, and CNV data; then privacy budgets are dynamically allocated based on correlation scores to inject Laplacian noise into the gradient update process—features with higher correlations receive less noise, preserving key prediction information while suppressing privacy leakage risks.

In summary, the novelty of our work is not merely the combination of MHA and DP, but the creation of a synergistic framework where (1) the bilayer MHA structure provides the granular feature importance signals necessary for adaptive noise allocation and (2) the adaptive DP mechanism, guided by layer-wise relevance analysis, protects privacy in a targeted manner that minimizes damage to the model's core predictive capability. This integrated solution addresses the specific limitations of prior multimodal models (which neglect privacy) and prior DP applications (which impair use in complex models), paving the way for clinically viable, privacy-preserving pan-cancer prediction tools.

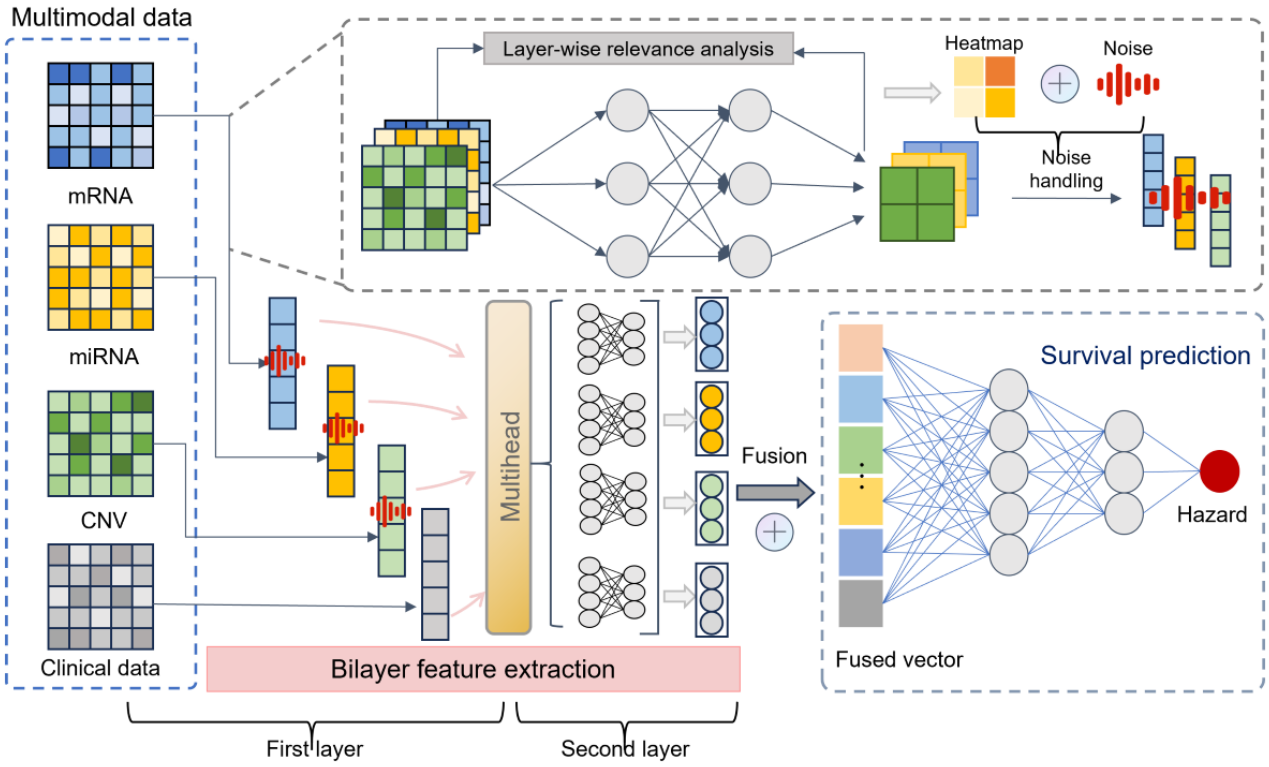
Methods

Overview

As shown in Figure 1, the proposed model architecture consists of 2 parts: a bilayer feature extraction module based on MHA and an adaptive DP protection module based on layer-wise relevance analysis. The bilayer feature extraction part, based on MHA, first performs first-layer feature extraction through FC layers and embedding layers, and then uses the MHA mechanism in the second layer for further feature extraction and multimodal data fusion. By integrating the MHA mechanism, the bilayer feature extraction module fully uses the advantage of parallel heads to capture different correlations, comprehensively and deeply mining key information from multimodal data. Additionally, a privacy protection component is added during the first-layer feature extraction, integrating layer-wise relevance analysis into the FC layers to deeply analyze the correlation between features and survival prediction outcomes through backpropagation. Noise is adaptively added based on the obtained correlation results. This study introduces a bilayer feature extraction module into the multimodal survival prediction model to better mine data features and provides privacy protection, effectively preserving the privacy information in the training data. The data preprocessing and

detailed method descriptions are presented in the following sections.

Figure 1. The architecture of the proposed model. CNV: gene copy number variation; miRNA: microRNA; mRNA: messenger RNA.



Data Preprocessing

In this study, 4 types of data were used: clinical data, gene expression (mRNA) data, miRNA data, and CNV data.

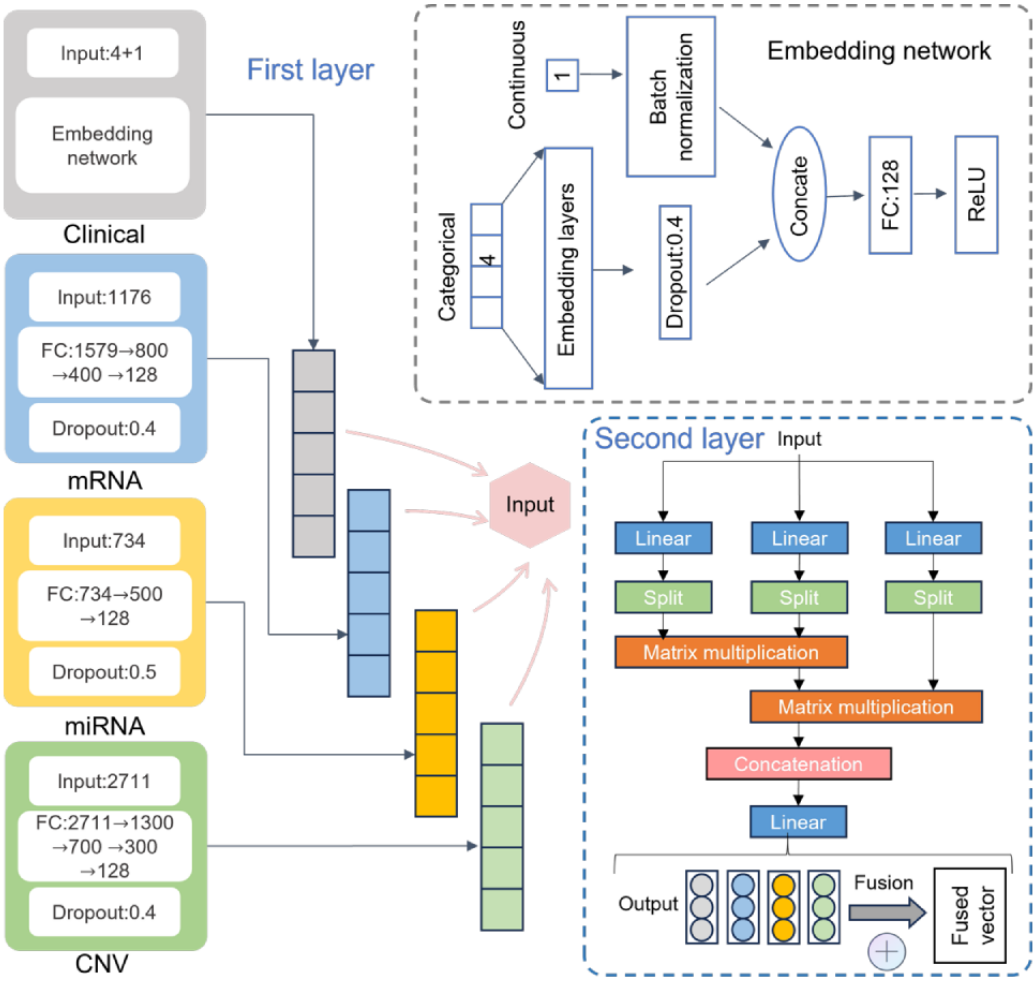
This section performs data screening for different data features. For clinical features, patient data with missing values and/or missing follow-up times were excluded. For gene expression data, a variance threshold method was used to select features with variances greater than a given threshold calculated from all patients [18,30]. In this study, the same thresholds as Fan et al [10] were used for mRNA and CNV modalities: 7 and 0.2, respectively. This processing retained 1579 mRNA features and 2711 CNV features. Subsequently, all continuous variables were normalized to the interval (0, 1) using minimum-maximum normalization [31].

Considering the variations in diagnosis and treatment methods across different cancer types, samples in the dataset may lack some required modalities. This study adopts zero-vector imputation for missing modalities, with its rationale mainly reflected in 2 aspects: first, zero vectors can explicitly indicate "no measured data for the modality," avoiding false biological signals introduced by statistical imputation methods. Second, zero-vector imputation ensures the consistency of feature dimensions across all samples, simplifies the model's input structure, and avoids additional computational overhead from complex missing-value processing modules, providing stability guarantees for gradient updates in the integration of adaptive DP mechanisms.

Bilayer Feature Extraction Based on MHA

To improve the accuracy and efficiency of feature extraction, this study designs a bilayer feature extraction method based on MHA, which can more comprehensively and deeply mine key information from multimodal data, as shown in Figure 2.

Figure 2. The architecture of bilayer feature extraction based on multihead attention. CNV: gene copy number variation; FC: fully connected; miRNA: microRNA; mRNA: messenger RNA; ReLU: rectified linear unit.



In recent years, neural networks with FC layers have been widely used for representation extraction [15]. In this study, different neural network settings with FC layers were used as the first-layer feature extraction to obtain representations from the raw numerical data of different modalities. Batch normalization was adopted to normalize the layer inputs and accelerate neural network training [30]. For the clinical data, we used categorical embedding layers for the 4 categorical variables, namely, cancer type, gender, race, and histological type, with dropout to encode the categorical variables into a numerical vector. The normalized continuous variables (ie, age) were then concatenated with these numerical vectors and fed into an FC layer with a fixed representation length. For miRNA, mRNA, and CNV modalities, 2-4 FC layers with rectified linear unit activation and batch normalization were used to extract fixed-length representations. The representation vectors $\beta_1, \beta_2, \ldots, \beta_n \in \mathbb{R}^{m\times 1}$ were obtained from modality-specific configurations to form the representation matrix $B \in \mathbb{R}^{m\times d}$ where m is the sample size and d is the feature dimension.

To effectively capture the complex interaction relationships between multimodal data and enhance feature representation capabilities, the MHA mechanism was introduced in the second-layer feature extraction stage.

The proposed MHA mechanism consists of 3 components similar to the self-attention in the transformer model: query (Q), key (K), and value (V). The matrix B is projected to generate keys K and values V. Specifically, V is processed using an identity function because the features extracted in the first layer already have sufficient representational power. For K, it is generated by:

$$K = Relu\left(Dropout\left(BW^K\right)\right) \quad (1)$$

Where $W^K \in \mathbb{R}^{d\times d}$ (d represents the feature dimension) is a trainable matrix randomly initialized from a standard Gaussian distribution and updated during training; B is the representation matrix obtained from the first-layer feature extraction. The dropout layer and rectified linear unit activation function are used to prevent overfitting and introduce nonlinearity. The query $Q \in \mathbb{R}^{1\times d}$ is a learnable vector initialized randomly, with elements sampled from a uniform distribution $\mathcal{U}\left(-\frac{1}{\sqrt{d}}, 1/\sqrt{d}\right)$.

The attention aggregation feature is calculated as:

$$G = Attention(Q,K,V) = softmax(QK^T)V = \sum_i^N A_i V_i \quad (2)$$

Where A is the attention weight. After softmax normalization along the sample dimension, the sum of the attention weights A for N samples is 1, enabling dynamic focusing on key survival-related features.

To capture multidimensional survival associations, the attention layer is extended to H heads. Specifically, Q, K, and V are split into blocks of size d/H along the embedding dimension. The attention for each head is calculated as:

$$head_{(h)} = Attention\left(Q_{(h)}, K_{(h)}, V_{(h)}\right) \quad (3)$$

These heads are then concatenated into a multihead output:

$$MHA(Q,K,V) = concat\left(head_{(1)}, head_{(2)}, \ldots, head_{(H)}\right) \quad (4)$$

This process allows different heads to capture interactions between different types of data separately.

Finally, the fused features are obtained by:

$$S = W^T droupout(G) \quad (5)$$

Where $W \in \mathbb{R}^{d\times 1}$ is the weight of the last layer. In evaluation mode, the dropout layer is disabled. Thus, the fused feature Y can be written as:

$$Y = W^T G = W^T concat\left(\sum_i^N A_{(1)i} V_{(1)i}, \ldots, \sum_i^N A_{(H)i} V_{(H)i}\right) = \sum_{h=1}^{H} \left(W_{(h)}^H \sum_i^N A_{(h)i} V_{(h)i}\right) \quad (6)$$

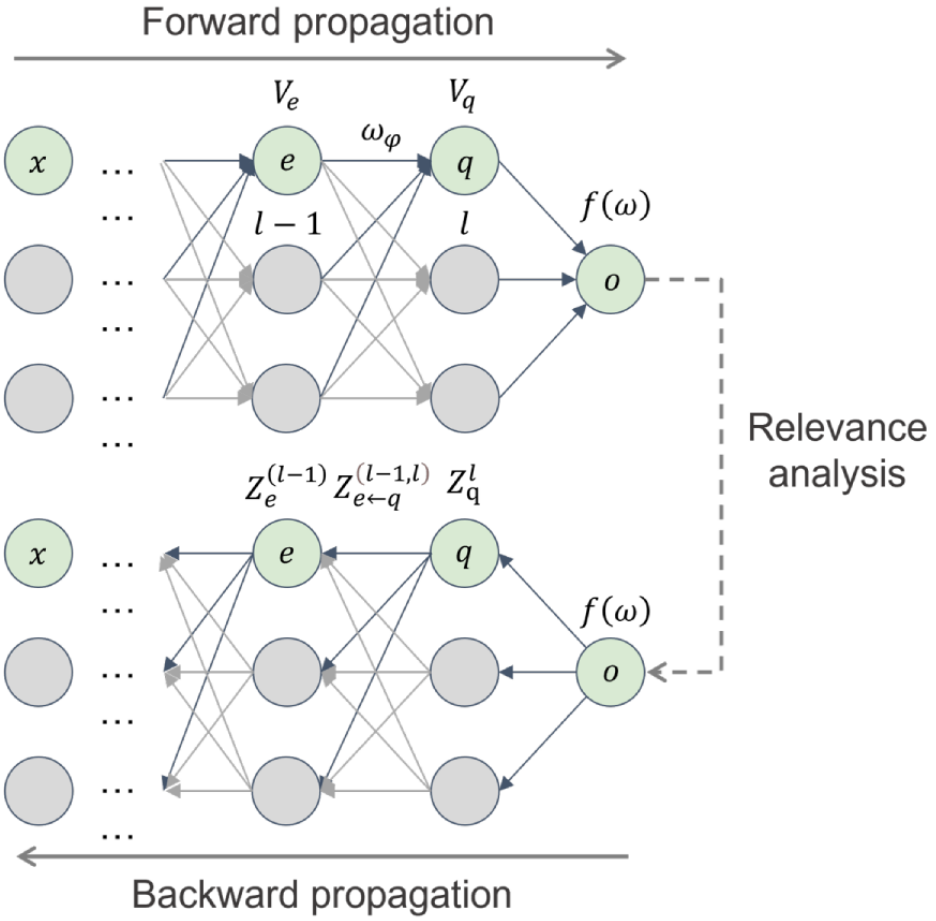
The attention mechanism focuses on key information through dynamic weight allocation, efficiently handling long-range dependencies in sequence data and suppressing noise interference. However, its single-head structure has limitations in modeling multidimensional correlations. Therefore, based on the features extracted from the 4 modalities in the first layer, this study introduces the MHA mechanism. By parallelizing multiple independent attention heads, it captures differentiated correlations between features from different subspaces, providing more comprehensive feature representations for pan-cancer survival prediction. The above method achieves deep fusion of multimodal survival features, offering high-precision feature representations for pan-cancer survival prediction.

Adaptive DP Based on Layer-Wise Relevance Analysis

In the field of cancer survival prediction, privacy protection of medical data is of utmost importance. Genetic data, such as mRNA, miRNA, and CNV data, contain highly sensitive personal information, posing a risk of privacy leakage. Therefore, during the feature extraction process, this study designs a method based on layer-wise relevance analysis to construct a noise addition mechanism by adding Laplacian noise for privacy protection, ensuring data security.

First, during the extraction of mRNA, miRNA, and CNV features in the first layer, this paper adds appropriate Laplace noise to the training gradient of neurons based on analyzing the correlation between each neuron layer and the output layer. The core theoretical operation process is shown in Figure 3. Relevance analysis begins after forward and backward propagation, calculating the correlation results of each feature in a layer of neurons. Importantly, the correlation analysis results are adjusted with the forward and backward propagation processes until the propagation ends.

Figure 3. The process of the layer-wise relevance analysis method of the proposed model based on the backpropagation algorithm.



The correlation between each input feature x_{ij} and the output $Hx_i(0)$ is calculated by decomposing the neurons of the previous layer. Given $Z_e^{(l)}(x_i)$ as the correlation resul $^{Z_{e\leftarrow q}^{(l-1,l)}(x_i)}$ t between x_i and neuron e in layer l, define as the process of neuron e sending information to q. The neuron correlation is:

$$Z_e^{(l-1)}(x_i) = \sum_{q\in h_i} Z_{e\leftarrow q}^{(l-1,l)}(x_i) \quad (7)$$

The decomposition rule for layer-wise relevance analysis is:

$$Z_{e\leftarrow q}^{(l-1)}(x_i) = \begin{cases} \frac{R_{eq}}{R_q+\varphi} Z_e^{(l)}(x_i), R_q \geq 0 \\ \frac{R_{eq}}{R_q-\varphi} Z_e^{(l)}(x_i), R_q \geq 0 \end{cases} \quad (8)$$

Where is a predefined stabilizer to address the unboundedness of $Z_e^{(l)}(x_i)$. R_q is the affine transformation of neuron e, defined as:

$$R_{eq} = v_e \omega_{eq} \quad (9)$$

$$R_q = \sum_e R_{eq} + u_q \quad (10)$$

Where v_e is the value of neuron e,$_{eq}$ is the weight between neuron e and q,u_q is the bias term.

In the last hidden layer, for the output variable o, the correlation is calculated as:

$$Z_q^{(l)}(x_i) = \begin{cases} \frac{R_{qo}}{R_o+\varphi} f_{x_i}(\omega), R_o \geq 0 \\ \frac{R_{eo}}{R_o-\varphi} f_{x_i}(\omega), R_o \geq 0 \end{cases} \quad (11)$$

This study directly calculates the correlation score $Z_j(x_{ij})$ between each feature x_{ij} inlayer j and the survival prediction result for mRNA, miRNA, and CNV features.

Next, a noise addition rule is defined. Through the above layer-wise relevance analysis, the value range of these correlation scores is divided into n nonoverlapping threshold intervals $^{\mathcal{T}=\{T_1,T_2,\ldots,T_n\}}$, and T_1 corresponds to the features with the highest relevance scores.

A privacy-budget mapping function $^{\eta:\eta(\mathcal{T})\to\mathbb{R}^+}$ is then defined, which assigns a specific privacy parameter $_k$ to each interval T_k. The function is designed to be strictly monotonic, satisfying the following condition: for any $g<h$,$\eta(T_g)<\eta(T_h)$, thereby ensuring that features with stronger predictive correlations are allocated a higher privacy budget (less noise), while features with weaker correlations receive a lower privacy budget (more noise).

Formally, for an interval index $k\in\{1,2,...,n\}$,the privacy parameter $_k$ is given by:

$$\epsilon_k = \epsilon_{max} - \frac{(\epsilon_{max} - \epsilon_{min})(k-1)}{n-1} + \delta_k \quad (12)$$

Where $_{max}$=0.5 and $_{min}$=0.1 define the permissible range of the privacy budget; $\delta_k \sim U(-\Delta,\Delta)$ is a bounded random perturbation term that preserves the strict monotonicity $_1 > _2 > \ldots > _n$.

Privacy budgets are adaptively allocated according to the function :

$$\epsilon_{ij} = \eta\left(Z_j(x_{ij})\right) \times \epsilon = \epsilon_k \times \epsilon, if\ Z_j(x_{ij}) \in T_k \quad (13)$$

Where denotes the global privacy budget.

Laplacian noise is added to the training gradients of neurons, where is the total privacy budget for protecting mRNA, miRNA, and CNV features.

At the start of training, the gradient update objective function of the general optimization method is defined. In each training step, using a set of random training samples L from the genetic data feature set X (where $X = \{x_{11}, x_{12}, \ldots, x_{ij}\}$), starting from the initial point $_0$, the parameters are updated at step t as:

$$\phi_{t+1} = \phi_t - \theta_t\left(\sigma\phi_t + \frac{1}{L}\sum_{i=1}\mathcal{L}\left(\phi_t, x_{ij}\right)\right) \quad (14)$$

where is the learning rate at step t, is the regularization parameter, and $^{\mathcal{L}}$ is the loss function.

Subsequently, the training gradients are perturbed to ensure the security of genetic data during training and sharing:

$$\phi_{t+1} = \phi_t - \theta_t\left(\sigma f_t + \frac{1}{L}\left(\sum_{i=1}\mathcal{L}\left(x_{ij}\right) + Y_t\right)\right) \quad (15)$$

Where Y_t is the Laplacian noise.

After detailing the process of adding Laplacian noise to features based on layer-wise relevance analysis, this section proceeds to prove the implementation of -DP. The following proof shows how the proposed method satisfies -DP requirements:

Proof: assume L and L' are 2 adjacent batches, and $_{t+1(L)}$ and $_{t+1(L')}$ are the parameters for L and L', respectively. The formula is expressed as follows:

$$\phi_{t+1(L)} = \phi_t - \theta_t\left(\sigma\phi_t + \frac{1}{L}\left(\sum_{i=1}^{L}\mathcal{L}\left(x_{ij}\right)\right)\right) \quad (16)$$

$$\phi_{t+1(L')} = \phi_t - \theta_t\left(\sigma\phi_t + \frac{1}{L}\left(\sum_{i=1}^{L}\mathcal{L}\left(x_{ij}{}'\right)\right)\right) \quad (17)$$

Then, the inequality for the difference between the 2 output results is:

$$\begin{aligned}\Delta_{\phi_t} &= \frac{\theta_t}{|L|}\sum_{\phi\in\phi_t}\left\|\sum_{x_{ij}\in\phi_t}\mathcal{L}\left(x_{ij}\right) - \sum_{x_{ij}{}'\in\phi'_t}\mathcal{L}\left(x_{ij}{}'\right)\right\|_1 \\ &\le \frac{\theta_t}{|L|}\sum_{\phi\in\phi_t}\left\|\sum_{x_{ij}\in\phi_t}\mathcal{L}\left(x_{ij}\right)\right\|_1 \\ &+ \left\|\frac{\theta_t}{|L|}\sum_{x_{ij}\in\phi_t}\left\|\sum_{x_{ij}{}'\in\phi'_t}\mathcal{L}\left(x_{ij}{}'\right)\right\|_1\right\|_1 \\ &\le 2\frac{\theta_t}{|L|}\max_{x_{ij}\in\phi_t}\sum_{\phi\in\phi_t}\left\|\mathcal{L}(x_{ij})\right\|_1\end{aligned} \quad (18)$$

From the above formula and DP, $_t$ is the sensitivity of the neural network **Inline graphic 27**. To protect the privacy information of the neural network, the gradient is perturbed based on relevance analysis, and the noise can be written as:

$$\phi_{t+1} = \phi_t - \theta_t\left(\sigma\phi_t + \frac{1}{L}\left(\sum_{i=1}^{L}\mathcal{L}\left(X_i\right) + Lap\left(\frac{\Delta_{\phi_t}}{\epsilon_{ij}}\right)\right)\right) \quad (19)$$

We have:

$$\begin{aligned}&\frac{Pr[\phi_{t+1(L)}]}{Pr[\phi_{t+1(L')}]} \\ &= \frac{\prod_{\phi\in\phi_t}\prod_{i=1}n\,exp\left(\frac{\epsilon_{ij}\frac{\theta_t}{|L|}\left\|\sum_{x_{ij}\in\phi_t}\mathcal{L}\left(x_{ij}\right) - \left(\sum_{x_{ij}\in\phi_t}\mathcal{L}\left(x_{ij}\right) + Lap\left(\frac{\Delta_{\phi_t}}{\epsilon_{ij}}\right)\right)\right\|_1}{\Delta_{\phi_t}}\right)}{\prod_{\phi\in\phi_t}\prod_{i=1}n\,exp\left(\frac{\epsilon_i\frac{\theta_t}{|L|}\left\|\sum_{x_{ij}'\in\phi'_t}\mathcal{L}\left(x_{ij}'\right) - \left(\sum_{x_{ij}\in\phi_t}\mathcal{L}\left(x_{ij}\right) + Lap\left(\frac{\Delta_{\phi_t}}{\epsilon_{ij}}\right)\right)\right\|_1}{\Delta_{\phi_t}}\right)} \\ &\le \prod_{\phi\in\phi_t}\prod_{i=1}n\,exp\left(\frac{\epsilon_{ij}\frac{\theta_t}{|L|}}{\Delta_{\phi_t}}\left\|\sum_{x_{ij}\in\phi_t}\mathcal{L}\left(x_{ij}\right) - \sum_{x_{ij}'\in\phi'_t}\mathcal{L}\left(x_{ij}'\right)\right\|_1\right) \\ &\le \prod_{\phi\in\phi_t}\prod_{i=1}n\,exp\left(\frac{\epsilon_{ij}\frac{\theta_t}{|L|}}{\Delta_{\phi_t}}2\max_{x_{ij}\in\phi_t}\left\|\mathcal{L}(x_{ij})\right\|_1\right) \\ &\le \prod_{\phi\in\phi_t}\prod_{i=1}n\,exp\left(\epsilon\frac{2\frac{\theta_t}{|L|}\frac{r_j}{Z_j(x_{ij})}}{\Delta_{\phi_t}}\right) = exp(\epsilon)\end{aligned} \quad (20)$$

This proves that the method satisfies -DP, ie, $^{\Pr[\mathcal{M}(L)=o] \le e^{\epsilon}\Pr[\mathcal{M}(L')=o]}$, where M is the model after adding Laplacian noise to features. It ensures that the model's noise addition mechanism is consistent with the definition of DP while protecting the privacy of mRNA, miRNA, and CNV features.

This method not only effectively reduces the risk of data leakage but also maximizes data usability, providing a more secure and reliable solution for cancer survival prediction and promoting the efficient use and sharing of medical data under privacy protection.

Ethical Considerations

All data used in this study were obtained from The Cancer Genome Atlas (TCGA), a publicly available database. The TCGA project obtained informed consent from all participants and received ethical approval from the appropriate institutional review boards. As this study involved only the use of deidentified, publicly accessible data, no additional ethical approval was required.

Results

Data

The clinical, mRNA, and miRNA data were downloaded from the Pan-Cancer Atlas of TCGA project (publicly accessible at the Genomic Data Commons), where more than 11,000 patient samples across 33 tumor types [32,33] were collected. The clinical dataset provides annotations for 11,160 patients, using 5 variables consistent with Silva and Rohr [17]: cancer type, gender, race, histological type, and age. The miRNA data set contains 743 miRNA feature records, and the mRNA dataset contains RNA sequencing counts for 20,531 mRNAs. The CNV data were downloaded from the University of California Santa Cruz Xena [34].

Table 1 describes the data information after preprocessing in detail. In this study, the pan-cancer datasets were partitioned

into training (6656/11094, 60%), validation (2219/11094, 20%), and testing (2219/11094, 20%) subsets, with model performance evaluated using 5-fold cross-validation.

Table 1. Summary information of the different modalities after preprocessing.

Modality	Patients, n	Features, n		All zero vector, n/N (%)
		Continuous	Categorical	
Clinical	11,094	1	4	—[a]
miRNA[b]	11,094	743	—	72/743 (9.7)
mRNA[c]	11,094	1579	—	49/1579 (3.1)
CNV[d]	11,094	2711	—	225/2711 (8.3)

[a]Not available.

[b]miRNA: microRNA.

[c]mRNA: messenger RNA.

[d]CNV: gene copy number variation.

Experimental Settings

We conduct all experiments on a simulation environment equipped with a 64-bit Intel(R) Xeon(R) Silver 4210R CPU @ 2.40 GHz processor, 32 GB RAM, and an NVIDIA GeForce RTX 3080 GPU for accelerated computation. The experiments are implemented on the Windows 11 operating system, with Python (version 3.10; Python Software Foundation) as the primary programming language and PyTorch (version 2.6.0; Meta AI) as the deep learning framework.

For model training, the Adam optimizer is adopted with a fixed learning rate of 0.001. Key hyperparameters are configured as follows: sequence length is set to 128, batch size is 256, and total training epochs are 100.

Evaluation Metrics

In this study, we assessed the performance of our model by using the C-index, a widely adopted metric for evaluating survival predictions in censored survival data [2,35]. The C-index measures the proportion of concordant pairs among all possible evaluation pairs, defined as:

$$C-index=\frac{\sum_{i\neq j}1\left(b(x_i)<b(x_j)\right)1(T_i>T_j)E_j}{\sum_{i\neq j}1(T_i>T_j)E_j} \quad (21)$$

where 1 is the indicator function of whether the expression in parentheses is true or false. The C-index ranges from 0 to 1. The closer the C-index is to 1, the closer the prediction order is to the real one; the closer the C-index is to 0.5, the closer the model's prediction is to a random prediction. Notably, the C-index focuses on the ordinal relationship of predictions rather than the accuracy of individual sample forecasts, making it particularly suitable for evaluating proportional hazards models.

Accuracy was also used for performance evaluation. The metric is evaluated as follows:

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \quad (22)$$

Layer-Wise Relevance Analysis

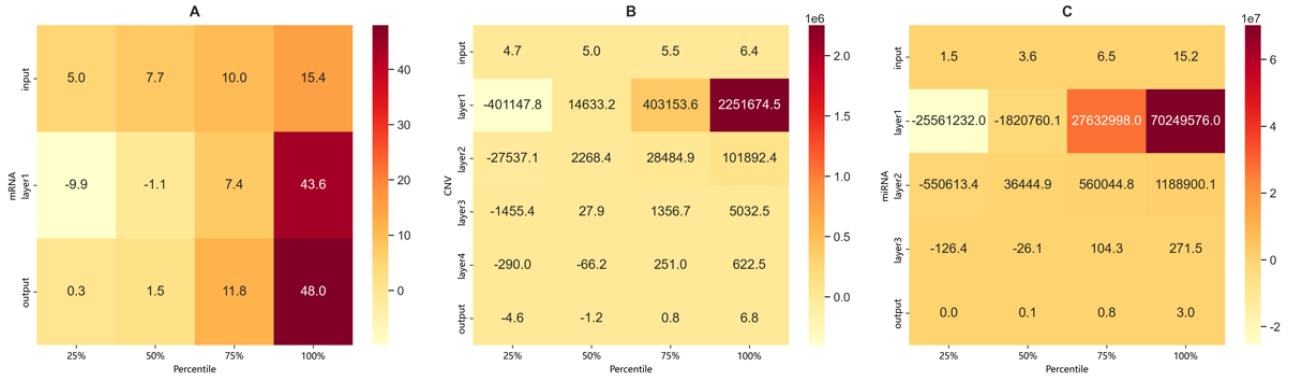
To evaluate the model's performance in privacy preservation, this study determines the privacy budget through hierarchical correlation analysis and subsequently adds Laplacian noise based on the results. Specifically, this study analyzed the strength of correlations between features and different FC layers using heatmaps for the 3 modalities: mRNA, miRNA, and CNV. The computed correlation scores are sorted in descending order, and the feature scores at the 25th, 50th, 75th, and 100th percentiles are visualized. The heatmaps show that higher correlation scores indicate stronger associations between the corresponding features and survival prediction outcomes.

A more detailed analysis of each modality's heatmap reveals that, as shown in Figure 4A, the output layer and mRNA layer 1 exhibit significantly higher correlation scores at the 100th percentile (48.0 and 43.6, respectively), with notable increases compared to the 25th, 50th, and 75th percentiles. This suggests that these 2 layers are particularly effective in capturing high-correlation features, which are strongly associated with survival outcomes. These findings indicate that privacy-preserving noise injection should focus on these layers to avoid the leakage of critical predictive information.

For the CNV and miRNA modalities, as illustrated in Figures 4B and 4C, layer 1 demonstrates markedly higher correlation scores at the 100th percentile compared to subsequent layers and the output layer. This suggests that the model primarily extracts key features from layer 1, with diminishing correlation strength in deeper layers. Therefore, noise injection strategies should prioritize high-correlation features in layer 1, enhancing privacy protection while maintaining the model's feature extraction performance.

Overall, the heatmaps of layer-feature correlations across the 3 modalities not only identify the core layers and key percentile-based features involved in the model's internal processing, but also provide empirical support for refining noise injection strategies. This contributes to achieving a better balance between privacy protection and model performance.

Figure 4. Heatmaps of correlation scores between layers and features in (A) messenger RNA (mRNA), (B) gene copy number variation (CNV), and (C) microRNA (miRNA) modalities.



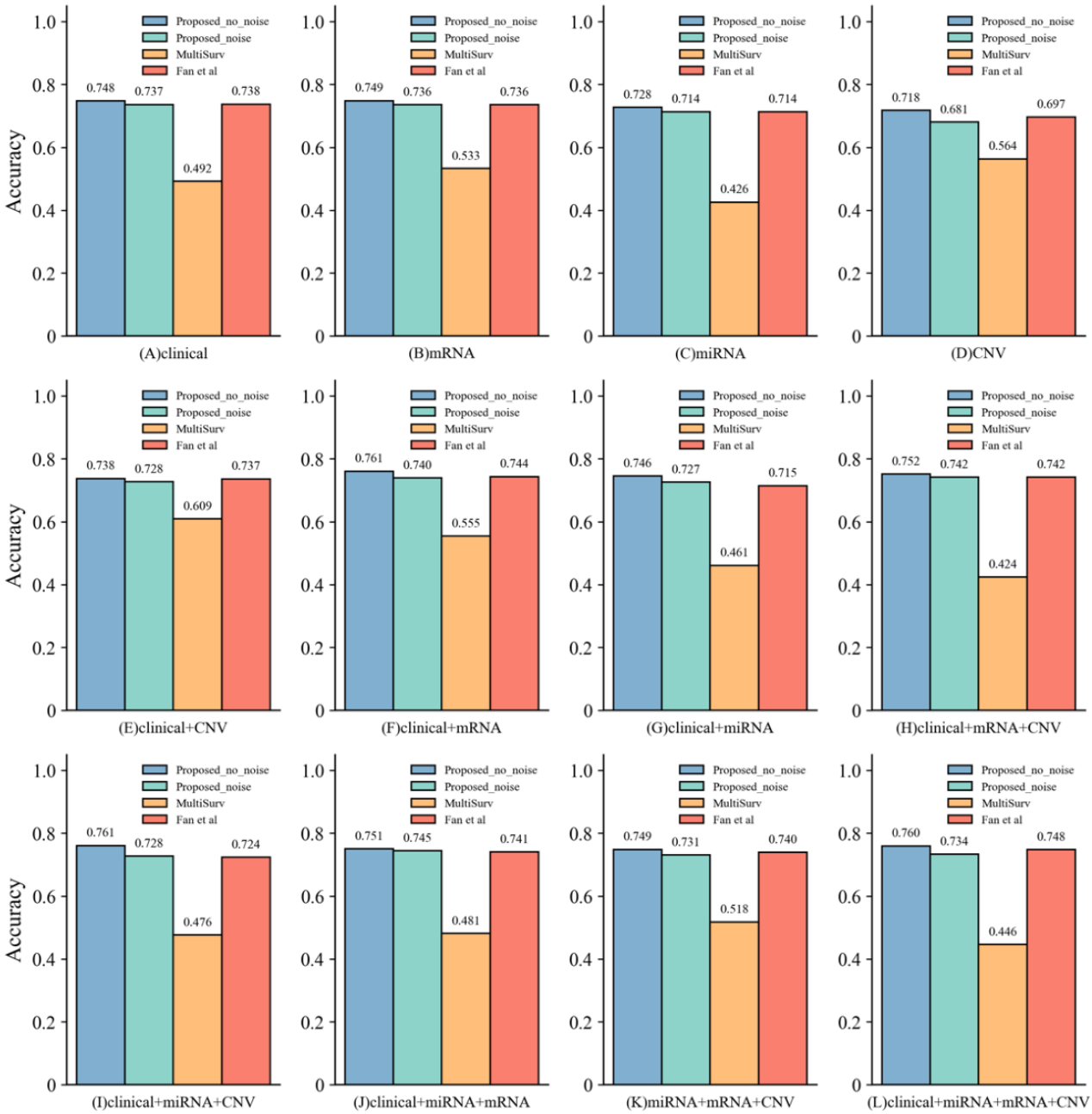
Privacy Protection Effect Comparison of Multimodal Methods

To further evaluate the impact of noise addition on model performance, this study uses accuracy as an auxiliary metric, with patients' survival status as the classification criterion (where deceased patients are labeled as 1 and surviving patients as 0), and implements the classification and prediction of survival status through a neural network with 3 FC layers. Leveraging the pan-cancer dataset, a systematic analysis is conducted on the model's survival classification performance across 4 data combinations: clinical data, mRNA, miRNA, and CNV. The proposed model is compared with 2 state-of-the-art attention-based models for pan-cancer survival prediction: MultiSurv [17] and Fan et al [10].

In unimodal settings, as shown in Figures 5A-5D, the proposed model demonstrates strong robustness. For instance, in the miRNA modality, accuracy slightly decreases from 0.728 before noise injection to 0.714 after; in the mRNA modality, accuracy drops modestly from 0.749 to 0.736.

Experiments involving multimodal combinations were conducted in dual-, tri-, and 4-modality settings. As shown in Figures 5E-5L, the model also maintains stable performance under multimodal conditions. For example, the clinical + miRNA + mRNA combination achieves an accuracy of 0.751 before noise addition and 0.746 afterward; for the clinical + mRNA combination, accuracy declines from 0.761 to 0.740 after noise injection.

Figure 5. Accuracy comparison results between "Proposed_no_noise" (blue bars), "Proposed method" (green bars), "MultiSurv" (orange bars), and Fan et al [10] (red bars) across (A) clinical, (B) messenger RNA (mRNA), (C) microRNA (miRNA), (D) gene copy number variation (CNV), (E) clinical + CNV, (F) clinical + mRNA, (G) clinical + miRNA, (H) clinical + mRNA + CNV, (I) clinical + miRNA + CNV, (J) clinical + miRNA + mRNA, (K) miRNA + mRNA + CNV, and (L) miRNA + mRNA + CNV modalities.



Survival Prediction Using Pan-Cancer Training Dataset

To evaluate the survival prediction advantages of the proposed model across different modalities, this study adopted the C-index as the performance metric and compared it against 6 state-of-the-art models: the CPH [3], RSF [2], MultiSurv, DeepHit [15], CMTA [16], and Fan et al [10]. The experiments are conducted using clinical data, mRNA, miRNA, and CNV data in both unimodal and multimodal settings. All results were obtained via 5-fold cross-validation and are summarized in Table 2.

Table 2. Model performance on mixed cancer types using single-modal and multimodal training datasets by the measurements of concordance index (C-index).

Modality	C-index						
	CPH[a] [3]	RSF[b] [2]	MultiSurv [17]	DeepHit [15]	CMTA[c] [16]	Fan et al [10]	Proposed model
Clinical	0.633 (±0.010)	0.745 (±0.008)	0.749 (±0.021)	0.742 (±0.013)	0.758 (±0.009)	0.756 (±0.013)	0.775 (±0.008)[d]
mRNA[e]	0.715 (±0.013)	0.739 (±0.013)	0.747 (±0.012)	0.744 (±0.008)	0.760 (±0.010)	0.764 (±0.013)	0.779 (±0.007)[d]
miRNA[f]	0.715 (±0.010)	0.735 (±0.005)	0.725 (±0.012)	0.701 (±0.008)	0.712 (±0.007)	0.738 (±0.011)	0.760 (±0.011)[d]
CNV[g]	0.589 (±0.008)	0.626 (±0.007)	0.649 (±0.006)	0.579 (±0.006)	0.618 (±0.0011)	0.646 (±0.009)	0.659 (±0.009)[d]
Clinical + CNV	—[h]	—	0.752 (±0.008)	0.747 (±0.006)	0.758 (±0.012)	0.753 (±0.005)	0.763 (±0.006)[d]
Clinical + mRNA	—	—	0.773 (±0.008)	0.764 (±0.011)	0.785 (±0.010)	0.784 (±0.007)	0.792 (±0.006)[d]
Clinical + miRNA	—	—	0.758 (±0.011)	0.760 (±0.008)	0.763 (±0.007)	0.766 (±0.010)	0.783 (±0.007)[d]
Clinical + mRNA + CNV	—	—	0.766 (±0.012)	0.768 (±0.013)	0.774 (±0.013)	0.779 (±0.007)	0.785 (±0.004)[d]
Clinical + miRNA + CNV	—	—	0.760 (±0.009)	0.761 (±0.010)	0.762 (±0.009)	0.764 (±0.007)	0.775 (±0.008)[d]
Clinical + miRNA + mRNA	—	—	0.771 (±0.014)	0.773 (±0.014)	0.779 (±0.013)	0.777 (±0.013)	0.799 (±0.012)[i]
miRNA + mRNA + CNV	—	—	0.747 (±0.013)	0.752 (±0.012)	0.770 (±0.007)	0.767 (±0.008)	0.778 (±0.010)[d]
Clinical + miRNA + mRNA + CNV	—	—	0.768 (±0.013)	0.770 (±0.010)	0.772 (±0.012)	0.779 (±0.010)	0.787 (±0.007)[d]

[a]CPH: Cox proportional hazards model.

[b]RSF: random survival forest.

[c]CMTA: cross-modal translation and alignment.

[d]$P<.05$ vs MultiSurv, DeepHit, CMTA, and Fan et al [10] (paired *t* test with Bonferroni correction, n=1000 bootstrap samples).

[e]mRNA: messenger RNA.

[f]miRNA: microRNA.

[g]CNV: gene copy number variation.

[h]Not available.

[i]$P<.01$ vs CMTA (top competitor; paired *t* test with Bonferroni correction, n=1000 bootstrap samples).

To substantiate claims of superiority despite modest C-index differences, statistical significance testing was performed using bootstrapped resampling and 2-tailed paired *t* tests. For each model and modality combination, 1000 bootstrap samples were generated from the test set. Paired *t* tests were conducted on the bootstrapped C-index values (n=1000) to test the null hypothesis that the mean performance difference is 0, with Bonferroni correction applied for multiple comparisons. Additionally, 95% CIs were derived from the bootstrap distributions. Nonoverlapping CIs are interpreted as indicating statistically significant differences ($P<.05$).

In unimodal prediction scenarios, the proposed model achieves C-index scores of 0.775, 0.779, 0.760, and 0.659 for clinical, mRNA, miRNA, and CNV data, respectively—significantly outperforming the CPH and RSF, which are limited to unimodal survival prediction. Notably, the proposed model yields the highest performance on mRNA data, with a C-index of 0.779, compared to 0.715 for CPH and 0.739 for RSF.

Furthermore, in the multimodal setting, this study compared the performance of MultiSurv, DeepHit, CMTA, Fan et al [10], and the proposed model across various combinations of data modalities. As shown in Table 2, the proposed model consistently achieves C-index values around 0.78 across different combinations. The highest performance is observed in the trimodal configuration (clinical + mRNA + miRNA), which yields a C-index of 0.799—surpassing MultiSurv (0.771),

DeepHit (0.773), CMTA (0.779), and Fan et al [10] (0.777). To statistically validate these performance improvements, paired significance testing with multiple-comparison adjustment was conducted. For each model and modality combination, 1000 bootstrap samples were generated from the test set. Paired *t* tests were performed on the bootstrapped C-index values (n=1000) with Bonferroni correction applied for multiple comparisons. The proposed model demonstrated statistically significant improvements over all competitors in the optimal trimodal configuration (clinical + mRNA + miRNA) after multiplicity adjustment: vs CMTA (*P*=.008), Fan et al [10] (*P*=.006), MultiSurv (*P*=.003), and DeepHit (*P*=.004). These adjusted *P* values (<.05) confirm that the observed C-index improvements (0.020, 0.022, 0.028, and 0.026 relative to CMTA, Fan et al [10], MultiSurv, and DeepHit, respectively) are statistically robust and not attributable to random variation.

This superior performance is attributed to the model's dual-layer feature extraction architecture, which, leveraging multiview parallel processing, enables finer-grained automatic weight adjustment. Notably, our proposed model further integrates an adaptive DP protection mechanism to safeguard sensitive medical data, while still achieving a higher C-index compared to existing state-of-the-art multimodal models that lack privacy-preserving designs. These findings reinforce the advantage of multimodal inputs over unimodal inputs, as multimodal data provide more comprehensive and informative cues for survival prediction.

Comparison With Standard Privacy Protection Methods

To comprehensively validate the superiority of the proposed adaptive DP framework in balancing privacy protection and predictive utility, this study adopted 2 core metrics—C-index for survival outcome ranking accuracy and classification accuracy for binary survival status prediction (deceased=1 and surviving=0)—and systematically compared it against 3 standard DP-aware baseline models: DP-Stochastic Gradient Descent (SGD) [36], DP-Adam [37], and Private Aggregation of Teacher Ensembles (PATE) [38]. The experiments were conducted under the optimal multimodal setting (clinical + mRNA + miRNA), with a fixed global privacy budget of = 0.8 (δ = 1e-5 for DP-SGD and DP-Adam, as per their standard implementations) to ensure a fair comparison of use under identical formal privacy guarantees. Consistent data preprocessing (zero-vector imputation for missing modalities, minimum-maximum normalization) and model training configurations (Adam optimizer, learning rate=0.001, batch size=256, and 100 epochs) were used across all methods to eliminate confounding variables. All results were derived from 5-fold cross-validation on the pan-cancer test set and supplemented with 1000 bootstrap resamples for statistical significance testing (paired *t* test with Bonferroni correction, *P*<.05). Detailed performance metrics and comparative analysis are summarized in Table 3.

Table 3 clearly shows that the proposed adaptive DP framework outperforms all standard DP baselines in maintaining predictive utility under the same formal privacy guarantee (=0.8). Regarding the C-index—the gold standard for survival prediction that assesses the ordinal consistency between predicted risk scores and actual survival times—the proposed model achieves 0.799 (±0.012), which is 4.3, 3.7, and 6.5 percentage points higher than DP-SGD (0.756 ± 0.015), DP-Adam (0.762 ± 0.014), and PATE (0.734 ± 0.016), respectively. Its utility loss compared to the non-DP counterpart (0.821 ± 0.008) is only 0.022, which is less than half of DP-SGD's loss (0.065) and about 1/3 of PATE's loss (0.087). This confirms that the layer-wise relevance propagation–guided targeted noise injection strategy more efficiently uses the privacy budget, preserving high-relevance prognostic features while satisfying -DP.

In terms of binary survival status classification accuracy, the proposed model maintains a high level of 0.745 (±0.013), outperforming the aforementioned baselines by 0.8, 1.0, and 1.4 percentage points, respectively. Its accuracy drop (0.006) is also significantly smaller than that of the baselines (0.014-0.020), further validating the efficiency of adaptive noise allocation in minimizing utility degradation for a given . Statistical analysis (paired *t* test with Bonferroni correction, n=1000 bootstrap samples) reveals that the proposed model's C-index and accuracy are significantly different from all baselines (*P*<.05), with nonoverlapping 95% CIs, confirming the robustness of the observed improvements.

Table 3. Use comparison of proposed model vs standard differential privacy (DP) baselines (multimodal: clinical + messenger RNA + microRNA; =0.8).

Modality	C-index[a]	Accuracy
Proposed_no_noise	0.821 (±0.008)	0.751 (±0.008)
Proposed model	0.799 (±0.012)	0.745 (±0.013)
DP-SGD[b]	0.756 (±0.015)	0.737 (±0.010)
DP-Adam	0.762 (±0.014)	0.735 (±0.008)
PATE[c]	0.734 (±0.016)	0.731 (±0.009)

[a]C-index: concordance index.

[b]SGD: Stochastic Gradient Descent

[c]PATE: Private Aggregation of Teacher Ensembles.

Analysis of Privacy Protection Effectiveness

The proposed adaptive DP mechanism is designed to satisfy -DP (=0.8) as formally proven in equation 20. Compared with traditional uniform noise injection methods (DP-SGD and DP-Adam) under the same global privacy budget, our method significantly improves the accuracy of survival prediction (with the C-index increased by 3.7-6.5 percentage points; Table 3). This result indicates that, for the same formal privacy guarantee ( ), our adaptive allocation strategy achieves higher predictive utility. The adaptive mechanism identifies features with high relevance to prediction and allocates a larger portion of the privacy budget to them (injecting less noise), thereby preserving critical prognostic information. Conversely, features with low predictive contribution receive a smaller budget (more noise), which theoretically enhances privacy protection for those elements without harming overall model utility. Therefore, the proposed method provides a stronger formal privacy guarantee for low-relevance features and achieves superior utility for high-relevance features under the same constraint, representing a more efficient use of the privacy budget.

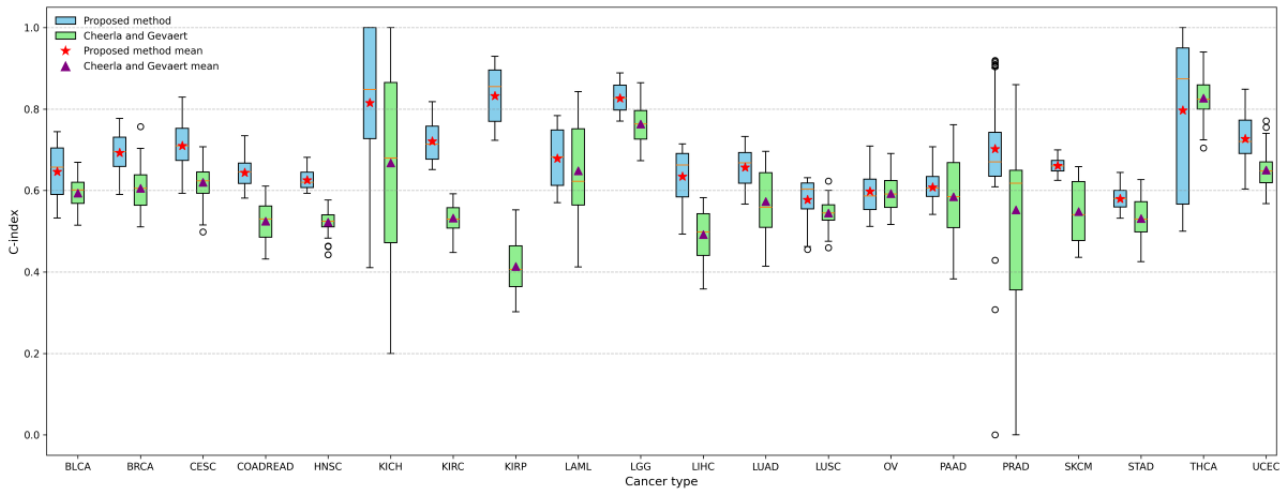
Comparison of Survival Prediction Performance for Individual Cancer Types

To investigate the effectiveness of the proposed model in predicting survival outcomes for individual cancer types, this study conducted a comparison between our proposed model and the state-of-the-art deep learning–based model (Cheerla and Gevaert [9]) for single-cancer prediction using pan-cancer datasets. For consistency, this study adopted a multimodal input comprising clinical, miRNA, and mRNA data. The results are presented in Figure 6.

Compared with Cheerla and Gevaert [9], our proposed method achieved a higher C-index in the majority of cancer types—18 out of 20. Notably, in 7 of these 18 cancer types, the performance improvement was substantial, with C-index differences exceeding 0.1.

Furthermore, as illustrated in Figure 6, our model exhibits higher median C-index values for several cancer types and demonstrates more concentrated distributions in the upper range. This indicates both superior and more stable predictive performance. In contrast, Cheerla and Gevaert [9] show relatively lower C-index scores in some cancer types, with several boxplots revealing greater data dispersion and reduced stability.

Figure 6. Concordance index (C-index) of the proposed model (blue bars) and the previous work (Cheerla and Gevaert [9]; green bars) on the 20 cancer types using the modality combination of clinical, microRNA, and messenger RNA. BLCA: bladder urothelial carcinoma; BRCA: breast invasive carcinoma; CESC: cervical squamous cell carcinoma; COADREAD: colon adenocarcinoma/rectum adenocarcinoma; HNSC: head and neck squamous cell carcinoma; KICH: kidney chromophobe; KIRC: kidney renal clear cell carcinoma; KIRP: kidney renal papillary cell carcinoma; LAML: acute myeloid leukemia; LGG: low-grade glioma; LIHC: liver hepatocellular carcinoma; LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma; OV: ovarian serous cystadenocarcinoma; PAAD: pancreatic adenocarcinoma; PRAD: prostate adenocarcinoma; SKCM: skin cutaneous melanoma; STAD: stomach adenocarcinoma; THCA: thyroid carcinoma; UCEC: uterine corpus endometrial carcinoma.



Comparison Between Using Single-Cancer and Pan-Cancer Training Datasets

To further investigate the predictive performance of the proposed model across different types of datasets, this study conducted experiments using both pan-cancer and single-cancer datasets. The first objective was to examine whether training on pan-cancer data improves survival prediction accuracy for individual cancer types. In this study, we selected the same 20 cancer types as those used by Cheerla and Gevaert [9], where patients have significantly different survival patterns.

This study first compared the performance of models trained solely on single-cancer data with those trained on all pan-cancer samples, using consistent test sets for each cancer type. For single-cancer experiments, this study selected patients with the same cancer type in pan-cancer training-validation-test sets to form the single-cancer training-validation-test sets. The model was trained using integrated clinical, mRNA, and miRNA modalities, and the results are shown in Figure 7.

As shown by the bar plots labeled “Single cancer” and “Pan-cancer (all)” in Figure 7, the model trained on pan-cancer data generally outperformed the single-cancer–trained model

on the same test sets, with the exception of 4 cancer types: kidney chromophobe, acute myeloid leukemia, prostate adenocarcinoma, and stomach adenocarcinoma. For example, in the case of bladder urothelial carcinoma the pan-cancer–trained model achieved a C-index of 0.665, significantly higher than the 0.612 achieved by the model trained only on bladder urothelial carcinoma data.

To control for potential biases due to varying training sample sizes, this study further evaluated model performance when trained on an equal number of samples from the pan-cancer and single-cancer datasets. Specifically, we selected subsets of pan-cancer data matching the sample size of each individual cancer dataset. As illustrated by the bar plots labeled "Single cancer" and "Pan-cancer (same)" in Figure 7, the single-cancer–trained models generally outperformed the pan-cancer–trained counterparts under equal training size conditions.

Figure 7. Concordance index (C-index) scores of the proposed model trained on single-cancer and pan-cancer datasets on 20 cancer types using clinical, messenger RNA, and microRNA modalities. Red bars represent models trained on single-cancer datasets, where only patients with the same cancer type were used. Blue bars indicate models trained on pan-cancer datasets with sample sizes matched to each single-cancer dataset to control for sample size bias. Orange bars correspond to models trained on the full pan-cancer dataset. BLCA: bladder urothelial carcinoma; BRCA: breast invasive carcinoma; CESC: cervical squamous cell carcinoma; COADREAD: colon adenocarcinoma/rectum adenocarcinoma; HNSC: head and neck squamous cell carcinoma; KICH: kidney chromophobe; KIRC: kidney renal clear cell carcinoma; KIRP: kidney renal papillary cell carcinoma; LAML: acute myeloid leukemia; LGG: low-grade glioma; LIHC: liver hepatocellular carcinoma; LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma; OV: ovarian serous cystadenocarcinoma; PAAD: pancreatic adenocarcinoma; PRAD: prostate adenocarcinoma; SKCM: skin cutaneous melanoma; STAD: stomach adenocarcinoma; THCA: thyroid carcinoma; UCEC: uterine corpus endometrial carcinoma.

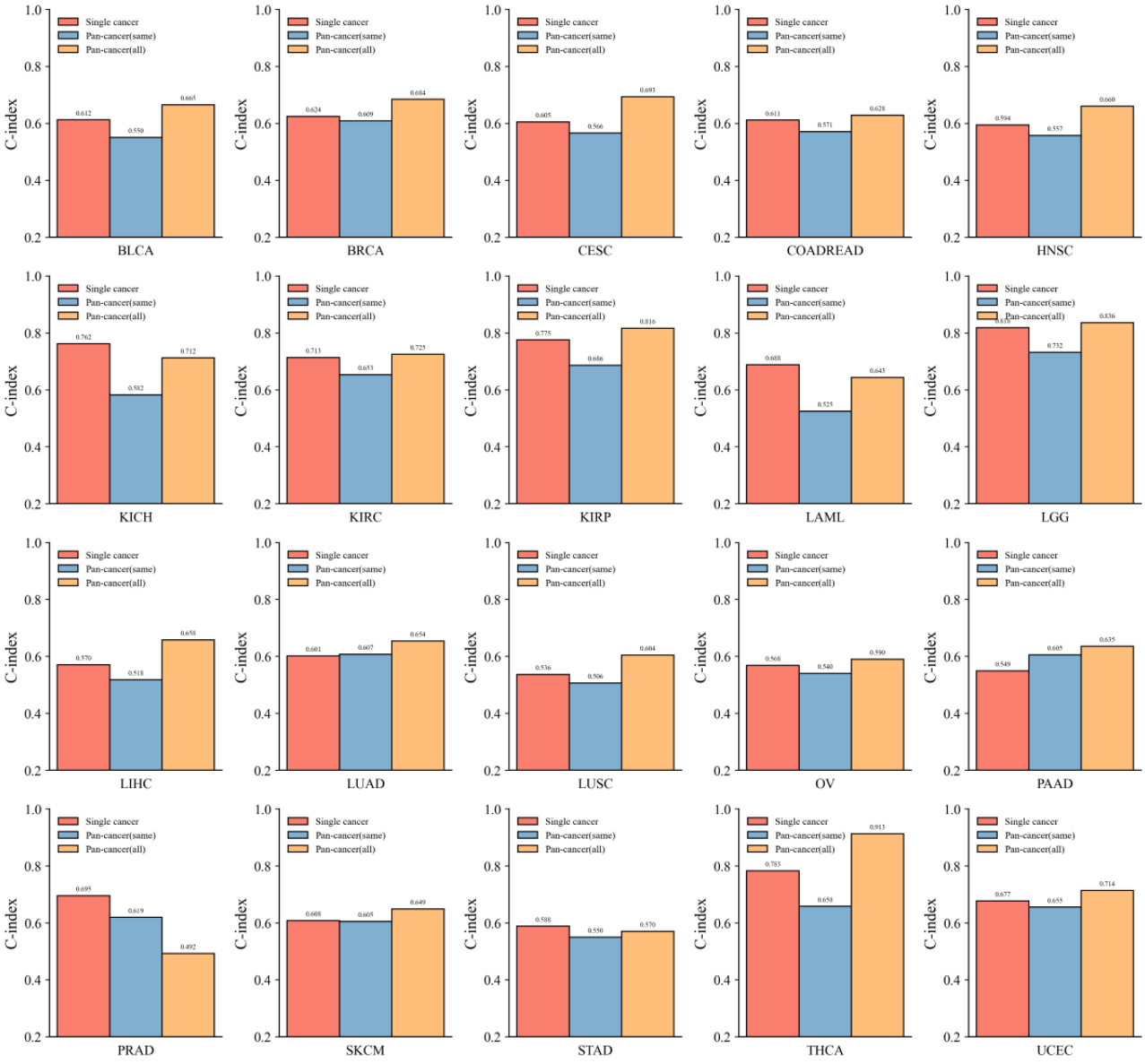
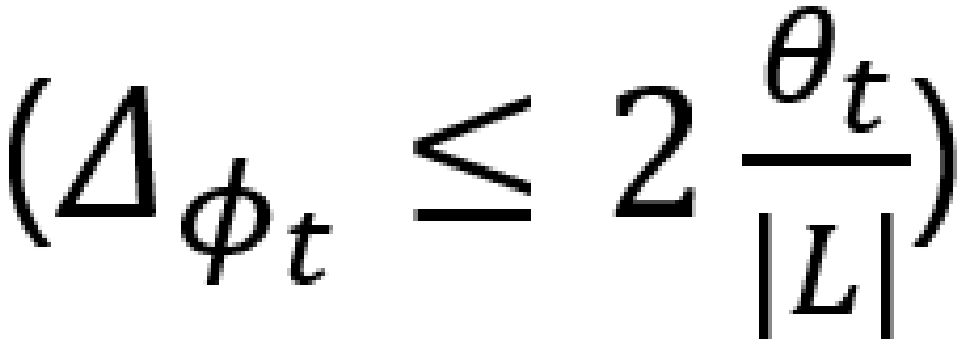


Figure 34. Inline graphic 27.



External Validation on Independent Cohorts

To address the critical question of generalizability beyond the TCGA dataset, we conducted an external validation on 2 independent cancer cohorts from the International Cancer Genome Consortium (ICGC) database [39]. Specifically, we selected the hepatocellular carcinoma (HCC; ICGC-HCC) and clear cell renal cell carcinoma (ccRCC; ICGC-ccRCC) projects, which provide clinical survival information and mRNA expression data compatible with our model's input requirements. To maintain consistency with our multimodal input configuration (clinical + mRNA + miRNA), the same zero-vector imputation strategy described in the Methods section ("Data Preprocessing") was applied to handle the missing miRNA modality and any mRNA features absent in the external data. The ICGC-HCC cohort comprised 242 patients with overall survival data, and the ICGC-ccRCC cohort comprised 530 patients. For each cohort, we preprocessed the mRNA expression data by aligning gene features to the 1579 mRNA features used in our TCGA training set. All continuous features were normalized using the minimum and maximum values derived from the TCGA training set to ensure consistent scaling.

This study applied our proposed model (trained on the pan-cancer TCGA dataset with clinical + mRNA + miRNA modalities) directly to the preprocessed ICGC cohorts without any retraining or fine-tuning. The model's survival prediction performance was evaluated using the C-index.

Our proposed model, trained on the pan-cancer TCGA dataset, achieved a C-index of 0.721 on the independent ICGC-HCC cohort and 0.708 on the ICGC-ccRCC cohort. This demonstrates that the model maintains its generalizable prognostic discriminatory power on unseen, external patient populations.

Discussion

Principal Results

In this study, we evaluated the model's classification performance in survival prediction across 4 data combinations—clinical data, mRNA, miRNA, and CNV—using the pan-cancer dataset. Compared with MultiSurv [17], the proposed model exhibits an approximately 0.3 improvement in prediction accuracy across all modality combinations after noise injection. When compared with Fan et al [10], the proposed model achieves higher accuracy in all settings before noise addition, and maintains comparable accuracy after noise injection, despite the latter model not incorporating any privacy-preserving mechanisms.

The improvement in survival prediction across different modalities stems from the model's ability to capture key information from different subspaces in parallel, thereby overcoming the representational limitations of single-attention mechanisms. By effectively integrating core features from each modality, the model enhances the flexibility and expressiveness of the fusion process. Notably, the inclusion of the CNV modality generally leads to a decline in predictive performance, suggesting that CNV data contain relatively fewer informative features relevant to patients' survival. This further underscores the value of multimodal data fusion in enhancing the generalizability of survival prediction models.

Comparisons in single-cancer survival prediction further demonstrate that the model achieves higher and more stable C-index values in 18 out of 20 cancer types compared with existing deep learning methods, highlighting its robustness in individualized survival prediction. Our study not only confirms the effectiveness and superiority of the proposed model in individualized cancer survival prediction but also demonstrates its ability to preserve patients' privacy. Collectively, these results provide a more reliable and privacy-conscious solution for precision oncology applications.

In addition, experiments conducted on both pan-cancer and single-cancer datasets show that the proposed model maintains strong predictive capability across diverse data settings, demonstrating robust generalization ability.

Challenges

While the model demonstrates strong privacy-preserving capabilities through DP, the integration of DP mechanisms into clinical artificial intelligence models introduces certain challenges. First, DP may impact the interpretability of the model, as the noise added to the gradients or model parameters can obscure the relationships between input features and predicted outcomes. This could reduce the transparency of the model, making it more difficult for clinicians to trust and understand the model's decisions, which is critical in medical applications. In addition, DP mechanisms may complicate auditability, as it becomes harder to trace the influence of individual data points on the model's predictions due to the noise injected for privacy protection.

Second, the integration of privacy mechanisms such as DP into clinical workflows or federated research environments can be practically challenging. In clinical settings, where data are highly sensitive, ensuring that privacy is preserved while maintaining the usability and accuracy of predictive models requires careful consideration of computational resources and model deployment strategies. In federated research environments, where data are distributed across multiple institutions, ensuring that DP mechanisms are applied consistently across all sites without compromising the model's predictive performance can be technically difficult. Moreover, adapting these privacy-preserving mechanisms to existing clinical systems and ensuring compliance with data protection regulations (eg, General Data Protection Regulation requires close collaboration with regulatory bodies and careful planning of data-sharing protocols.

These considerations highlight the need for further research into improving the balance between privacy protection, model interpretability, and practical integration in clinical and research settings.

Limitations

The proposed model has certain limitations. First, while we have demonstrated the model's effectiveness on the TCGA pan-cancer dataset and provided initial evidence of generalizability through external validation on ICGC cohorts, the performance on independent datasets shows a nonnegligible decline. This underscores the inherent challenge of translating models trained on 1 cohort (even a large and diverse one like TCGA) to others due to population, treatment, and technical variability. More extensive validation across multiple, prospectively collected cohorts is needed to fully establish clinical readiness. For example, training efficiency requires improvement. Analysis suggests that one possible reason is the introduction of the bilayer feature extraction module based on MHA and the adaptive DP protection mechanism, which increases algorithmic computational complexity. Additionally, the model exhibits increased inference time compared with simpler architectures such as CPH or RSF, due to the MHA mechanism and the privacy-preserving noise injection process. This may limit its deployment in real-time clinical settings where rapid predictions are required. Meanwhile, comparative experiments with other advanced models on larger-scale multimodal datasets are necessary. Indeed, in the era of big data and multimodal data fusion, new challenges lie ahead.

Future Work

Future work will expand in several directions. On the one hand, integrating pathological images and proteomics data will enhance multimodal feature fusion, extracting and fusing more information to further improve prediction performance. On the other hand, the model will be applied to additional tasks such as cancer subtype classification and biomarker discovery. Finally, research on other noise mechanisms based on DP is needed to protect sensitive information from multiple perspectives and ensure the security of model training and sharing. Moreover, efforts will be made to optimize the computational efficiency and reduce inference time through techniques such as model pruning, quantization, or lightweight attention mechanisms, thereby enhancing the model's practical applicability in clinical environments. To address potential limitations of zero-vector imputation, future work will also explore advanced missing-modality handling strategies, including masking strategies and missing-modality–aware attention mechanisms.

Conclusions

To address the key challenges of inefficient multimodal integration, insufficient privacy protection, and limited generalizability in pan-cancer survival prediction, this study proposes a bilayer feature fusion model integrating the MHA mechanism and adaptive DP. Innovatively, the framework uses modality-specific FC networks in the first layer to extract features from clinical, mRNA, miRNA, and CNV data. The second layer uses MHA to model cross-modal interactions through parallel attention heads, dynamically allocating weights to focus on survival-related features and overcoming the limitations of static fusion. Additionally, layer-wise relevance analysis is used during first-layer feature extraction to quantify the correlation between features and outcomes, enabling adaptive DP—injecting less Laplacian noise into the gradients of high-correlation features and more noise into low-correlation features to balance privacy and utility. Experimental results on pan-cancer datasets from TCGA demonstrate the feasibility and superiority of our proposed method. Furthermore, external validation on independent ICGC cohorts provided preliminary evidence of the model's generalizability, showing retained predictive power superior to clinical-only models despite expected performance attenuation due to cohort heterogeneity.

Acknowledgments

The authors sincerely thank all colleagues and collaborators who provided helpful discussions and valuable suggestions during the preparation of this study. In addition, the authors note that XW and HW contributed equally as co-corresponding authors.

Funding

This work was supported by the Yunlong Lake Laboratory of Deep Underground Science and Engineering Project under grant 104024005, Jiangsu Provincial Innovation Capacity Building Program under grants BM2022009 and BM2023017, the Integrated Traditional Chinese and Western Medicine Chronic Disease Management Research Project under grant CXZH2024088, and the Serving the “343” Industrial Development Project for Universities in Xuzhou under grant gx2024017.

Data Availability

The data that support the findings of this study are openly available in the Xena database [40] and the TCGA database [41].

Authors' Contributions

Conceptualization: YC
Data curation: ZD
Formal analysis: YC (lead), LW (supporting)
Funding acquisition: XW
Investigation: ZD (lead), LW (supporting)
Methodology: XW
Project administration: XW
Resources: XW
Supervision: LW (lead), HW (supporting)
Validation: YC (lead), HW (supporting)
Visualization: YC (lead), ZD (supporting)
Writing—original draft: YC (lead), ZD (supporting)
Writing—review & editing: XW (lead), HW (supporting)

Conflicts of Interest

None declared.

References

1. Vale-Silva LA, Rohr K. Long-term cancer survival prediction using multimodal deep learning. Sci Rep. 2021;11(1):13505. [FREE Full text] [doi: 10.1038/s41598-021-92799-4] [Medline: 34188098]
2. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann. Appl. Stat. 2008;2(3):841-860. [doi: 10.1214/08-AOAS169]
3. Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological). 2018;34(2):187-202. [doi: 10.1111/j.2517-6161.1972.tb00899.x]
4. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol. 2018;18(1):24. [FREE Full text] [doi: 10.1186/s12874-018-0482-1] [Medline: 29482517]
5. Kim DW, Lee S, Kwon S, Nam W, Cha I, Kim HJ. Deep learning-based survival prediction of oral cancer patients. Sci Rep. 2019;9(1):6994. [FREE Full text] [doi: 10.1038/s41598-019-43372-7] [Medline: 31061433]
6. Wulczyn E, Steiner DF, Moran M, Plass M, Reihs R, Tan F, et al. Interpretable survival prediction for colorectal cancer using deep learning. NPJ Digit Med. 2021;4(1):71. [FREE Full text] [doi: 10.1038/s41746-021-00427-2] [Medline: 33875798]
7. Wulczyn E, Steiner DF, Xu Z, Sadhwani A, Wang H, Flament-Auvigne I, et al. Deep learning-based survival prediction for multiple cancer types using histopathology images. PLoS One. 2020;15(6):e0233678. [FREE Full text] [doi: 10.1371/journal.pone.0233678] [Medline: 32555646]
8. Yeoh PSQ, Lai KW, Goh SL, Hasikin K, Wu X, Li P. Transfer learning-assisted 3D deep learning models for knee osteoarthritis detection: Data from the osteoarthritis initiative. Front Bioeng Biotechnol. 2023;11:1164655. [FREE Full text] [doi: 10.3389/fbioe.2023.1164655] [Medline: 37122858]
9. Cheerla A, Gevaert O. Deep learning with multimodal representation for pancancer prognosis prediction. Bioinformatics. 2019;35(14):i446-i454. [FREE Full text] [doi: 10.1093/bioinformatics/btz342] [Medline: 31510656]
10. Fan Z, Jiang Z, Liang H, Han C. Pancancer survival prediction using a deep learning architecture with multimodal representation and integration. Bioinform Adv. 2023;3(1):vbad006. [FREE Full text] [doi: 10.1093/bioadv/vbad006] [Medline: 36845202]

11. Guo W, Liang W, Deng Q, Zou X. A multimodal affinity fusion network for predicting the survival of breast cancer patients. Front Genet. 2021;12:709027. [FREE Full text] [doi: 10.3389/fgene.2021.709027] [Medline: 34490038]
12. Tan K, Huang W, Liu X, Hu J, Dong S. A multi-modal fusion framework based on multi-task correlation learning for cancer prognosis prediction. Artif Intell Med. 2022;126:102260. [doi: 10.1016/j.artmed.2022.102260] [Medline: 35346442]
13. Na H, Wang L, Zhuang X. Attention-based multimodal bilinear feature fusion for lung cancer survival analysis. IEEE; 2023. Presented at: IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE); 2023 December 04-06:219-225; Dayton, OH, USA. [doi: 10.1109/bibe60311.2023.00042]
14. Li R, Wu X, Li A, Wang M. HFBSurv: Hierarchical multimodal fusion with factorized bilinear models for cancer survival prediction. Bioinformatics. 2022;38(9):2587-2594. [FREE Full text] [doi: 10.1093/bioinformatics/btac113] [Medline: 35188177]
15. Lee C, Yoon J, Schaar MVD. Dynamic-deepHit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. IEEE Trans Biomed Eng. 2020;67(1):122-133. [doi: 10.1109/TBME.2019.2909027] [Medline: 30951460]
16. Kim T, Cho H, Yoon KJ. Cmta: Cross-modal temporal alignment for event-guided video deblurring. Cham. Springer Nature Switzerland; 2024. Presented at: European Conference on Computer Vision; 2024 September 29–October 4:1-19; Milan, Italy. [doi: 10.1007/978-3-031-72943-0_1]
17. Silva LAV, Rohr K. Pan-cancer prognosis prediction using multimodal deep learning. 2020. Presented at: IEEE 17th International Symposium on Biomedical Imaging (ISBI); 2020 April 3-7:568-571; Iowa City, IA, USA. [doi: 10.1109/isbi45749.2020.9098665]
18. Zhang Z, Yin W, Wang S, Zheng X, Dong S. MBFusion: Multi-modal balanced fusion and multi-task learning for cancer diagnosis and prognosis. Comput Biol Med. 2024;181:109042. [doi: 10.1016/j.compbiomed.2024.109042] [Medline: 39180856]
19. Wang C, Guo J, Zhao N, Liu Y, Liu X, Liu G, et al. A cancer survival prediction method based on graph convolutional network. IEEE Trans Nanobioscience. 2020;19(1):117-126. [doi: 10.1109/TNB.2019.2936398] [Medline: 31443039]
20. Flack D, Tripathi A, Waqas A, Rasool G, Dera D. Robust multimodal fusion for survival prediction in cancer patients. Cancer Inform. 2025;24:11769351251376192. [FREE Full text] [doi: 10.1177/11769351251376192] [Medline: 41024938]
21. Venkatesaramani R, Malin BA, Vorobeychik Y. Re-identification of individuals in genomic datasets using public face images. Sci Adv. 2021;7(47):eabg3296. [FREE Full text] [doi: 10.1126/sciadv.abg3296] [Medline: 34788101]
22. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell. 2020;2(6):305-311. [doi: 10.1038/s42256-020-0186-1]
23. Chen H, Wang N, Zhou Y, Mei K, Tang M, Cai G. Breast cancer prediction based on differential privacy and logistic regression optimization model. Applied Sciences. 2023;13(19):10755. [doi: 10.3390/app131910755]
24. Chai H, Huang Y, Xu L, Song X, He M, Wang Q. A decentralized federated learning-based cancer survival prediction method with privacy protection. Heliyon. 2024;10(11):e31873. [FREE Full text] [doi: 10.1016/j.heliyon.2024.e31873] [Medline: 38845954]
25. Wu X, Wei Y, Mao Y, Wang L. A differential privacy DNA motif finding method based on closed frequent patterns. Cluster Comput. 2018;22(S2):2907-2919. [doi: 10.1007/s10586-017-1691-9]
26. Wu X, Zhang Y, Shi M, Li P, Li R, Xiong NN. An adaptive federated learning scheme with differential privacy preserving. Future Generation Computer Systems. 2022;127:362-372. [doi: 10.1016/j.future.2021.09.015]
27. Wang H, Zhang X, Xia Y, Wu X. An intelligent blockchain-based access control framework with federated learning for genome-wide association studies. Computer Standards & Interfaces. 2023;84:103694. [doi: 10.1016/j.csi.2022.103694]
28. Wang H, Zhang X, Xia Y, Wu X. An intelligent blockchain-based access control framework with federated learning for genome-wide association studies. Computer Standards & Interfaces. 2023;84:103694. [doi: 10.1016/j.csi.2022.103694]
29. Wang H, Wu X. IPP: An intelligent privacy-preserving scheme for detecting interactions in genome association studies. IEEE/ACM Trans Comput Biol Bioinform. 2023;20(1):455-464. [doi: 10.1109/TCBB.2022.3155774] [Medline: 35239492]
30. Al Fatih Abil Fida M, Ahmad T, Ntahobari M. Variance threshold as early screening to boruta feature selection for intrusion detection system. 2021. Presented at: 13th International Conference on Information & Communication Technology and System (ICTS); 2021 October 19-21:46-50; Surabaya, Indonesia. [doi: 10.1109/icts52701.2021.9608852]
31. Gajera V, Gupta R, Jana PK. An effective multi-objective task scheduling algorithm using min-max normalization in cloud computing. 2016. Presented at: 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT); 2016 July 21-23:812-816; SJB Institute of Technology, Bengaluru, Karnataka, India. [doi: 10.1109/icatcct.2016.7912111]
32. Hutter C, Zenklusen JC. The cancer genome atlas: Creating lasting value beyond its data. Cell. 2018;173(2):283-285. [FREE Full text] [doi: 10.1016/j.cell.2018.03.042] [Medline: 29625045]
33. Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, Cancer Genome Atlas Research Network, et al. Machine learning identifies stemness features associated with oncogenic dedifferentiation. Cell. 2018;173(2):338-354.e15. [FREE Full text] [doi: 10.1016/j.cell.2018.03.034] [Medline: 29625051]
34. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167. 2015. [doi: 10.5860/choice.189890]

35. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA. 1982;247(18):2543-2546. [Medline: 7069920]
36. Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K. Deep learning with differential privacy. 2016. Presented at: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security; 2016 October 24 - 28:308-318; Vienna Austria. [doi: 10.1145/2976749.2978318]
37. Tang Q, Lécuyer M. DP-Adam: Correcting DP bias in adam's second moment estimation. arXiv:2304.11208. 2023. [doi: 10.48550/arXiv.2304.11208]
38. Papernot N, Abadi M, Erlingsson U, Goodfellow I, Talwar K. Semi-supervised knowledge transfer for deep learning from private training data. arXiv:1610.05755. 2016. [doi: 10.48550/arXiv.1610.05755]
39. ICGC ARGO. URL: https://www.icgc-argo.org/ [accessed 2026-03-06]
40. UCSC Xena. URL: https://xenabrowser.net/ [accessed 2026-03-06]
41. GDC Data Portal. URL: https://portal.gdc.cancer.gov/ [accessed 2026-03-06]

Abbreviations

ccRCC: clear cell renal cell carcinoma
C-index: concordance index
CMTA: cross-modal translation and alignment
CNV: gene copy number variation
CPH: Cox proportional hazards model
DP: differential privacy
FC: fully connected
HCC: hepatocellular carcinoma
ICGC: International Cancer Genome Consortium
MHA: multihead attention
miRNA: microRNA
mRNA: messenger RNA
PATE: Private Aggregation of Teacher Ensembles
RSF: random survival forest
SGD: Stochastic Gradient Descent
TCGA: The Cancer Genome Atlas

Edited by J Klann; submitted 08.Sep.2025; peer-reviewed by Y Yang, F Amakye, D Dera, J Zhu; comments to author 11.Dec.2025; accepted 09.Feb.2026; published 30.Mar.2026

Please cite as:
Chen Y, Deng Z, Wang L, Wang H, Wu X
A Bilayer Feature Fusion Framework for Pan-Cancer Survival Prediction Based on Multihead Attention and Adaptive Differential Privacy: Model Development and Validation Study
JMIR Med Inform 2026;14:e83743
URL: https://medinform.jmir.org/2026/1/e83743
doi: 10.2196/83743
PMID: