doi: N/A
Paper

PAPER

Masked Multimodal Autoencoders with Contrastive Learning for Cancer Survival Representation Learning

Yancheng Liu,[1,*] David Ning[1] and Yang Xiang[1]

[1]Department Of Computer Science, Brown University, 02912, RI, USA
[*]Corresponding author. yancheng_liu@brown.edu

Abstract

Accurate survival prediction is critical for guiding cancer treatment decisions. While deep learning models have advanced this field, many depend on large, high-dimensional datasets like TCGA, limiting their accessibility. Emerging mutation profiling platforms offer binary gene mutation data, but their sparsity challenges traditional approaches. We propose a masked multimodal autoencoder with contrastive learning that encodes gene mutation profiles and clinical data into compact representations. The model captures intra- and inter-modal relations using modality-specific encoders and a learnable fusion mechanism. Contrastive learning objectives enhance robustness, and the resulting embeddings are used in a downstream survival model to perform survival analysis. Tested on a pan-cancer dataset of 22,284 patients from MSK, our method achieves an average c-index of 0.744 using far fewer features than existing models. Integrated gradient analysis further reveals biologically meaningful gene and clinical signal contributions. This framework provides a scalable, interpretable, and data-efficient approach for pan-cancer survival prediction.

Introduction

Accurate stratification of cancer patients by survival risk is essential for guiding therapeutic decisions. The growing availability of molecular and clinical data has inspired deep-learning approaches that jointly model heterogeneous modalities to improve prognostic accuracy. However, most state-of-the-art methods depend on large, high-dimensional datasets—e.g., RNA-seq profiles from TCGA—and extensive hand-crafted feature engineering, which limits reproducibility and clinical translation.

Meanwhile, targeted mutation panels such as Foundation One [10] have made it possible to obtain standardized binary profiles of gene alterations across diverse cancers. Yet these datasets are often extremely sparse—most patients harbor only a few mutations among hundreds of tested genes—and frequently include missing clinical and genomic variables. This sparsity and incompleteness challenge conventional deep learning models, leading to overfitting, reduced interpretability, and poor generalization across cancer types and patient populations.

To address these limitations, we introduce Surv-M^{2}AC, a **M**asked **M**ultimodal **A**utoencoder with **C**ontrastive. It jointly encodes gene mutation profiles and clinical features into compact latent embeddings by explicitly modeling missingness via stochastic input masking and promoting identity-preserving representations through contrastive regularization. By reducing reliance on large-scale training data and manual feature design, our approach offers a scalable and generalizable solution for pan-cancer survival prediction using sparse, real-world clinical-genomic inputs.

Related Work

Various deep learning approaches have been proposed to improve survival prediction beyond traditional methods like Cox regression and random forests. Architectures based on MLPs, CNNs, RNNs, and Transformers have been explored for their ability to model spatial, sequential, or structured dependencies in high-dimensional biological data [8, 16, 9, 15]. Autoencoders, in particular, have shown promise in cancer survival analysis for their ability to process multimodal inputs and learn compact, informative latent representations. Franco et al.'s work showed that different autoencoder structures can significantly impact downstream prognostic accuracy [3].

Recent models further highlight this trend. Ellen et al. used a denoising autoencoder combined with multitask prediction and achieved a C-index of 0.69 on lung cancer [2], but their model requires massive input dimensionality ($>$ 500K features). AUTOSurv improves interpretability by using a pathway-guided VAE [7], while SELECTOR adopts a masked autoencoder to address missing-modality issues [11]. Both models report peak C-index scores ranging from 0.75 to 0.79.

These studies underscore the potential of autoencoders for survival modeling, particularly in their ability to learn biologically meaningful, low-dimensional embeddings. Building on this foundation, we introduce Surv-M^{2}AC, which integrates and extends key ideas from AUTOSurv and SELECTOR with the following contributions:

- **Pan-cancer generalization**: Surv-M^{2}AC is designed to operate across multiple cancer types, enabling broader applicability beyond single-disease models.

- **Data efficiency**: Surv-M^{2}AC achieves strong performance with much smaller, lower-dimensional binarized datasets.
- **Architectural innovation**: Uses a masked autoencoder with modality-specific encoders and inter-modal fusion.
- **Contrastive regularization**: Adds a contrastive loss to enhance robustness under missing data and promote patient identity consistency.

Method

Dataset And Feature Engineering

We use data from Memorial Sloan Kettering Cancer Center, covering five major cancer types: non-small cell lung, colorectal, breast, prostate, and pancreatic cancer [6]. The dataset includes targeted sequencing data across 24,950 patients, with matched gene mutation samples from the MSK-IMPACT assay and detailed clinical annotations.

We filtered for patients with unique sample IDs and retained 16 clinical features relevant to survival analysis, including age at diagnosis, cancer type, cancer stage, metastatic indicators, and key biomarkers such as Tumor mutational burden (TMB). Features with high missingness, such as MSI Type and Prior Treatment, were excluded. Survival status and survival time (in months) served as labels for survival analysis modeling. Except for the targets, all features were binarized based on clinical criteria to simplify the input space.

For the gene mutation profiles, we defined mutation subtypes based on gene name, mutation type, variant type, and chromosome number, yielding 2,795 unique subtypes. Each subtype was binarized, with 1 indicating the presence of a specific mutation in a patient and 0 indicating its absence. Subtypes found in fewer than 10 patients (less than 0.05% of the cohort) were removed, leaving 1,165 retained subtypes. The overall structure of the dataset can be seen in Figure 5 (see Supplementary Information). After excluding patients with missing labels, the final cohort comprised 22,284 patients.

Surv-M^{2}AC Architecture

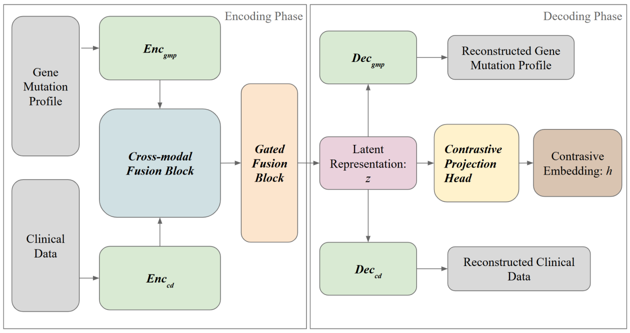


Fig. 1. Overview of the Surv-M^{2}AC Architecture. Modality-specific encoders extract intra-modal features, which are fused via a cross-modal block. The latent representation z is used for both masked reconstruction and contrastive supervision.

As a baseline, we initially followed the DeepSurv approach and trained a simple two-layer MLP with dropout on concatenated gene mutation and clinical features [8]. However, as shown in Figure 6 (see Supplement Information), this model rapidly overfits: while the training loss steadily decreases, the validation loss increases, and survival prediction metrics such as the concordance index (c-index) deteriorate. The final mean c-index is 0.706. This overfitting is likely due to the extreme sparsity of the data—on average, each patient exhibits mutations in only approximately 13 out of 1,165 possible genes.

This result highlights a critical limitation of standard survival models when applied to sparse, high-dimensional mutation data, motivating the need for more robust representation learning strategies. To address this, we introduce **Surv-M^{2}AC** (**Surv**ival **M**asked **M**ultimodal **A**utoencoder with **C**ontrastive learning), designed to learn compact, informative, and interpretable patient embeddings while reducing overfitting and improving generalization.

As shown in Figure 1, Surv-M^{2}AC comprises three key components: (1) modality-specific encoders to capture intra-modality structures, (2) fusion blocks for modeling cross-modality interactions, and (3) dual decoders for reconstructing masked inputs. Additionally, a contrastive learning objective enforces consistency across different corruption views, improving robustness under input sparsity.

Data Encoding: Modality-Specific Masked Autoencoder

Each modality—gene mutation profiles (GMP) and clinical data (CD)—is processed by a dedicated encoder to capture intra-modal structures and preserve modality-specific information before cross-modal fusion.

Specifically, the gene mutation encoder $\mathrm{Enc}_{\mathrm{gmp}}$ projects the input $x_{\mathrm{gmp}} \in \mathbb{R}^{B \times d_{\mathrm{gmp}}}$ into a hidden space of dimension d_{hidden} using a deeper architecture composed of two residual blocks followed by a Squeeze-and-Excitation (SE) block [4, 5]. This design enables expressive local transformations while emphasizing informative feature channels. In contrast, the clinical data encoder $\mathrm{Enc}_{\mathrm{cd}}$ adopts a lighter configuration with a single residual block and an SE block to prevent overfitting. The resulting embeddings are:

$$h_{\mathrm{gmp}} = \mathrm{Enc}_{\mathrm{gmp}}(x_{\mathrm{gmp}}), \quad h_{\mathrm{cd}} = \mathrm{Enc}_{\mathrm{cd}}(x_{\mathrm{cd}}).$$

This asymmetric encoder design reflects the differing complexity of each modality: gene mutations are high-dimensional and sparse, demanding deeper transformations, whereas clinical features are compact and semantically rich, benefiting from simpler models. Adapting encoder depth to each modality improves representation quality and efficiency, enhancing downstream fusion and predictive performance.

Multimodal Fusion: Bi-FiLM

While modality-specific encoders extract rich intra-modal representations, predictive performance hinges on capturing how mutation patterns condition clinical outcomes—and vice versa. To model cross-modal interactions between gene mutations and clinical features efficiently, we employ **Bi**directional **F**eature-w**i**se **L**inear **M**odulation (Bi-FiLM) [12].

Concretely, each modality embedding is first normalized:

$$\hat{h}_{\mathrm{gmp}} = \mathrm{LN}(h_{\mathrm{gmp}}), \quad \hat{h}_{\mathrm{cd}} = \mathrm{LN}(h_{\mathrm{cd}}).$$

Then, FiLM parameters $(\gamma_{\mathrm{cd}}, \beta_{\mathrm{cd}})$ are learned from $\hat{h}_{\mathrm{gmp}}$ to modulate clinical features, and vice versa:

$$(\gamma_{\mathrm{gmp}}, \beta_{\mathrm{gmp}}) = W_{\mathrm{cd}}(\hat{h}_{\mathrm{cd}}), \quad (\gamma_{\mathrm{cd}}, \beta_{\mathrm{cd}}) = W_{\mathrm{gmp}}(\hat{h}_{\mathrm{gmp}}).$$

where W_{cd} and W_{gmp} are modality-specific linear layers. The bidirectional FiLM modulation is applied as:

$$h'_{\mathrm{gmp}} = \gamma_{\mathrm{gmp}} \odot \hat{h}_{\mathrm{gmp}} + \beta_{\mathrm{gmp}}, \quad h'_{\mathrm{cd}} = \gamma_{\mathrm{cd}} \odot \hat{h}_{\mathrm{cd}} + \beta_{\mathrm{cd}}.$$

The modulated features are concatenated and passed through a lightweight projection layer to obtain the final shared latent representation:

$$z = W_{\text{proj}}(h'_{\text{gmp}} \oplus h'_{\text{cd}}) \in \mathbb{R}^{B \times d_{\text{latent}}}.$$

The Bi-FiLM mechanism enhances cross-modal fusion by enabling mutual conditioning between modalities, leading to more informative and robust joint embeddings.

Reconstruction Decoders and Contrastive Projection Head

From the shared latent embedding z, two independent decoders reconstruct the original gene mutation and clinical inputs, encouraging the model to retain modality-specific information and improving robustness to feature masking. Each decoder is a lightweight two-layer MLP with GELU activations, designed to keep reconstruction simple and shift representational complexity onto the encoders.

Additionally, a separate prediction head projects z into a contrastive embedding space, with outputs l_2-normalized to stabilize contrastive loss and maintain consistent similarity scaling across batches.

Masked Reconstruction with Contrastive Learning

While masked reconstruction encourages the model to recover local feature patterns, it does not guarantee that the latent embeddings of the same patient remain close under different input corruptions. To enforce identity-preserving representations and improve robustness to missing data, we augment the reconstruction loss with a contrastive regularizer inspired by SimCLR [1].

For each patient, we apply two independent random masks to the gene mutation and clinical inputs, creating two corrupted views. Surv-M^{2}AC processes each masked view separately, producing reconstructed inputs and latent representations. The reconstruction loss is computed independently for each view by comparing the masked predictions to the original inputs, using only the visible (unmasked) features based on Equation 1.

$$\mathcal{L}_{\text{recon}} = \frac{\sum_{i=1}^{d}(1 - m_i) \cdot \text{BCEWithLogits}(\hat{x}_i, x_i)}{\sum_{i=1}^{d}(1 - m_i) + \epsilon} \tag{1}$$

where m_i denotes the mask indicator for feature i, and ϵ is a small constant for numerical stability. Reconstruction losses from the gene mutation and clinical modalities are summed for each view and averaged across the two views to obtain the overall reconstruction loss.

In parallel, we apply a contrastive loss to the latent embeddings generated from the two masked views. Specifically, the two masked views of the same patient are treated as positive pairs, while embeddings from other patients in the batch serve as negatives. We encourage representations of the same patient under different corruptions to remain close, while pushing apart those from different patients, following the InfoNCE framework [14].

The total training objective combines the reconstruction loss and the contrastive loss, weighted by a coefficient β that balances the two terms as

$$\mathcal{L} = \frac{1}{2} \cdot (\mathcal{L}^{(1)}_{\text{recon}} + \mathcal{L}^{(2)}_{\text{recon}}) + \beta \cdot \mathcal{L}_{\text{cont}}.$$

This contrastive-augmented strategy encourages the model to learn consistent and semantically meaningful patient representations, improving robustness to missing or corrupted features and leading to more stable training dynamics.

Downstream Survival Prediction

To assess the quality of the learned patient embeddings, we train a lightweight survival prediction model on top of the frozen representations. Specifically, we fine-tune a two-layer MLP following the DeepSurv architecture, using the same loss function and training setup that previously suffered from overfitting. Model performance is evaluated on a held-out test set by computing the concordance index (C-index) and plotting Kaplan–Meier survival curves stratified by predicted risk groups. Improved predictive accuracy and more stable stratification would indicate that the learned embeddings effectively mitigate overfitting and enhance generalization.

Model Interpretability

To interpret the trained Surv-M^{2}AC model and its downstream survival predictor, we apply Integrated Gradients to the frozen patient embeddings and risk prediction network [13]. We build an integrated model that sequentially passes inputs through the encoder and the survival predictor. For each test sample, we compute feature attributions relative to the predicted risk score, using a zero baseline.

We highlight the top features influencing individual patient predictions and aggregate attributions across the test cohort to identify globally important features. This analysis provides insight into which variables most strongly drive survival predictions, supporting the biological interpretability of the model.

Results

C-index and Survival Prediction

The concordance index (C-index) is a standard metric for evaluating survival models, reflecting how well predicted risk scores align with actual survival outcomes. We compared several encoder variants and training objectives using a fixed two-layer MLP head. The models include: (1) no autoencoder (plain DeepSurv model), (2) a vanilla autoencoder, (3) Surv-M^{2}AC without cross-modal fusion, and (4) the full Surv-M^{2}AC architecture. We also tested training strategies with and without masking and contrastive loss. All models were trained 100 times per variant. Average C-index scores are summarized in Figure 2.

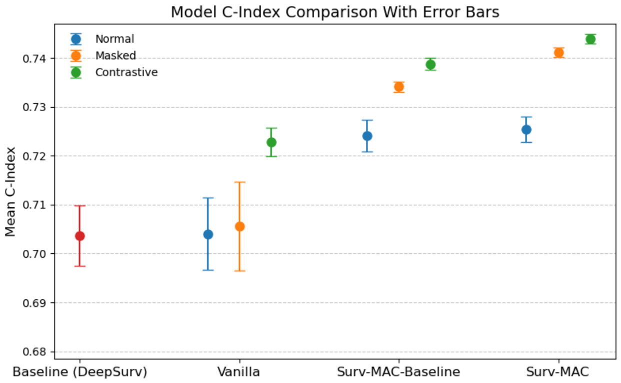


Fig. 2. Comparison of average C-index scores across encoder variants over 100 runs.

The results highlight the impact of architectural and training improvements. Even without masking or contrastive regularization, Surv-M^{2}AC outperforms the vanilla autoencoder by 3% (from 0.704 to 0.724). Adding both masking and contrastive objectives yields a final C-index of 0.744—a 5.8% overall gain.

In addition, Kaplan–Meier curves (Figure 7, see Supplementary Information) generated using Surv-M^{2}AC embeddings show clearer separation between high- and low-risk groups, with tighter confidence intervals. These results demonstrate that Surv-M^{2}AC produces more robust and informative representations for survival analysis, even under consistent downstream model configurations.

Interpretability of Surv-M^{2}AC

Figure 3 shows the most influential features identified by Integrated Gradients. Among them are well-established prognostic indicators such as metastatic sites (LIVER, LUNG, CNS BRAIN) and clinical staging variables (highest stage, Distant, Regional). Cancer types, including breast, colon, and pancreas, also appear frequently, indicating the model's ability to capture subtype-specific survival signals.

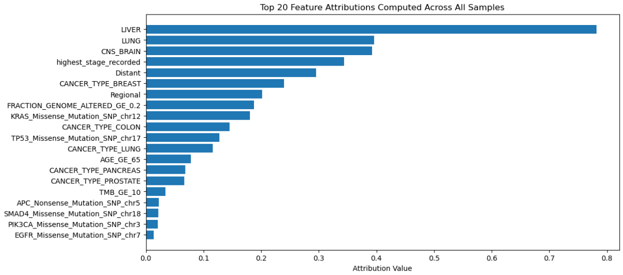


Fig. 3. Top features identified using Integrated Gradients across the dataset.

Crucially, Surv-M^{2}AC assigns high attribution scores to key somatic mutations (e.g., TP53, KRAS, EGFR, APC) and genomic instability markers (TMB, genome alteration fraction), aligning with known biomarkers and therapeutic targets. These results demonstrate that the model not only copes well with sparsity but also generates biologically grounded and interpretable predictions aligned with clinical knowledge.

Representation Structure

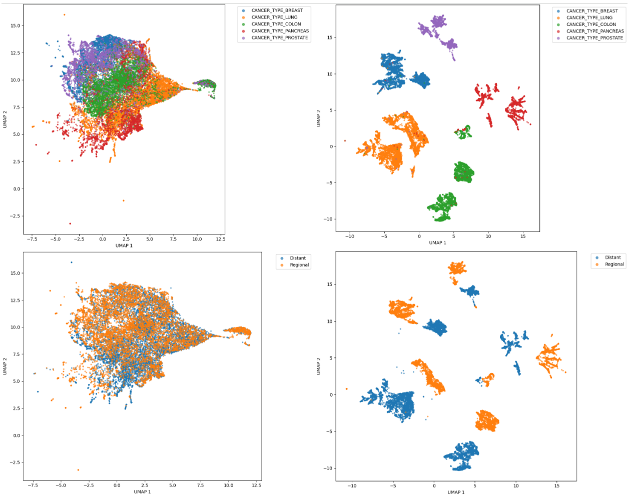


Fig. 4. UMAP visualizations of patient embeddings. Left: Vanilla autoencoder projections colored by cancer type (top) and metastasis spread (bottom). Right: Corresponding Surv-M^{2}AC embeddings show clearer separation and clinical coherence.

Figure 4 compares UMAP projections of latent spaces learned by a vanilla autoencoder and Surv-M^{2}AC, colored by cancer types and metastasis status. Surv-M^{2}AC produces more distinct and clinically meaningful clusters, demonstrating better alignment with both genomic and clinical signals.

Even without masking or contrastive loss, Surv-M^{2}AC produces more structured latent spaces than the baseline, underscoring the benefit of its modality-specific encoding and cross-modal fusion. These results confirm that Surv-M^{2}AC learns more discriminative and biologically meaningful patient representations.

Conclusion and Discussion

Our experiments demonstrate that Surv-M^{2}AC can produce meaningful latent representations, enabling a survival prediction model that achieves a mean C-index of 0.744. While this performance is slightly lower than that of AUTOSurv and SELECTOR, it is notable given our use of a simpler architectural backbone and a smaller, more accessible dataset. These results highlight the potential of lightweight designs for robust survival modeling in data-sparse clinical settings.

The model's strength comes from two core features: (1) modality-specific encoders and cross-modal fusion to capture intra- and inter-modal interactions, and (2) masked reconstruction with contrastive regularization to improve robustness under missing data.

Future work will explore more advanced autoencoder architectures, such as incorporating tokenization and attention mechanisms to enable fine-grained modeling of clinical and genomic features, capture long-range dependencies, and better handle irregular or missing inputs across modalities.

Contribution

Yancheng Liu led the conceptual development of the project, including dataset curation, feature engineering, and architectural design. He implemented baseline models such as the vanilla autoencoder and simple prediction heads, and developed the core Surv-M^{2}AC framework with multiple backbone variants and configurations. He also built the training and evaluation pipeline, conducting over 1,000 experiments for comparative analysis. In addition, Yancheng performed model interpretation using Integrated Gradients, generated UMAP visualizations, and identified clinically meaningful features. He authored the methods, results, and discussion sections across all drafts, and prepared the introduction, methodology, and results slides for the final presentation.

David Ning initially explored a diffusion-based approach for Hi-C data, though it was not pursued in the final project. For Surv-M^{2}AC, he contributed to the literature review and implemented additional downstream prediction heads (e.g., CNN1D and self-attention) to explore potential performance improvements. These alternatives did not yield significant gains but helped confirm the robustness of the current setup.

Yang Xiang supported the literature review, participated in model design discussions, assisted with data preprocessing, and contributed to interpretability testing. His input helped refine the architecture and guide the analysis.

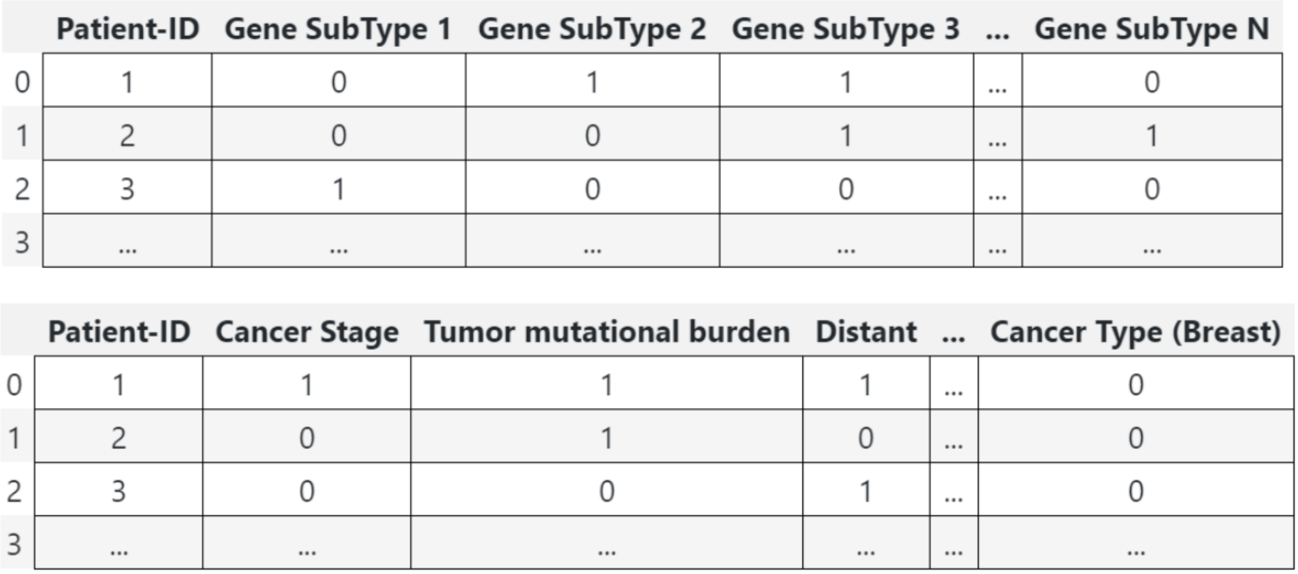
Code Availability

The code for Surv-M^{2}AC is available on GitHub at `https://github.com/Lyce24/Surv_MAC`. This repository includes the complete implementation and scripts needed to reproduce our results.

References

1. Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.
2. Jacob G. Ellen, Etai Jacob, Nikos Nikolaou, and Natasha Markuzon. Autoencoder-based multimodal prediction of non-small cell lung cancer survival. *Scientific Reports*, 13(1):15761, Sep 2023.
3. Edian F Franco, Pratip Rana, Aline Cruz, Víctor V Calderón, Vasco Azevedo, Rommel T J Ramos, and Preetam Ghosh. Performance comparison of deep learning autoencoders for cancer subtype detection using Multi-Omics data. *Cancers (Basel)*, 13(9), April 2021.
4. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.
5. Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks, 2019.
6. Justin Jee and Christopher et al Fong. Automated real-world data integration improves cancer outcome prediction. *Nature*, 636(8043):728–736, December 2024.
7. Lindong Jiang, Chao Xu, Yuntong Bai, Anqi Liu, Yun Gong, Yu-Ping Wang, and Hong-Wen Deng. Autosurv: interpretable deep learning framework for cancer survival analysis incorporating clinical and multi-omics data. *npj Precision Oncology*, 8(1):4, Jan 2024.
8. Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24, Feb 2018.
9. Changhee Lee, Jinsung Yoon, and Mihaela van der Schaar. Dynamic-DeepHit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Trans. Biomed. Eng.*, 67(1):122–133, January 2020.
10. Coren A Milbury, James Creeden, and Yip et al. Clinical and analytical validation of FoundationOne®CDx, a comprehensive genomic profiling assay for solid tumors. *PLoS One*, 17(3):e0264138, March 2022.
11. Liangrui Pan, Yijun Peng, Yan Li, Xiang Wang, Wenjuan Liu, Liwen Xu, Qingchun Liang, and Shaoliang Peng. Selector: Heterogeneous graph network with convolutional masked autoencoder for multimodal robust prediction of cancer survival, 2024.
12. Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer, 2017.
13. Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017.
14. Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.
15. Zifeng Wang and Jimeng Sun. Survtrace: transformers for survival analysis with competing events. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '22, page 1–9. ACM, August 2022.
16. Qingyan Yin, Wangwang Chen, Chunxia Zhang, and Zhi Wei. A convolutional neural network model for survival prediction based on prognosis-related cascaded wx feature selection. *Laboratory Investigation*, 102(10):1064–1074, 2022.

Supplementary Information



	Patient-ID	Gene SubType 1	Gene SubType 2	Gene SubType 3	...	Gene SubType N
0	1	0	1	1	...	0
1	2	0	0	1	...	1
2	3	1	0	0	...	0
3	...	...	...	...	...	...

	Patient-ID	Cancer Stage	Tumor mutational burden	Distant	...	Cancer Type (Breast)
0	1	1	1	1	...	0
1	2	0	1	0	...	0
2	3	0	0	1	...	0
3	...	...	...	...	...	...

Fig. 5. Final data format of gene mutation subtype profiles (upper) and clinical data (bottom) for patients.

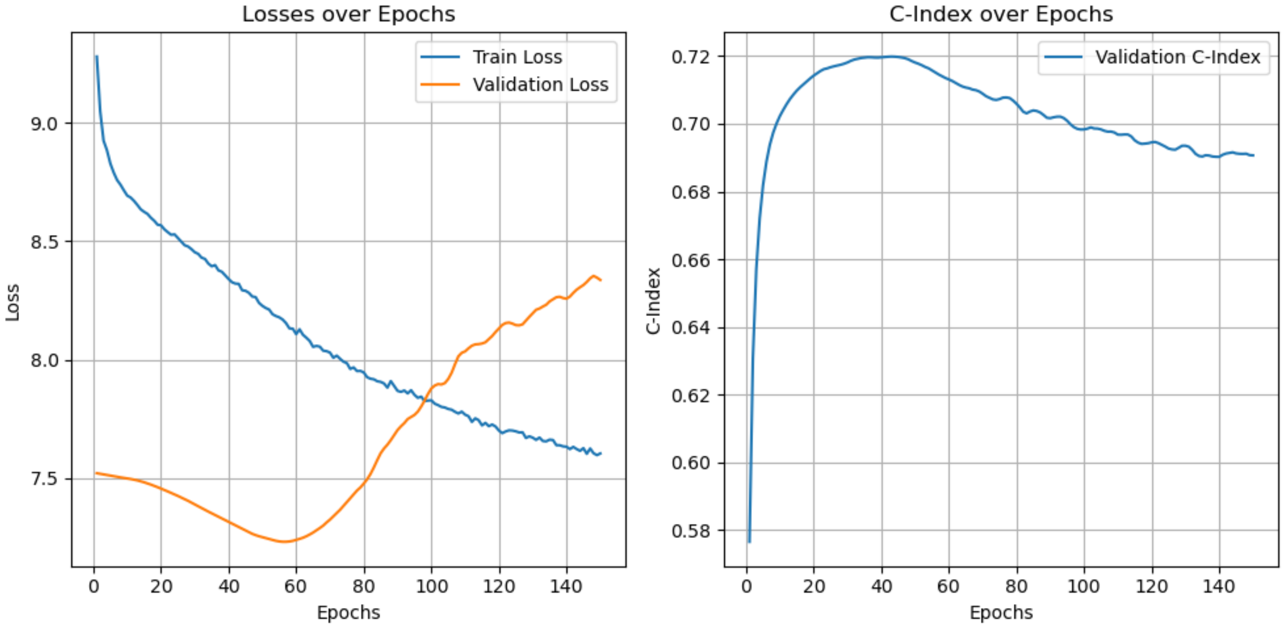


Fig. 6. Training and validation loss curves for DeepSurv based on our dataset.

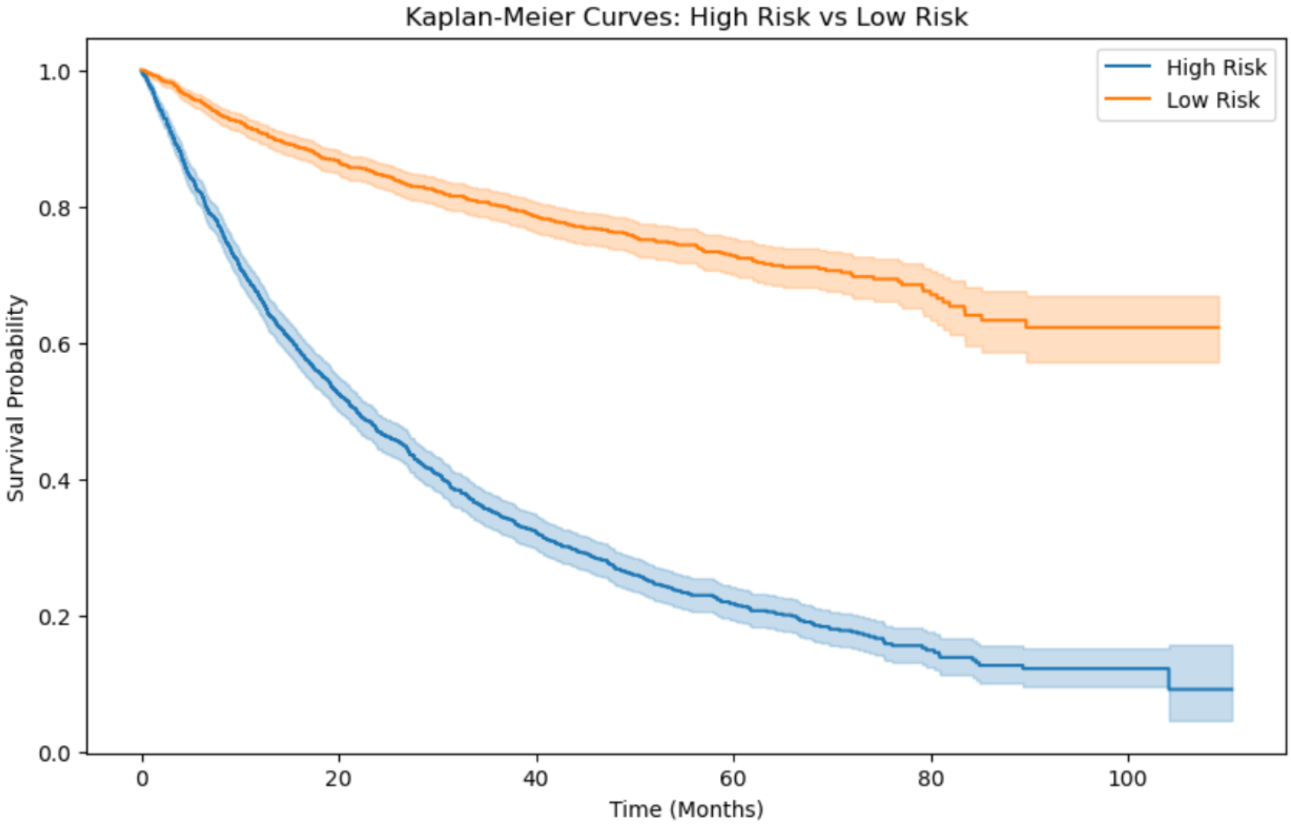


Fig. 7. The Kaplan–Meier curves generated using Surv-M^{2}AC