# MGCT: Mutual-Guided Cross-Modality Transformer for Survival Outcome Prediction using Integrative Histopathology-Genomic Features

Mingxin Liu[1], Yunzan Liu[1], Hui Cui[2], Chunquan Li[3,#], Jiquan Ma[1,#]

*Department of Computer Science and Technology, Heilongjiang University*
*Department of Computer Science and Information Technology, La Trobe University*
*The First Affiliated Hospital, Hengyang Medical School, University of South China*

# Background

I. Cancer ranks the leading cause of death worldwide and has become one of the five most common diseases in China and developing or developed countries.

II. In China, 55 people die of cancer in every 10 minutes.

III. There were an estimated 19,292,789 new cases and 9,958,133 cancer deaths worldwide in 2020. (excluding nonmelanoma, skin cancer, and basal cell carcinoma)

IV. In 2023, 1,958,310 new cancer cases and 609,820 cancer deaths are projected to occur in the United States.

V. Accurately diagnosing and prognosis the cancer is of paramount clinical importance.

[1] C. Xia et al. Cancer statistics in China and United States, 2022: profiles, trends, and determinants. Chinese Medical Journal, 2022.
[2] F. Bray et al. The ever-increasing importance of cancer as a leading cause of premature death worldwide. Cancer, 2021.
[3] H. Sung et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians, 2021.
[4] R. L. Siegel et al. Cancer statistics, 2023. CA: A Cancer Journal for Clinicians, 2023.

# Background

I. Survival analysis is a crucial topic in clinical research, which aims to predict the time elapsed from a known origin to an event of interest, such as death, relapse of disease, and development of an adverse reaction.

II. Traditionally, survival analysis relies on short term clinical indicators and long-term follow-up reports which are time-consuming and impractical in clinical applications.

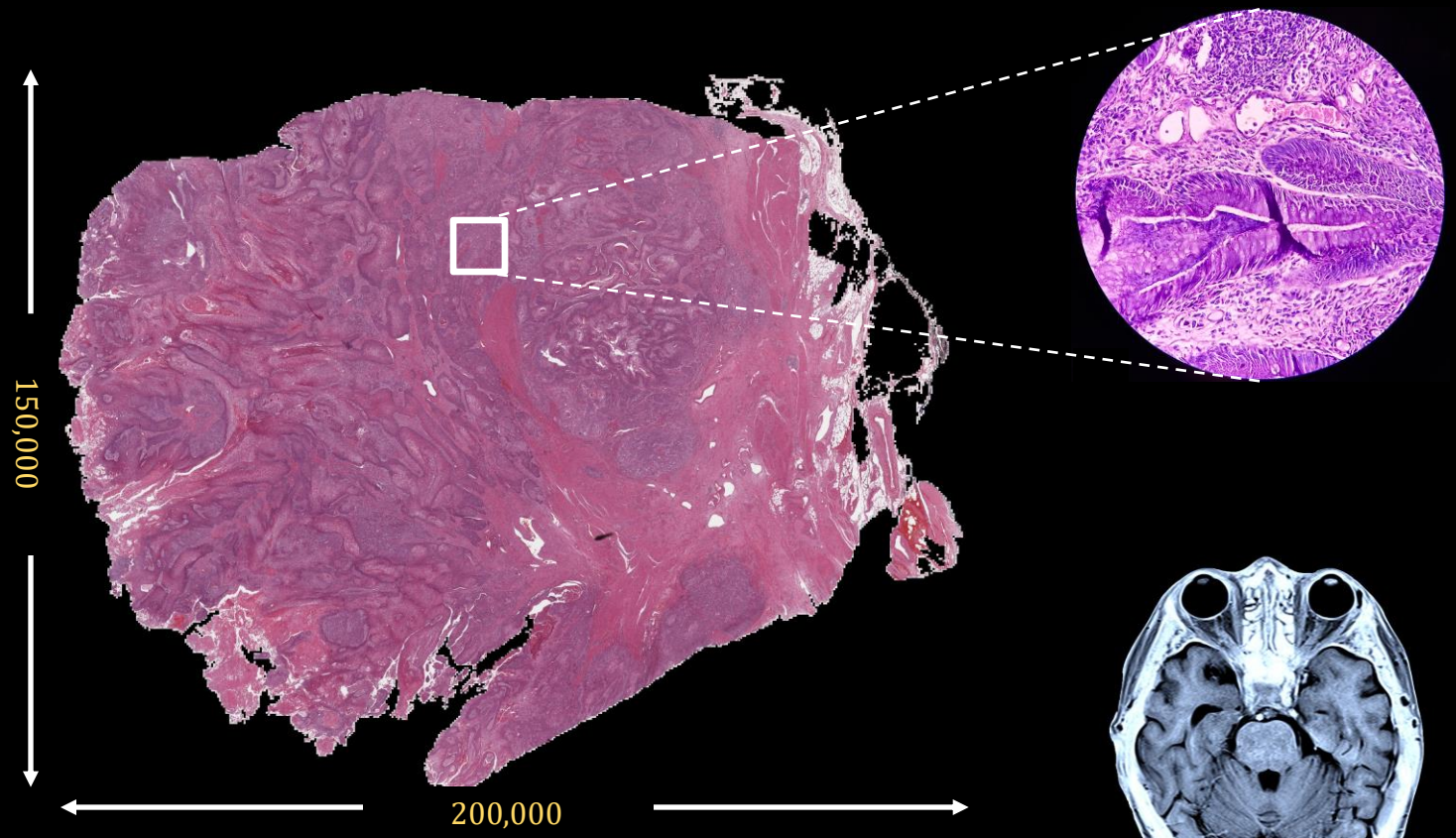
III. Recently, deep learning based medical image analysis is unfolding its infinity power.

IV. While current deep learning-based survival outcome prediction techniques are single-modality, pathology or genomics alone, which inevitably reduce their potential to accurately predict patient prognosis.

F. Zhou et al. Cross-Modal Translation and Alignment for Survival Analysis. ICCV, 2023.

# Challenges

— The enormous heterogeneity of gigapixel WSIs




Whole Slide Image (4GB)



MRI (50MB)

CT (1MB)

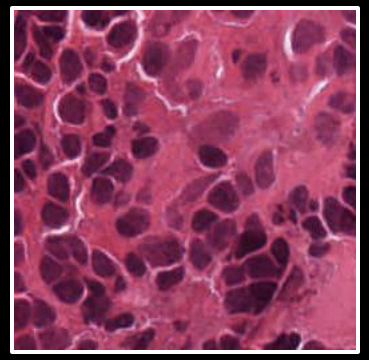
X-Ray (1MB)

TCGA: The Cancer Genome Atlas. [GDC Data Portal]

# Challenges

— The absence of spatially corresponding relationship

```
>chr1:90006571-90007309
CTGAAGGAAATAATTTTGCAAATAATTGAATATATTATAA
>chr1:230843894-230855664
CTCTAGGGTGGGACATGAGAAAGGACAGAAAAAAGAAAGA
>chr1:239632290-239678180
GATCCTTTTTCTAGTTGAGCTATTTCCTTGAAAAGGGGGA
>chr1:182489205-182492965
AGAGCAGTCACTGTAATTTTTTTGACCTTTACATATGGGC
>chr1:230236424-230236654
GTACTGACACTGATTTTATCCCTGTGTCTGGCTCTTCTCT
>chr1:51463798-51465258
CTACAATAAAAAACAAAATCACAAGTTAAATAATAAGAGA
>chr1:68047367-68050547
AAGAAAGTGAGACGGTGAGAGATGGTAGTGGTAACCTACA
>chr1:35895269-35901471
GTGAGATTGCCAAGTAATGGCTGGGGAATAGGCATTGTAT
```



Genomics

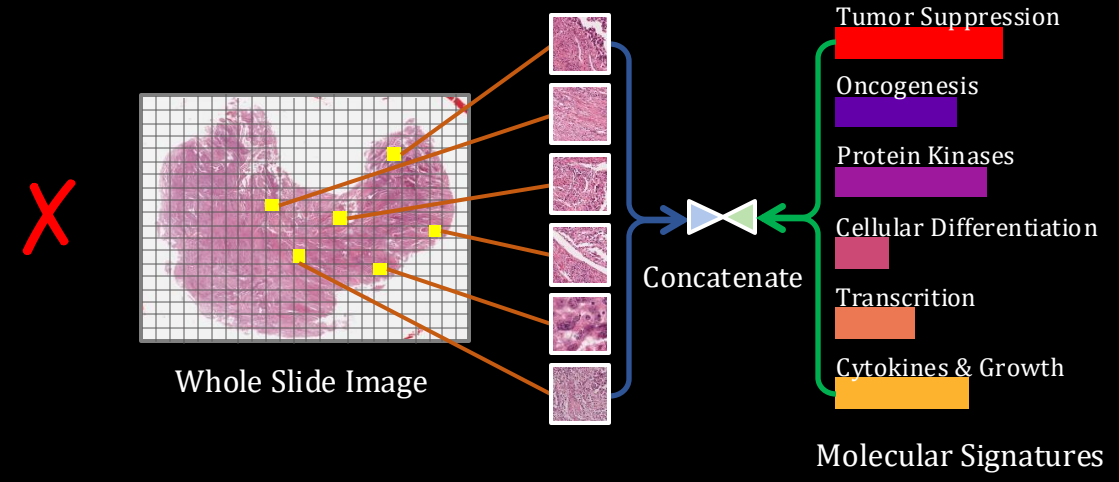
Pathological Image

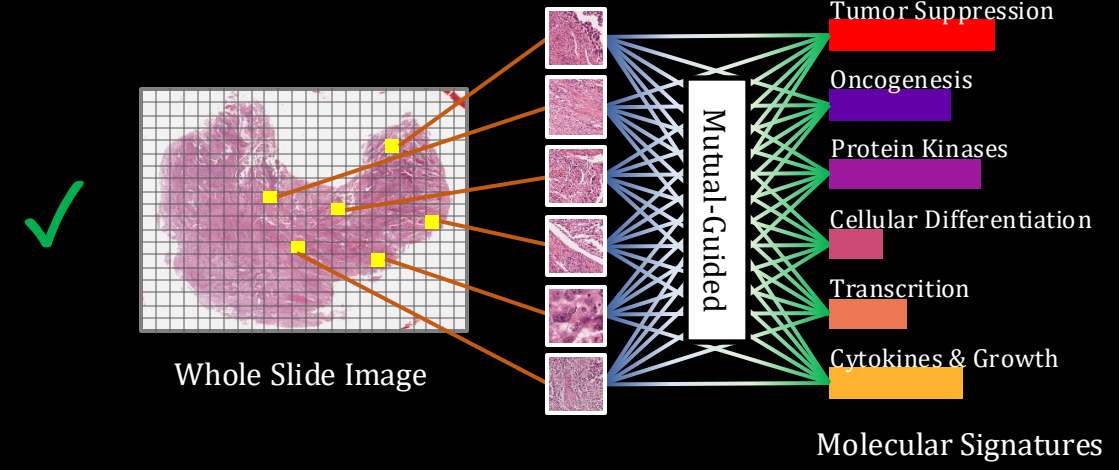
3D Pathology

A. H. Song et al. Weakly Supervised AI for Efficient Analysis of 3D Pathology Samples. arXiv Preprint, 2023.

# Challenges

— SOTAs struggle to capture the explicit interactions

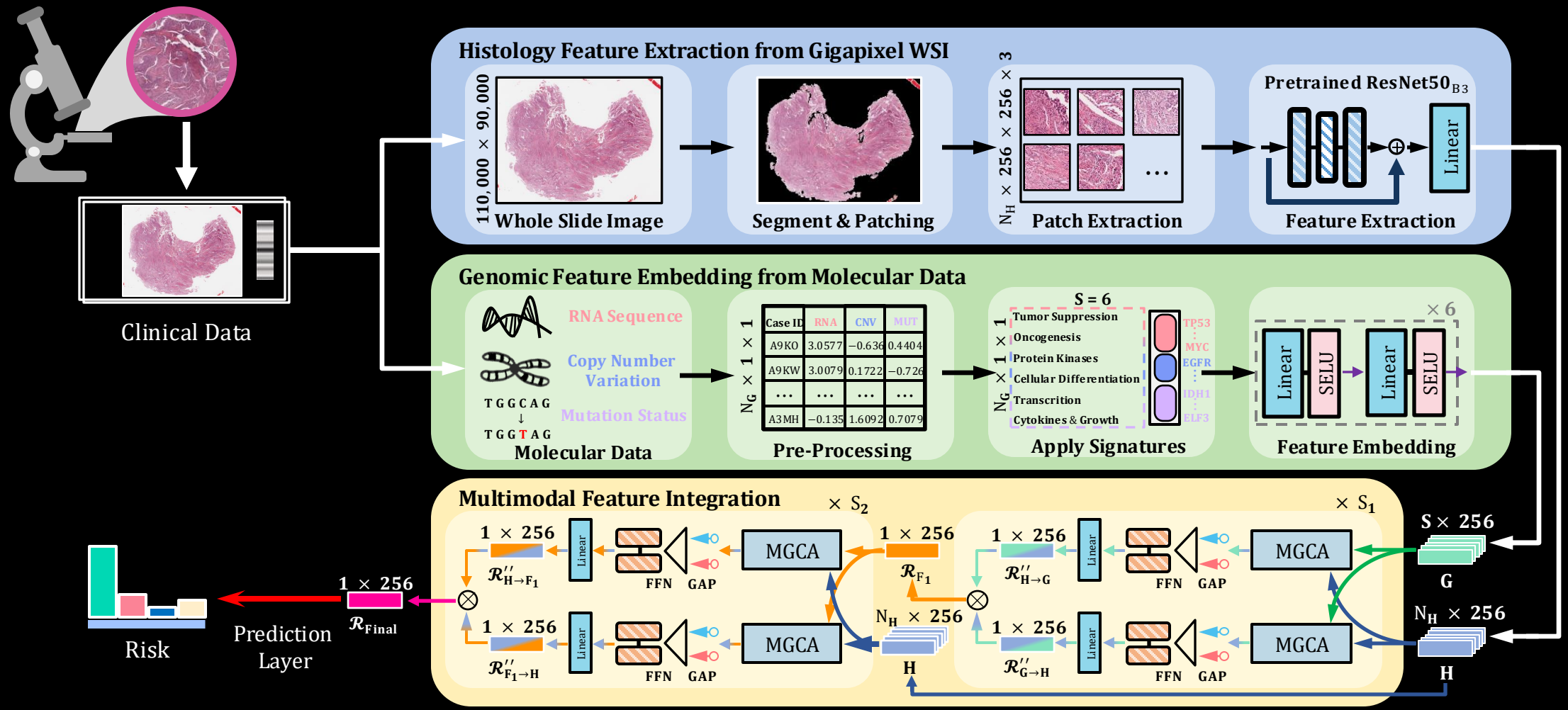



- Current SOTA methods are almost using early, late, intermediate multimodal feature fusion strategies which cannot fully exploit the crucial interactions between histopathology feature and genomic data.
- Some guided-fusion based approaches are solely using the genomic data as the guidance to integrate multimodal pathomic features.
- However, the gigapixel WSIs encompass abundant crucial information including cell appearance, tumor microenvironment (TME), geometrical characteristics.
- Therefore, we designed a novel framework to capture the genotype-phenotype interactions by making these two modalities' data guide each other mutually.

# Methodology

## — Mutual-Guided Cross-Modality Transformer



# Methodology

## — Problem Formulation

- We denote the input WSI as $\mathbb{X}_i$, the feature vector of genomic attributes with the WSI as $\mathbb{G}_i$, the overall survival time (in months) as $t_i \in \mathbb{R}^+$, and the right uncensorship status (death observed) as $c_i \in \{0,1\}$.
- Therefore, we can represent the observations for all patient samples as a quadruple $\{\mathbb{X}_i, \mathbb{G}_i, t_i, c_i\}_{i=1}^{N}$.
- The main objective is to develop and optimize $\mathcal{T}(\cdot)$ for integrating $\mathbb{X}_i$ and $\mathbb{G}_i$ to estimate the hazard function:

$$\widetilde{t_i} = \mathcal{T}(\mathbb{X}_i, \mathbb{G}_i) = \emptyset\left(\xi\left(\rho\left(\left[f(x_1), f(x_2), \cdots, f\left(x_{N_i}\right)\right], \mathbb{G}_i\right)\right)\right)$$

- $f(\cdot)$ is an instance-level encoder that processes features for each instance independently
- $\rho(\cdot)$ is the method for multimodal pathomic features integration
- $\xi(\cdot)$ is a permutation-invariant instance aggregator which aggregate and pools the features to a bag-level embedding
- $\emptyset(\cdot)$ is a bag-level classifier to make final survival outcome predictions

# Methodology

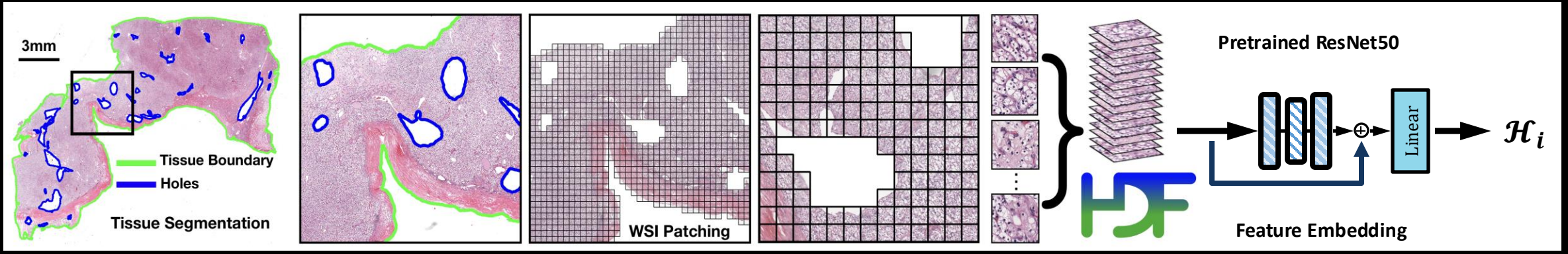
— Histopathology Feature Extraction

a) For input WSI $\mathbb{X}_i$, CLAM repository is employed for automated tissue segmentation.

b) Then we extract 256×256 image patches $\{x_k\}_{k=1}^{N_i}$ without spatial overlapping at the 20× magnification.

c) We further utilize an ImageNet-pretrained ResNet-50 to generate a 1024-dim feature embedding $\mathbf{h}_k \in \mathbb{R}^{1024}$.

d) Finally, we assemble the feature embeddings into a WSI-level bag representation $\mathcal{H}_i \in \mathbb{R}^{1024 \times N_i}$.

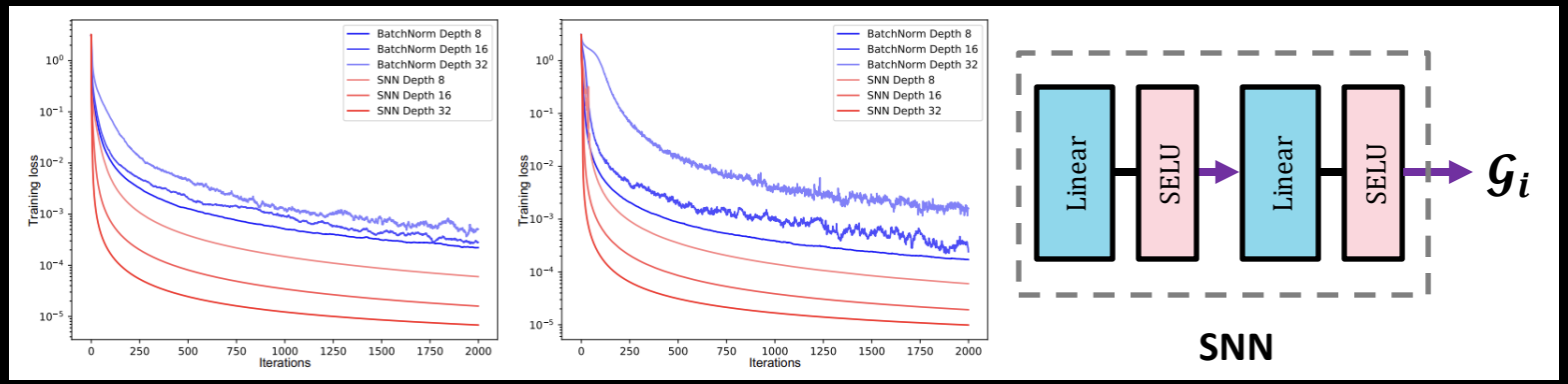



MY Lu et al. Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering, 2021.

# Methodology

— Genomic Feature Embedding

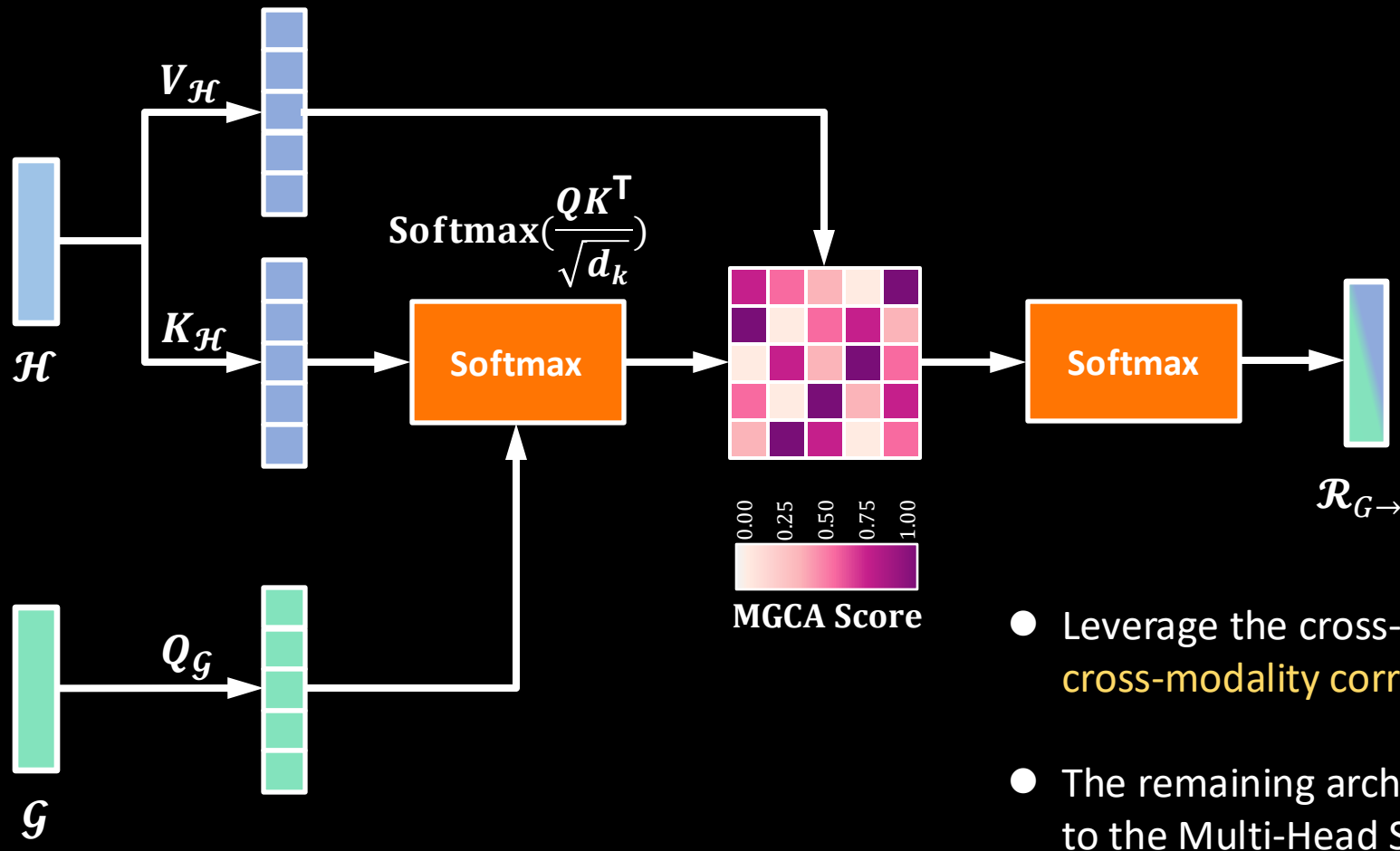
- We select transcript abundance (bulk RNA-Seq), gene mutation status, copy number variation as the input genomics.
- These $1 \times 1$ measurements exhibit a high-dimensional low-sample (HDLSS) nature which leads to overfitting problem.
- Therefore, we leverage the Self-Normalizing Neural Network (SNN) to formulate the genomic feature embedding.
- We further aggregate and structure the genomic embeddings based on $\boldsymbol{S}$ related biological functional impacts.
- Finally, we can generate the bag-level genomic feature embedding as $\mathcal{G}_i \in \mathbb{R}^{1024 \times S}$.




G. Klambauer et al. Self-normalizing neural networks. NeurIPS, 2017.

# Methodology

— Mutual-Guided Cross-Modality Attention




genomic-guided histopathology feature embedding

- Leverage the cross-attention module to explore the intrinsic cross-modality correlations and comprehensive interactions.
- The remaining architectures and calculation are analogous to the Multi-Head Self-Attention (MHSA) within vanilla transformer encoder layer.

A. Vaswani et al. Attention is all you need. NeurIPS, 2017.

# Methodology

— Mutual-Guided Cross-Modality Transformer

- The procedure for MGCT layer calculation:

$$\mathbf{MGCA}(\mathcal{G}_i, \mathcal{H}_i, \mathcal{H}_i) = \mathbf{Softmax}\left(\frac{\mathrm{Q}\cdot\mathrm{K}^{\intercal}}{\sqrt{d_k}}\right)$$

$$= \mathbf{Softmax}\left(\frac{\mathrm{W}_q\cdot\mathcal{G}_i\cdot\mathcal{H}_i^{\intercal}\cdot\mathrm{W}_k^{\intercal}}{\sqrt{d_k}}\right)\cdot\mathrm{W}_v\cdot\mathcal{H}_i \longrightarrow \mathcal{R}_{G\to H}$$

$$\mathcal{R}'_{G\to H} = \mathbf{AttnPool}\left(\sum_{i=1}^{N}\alpha_i\right)\cdot\mathcal{R}_{G\to H} \qquad where$$

$$\alpha_i = \frac{\exp\{\mathrm{W}(\mathbf{tanh}(\mathrm{V}\cdot\mathcal{R}_i^{\intercal})\odot\mathbf{sigm}(\mathrm{U}\cdot\mathcal{R}_i^{\intercal}))\}}{\sum_{j=1}^{N}\exp\{\mathrm{W}(\mathbf{tanh}(\mathrm{V}\cdot\mathcal{R}_i^{\intercal})\odot\mathbf{sigm}(\mathrm{U}\cdot\mathcal{R}_i^{\intercal}))\}}$$

$$\mathcal{R}''_{G\to H} = \zeta(\mathbf{MLP}(\mathcal{R}'_{G\to H})\mathrm{W}_{\mathbf{MLP}})\cdot\mathrm{W}_{\zeta}$$

$\zeta(\cdot)$ is a permutation-invariant instance aggregator

**Algorithm 1**: The proposed MGCT framework

**Input:**

I. WSI bag embedding $\mathcal{H}_i \in \mathbb{R}^{1024\times N_i}$.
II. Genomic bag embedding $\mathcal{G}_i \in \mathbb{R}^{1024\times \mathrm{S}}$.
III. # MGCT layers in two multimodal feature integration stages, $\mathrm{S}_1$ and $\mathrm{S}_2$.

1: **for** $\mathrm{s}_1 = 1$ to $\mathrm{S}_1$ **do**

2: $\mathcal{R}''_{G\to H} \leftarrow \mathbf{MGCT\text{-}Layer}(\mathcal{G}_i, \mathcal{H}_i, \mathcal{H}_i)$
$\mathcal{R}''_{H\to G} \leftarrow \mathbf{MGCT\text{–}Layer}(\mathcal{H}_i, \mathcal{G}_i, \mathcal{G}_i)$

3: **end for**

4: $\mathcal{R}_{F_1} \leftarrow \mathbf{Concatenate}(\mathcal{R}''_{G\to H}, \mathcal{R}''_{H\to G})$

5: **for** $\mathrm{s}_2 = 1$ to $\mathrm{S}_2$ **do**

6: $\mathcal{R}''_{F_1\to H} \leftarrow \mathbf{MGCT\text{-}Layer}(\mathcal{R}_{F_1}, \mathcal{H}_i, \mathcal{H}_i)$
$\mathcal{R}''_{H\to F_1} \leftarrow \mathbf{MGCT\text{–}Layer}(\mathcal{H}_i, \mathcal{R}_{F_1}, \mathcal{R}_{F_1})$

7: **end for**

8: $\mathcal{R}_{\mathrm{Final}} \leftarrow \mathbf{Concatenate}(\mathcal{R}''_{F_1\to H}, \mathcal{R}''_{H\to F_1})$

**Return** final multimodal feature embedding $\mathcal{R}_{\mathrm{Final}}$

# Experiments

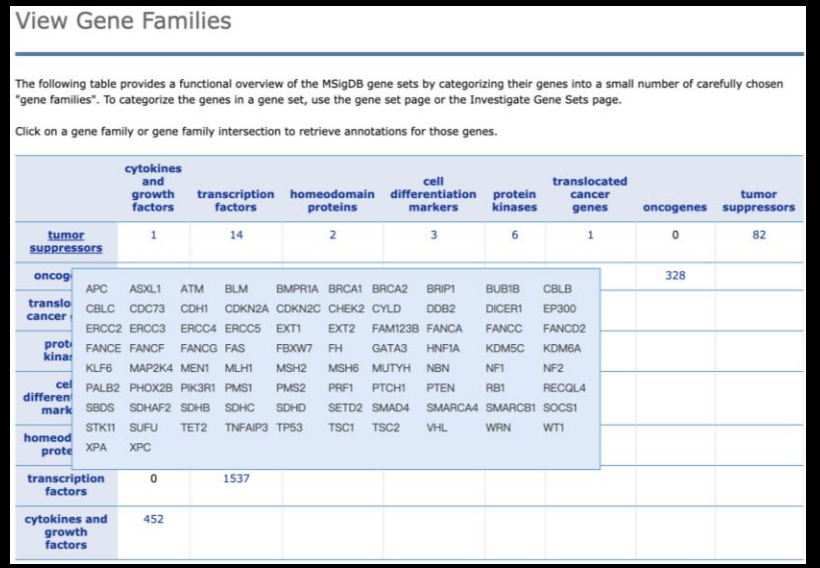
— Datasets

- Five benchmarks were used for model evaluation
- BLCA: Bladder Urothelial Carcinoma
- BRCA: Breast Invasive Carcinoma
- LUAD: Lung Adenocarcinoma.
- GBMLGG: Glioblastoma Multiforme & Brain Lower Grade Glioma
- UCEC: Uterine Corpus Endometrial Carcinoma

| Cancer | # Cases | # WSIs | # Patches | Censorship |
|---|---|---|---|---|
| BLCA | 373 | 437 | 7,648,953 | 0.547 |
| BRCA | 957 | 1,023 | 12,306,155 | 0.860 |
| LUAD | 453 | 516 | 6,717.757 | 0.651 |
| GBMLGG | 569 | 1,042 | 12,742,037 | 0.766 |
| UCEC | 480 | 539 | 9,136,545 | 0.844 |
| Overall | 2,832 | 3,557 | 48,551,447 | 0.734 |

Genomic profiles are grouped by:

a) Tumor Suppression
b) Oncogenesis
c) Protein Kinases
d) Cellular Differentiation
e) Transcription
f) Cytokines and Growth




A. Subramanian et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. PNAS, 2005.

# Experiments

## — Evaluation Metrics

- 5-fold Monte Carlo cross-validation for each cancer type was used for model evaluation.
- Concordance index (C-index) values were employed to measure the predictive ability of the model.
- Kaplan-Meier curves (KM curve) were leveraged to visually represent the quality of patient stratification.
- Log-rank test was introduced to determine the statistical significance of patient stratification.

## — Implementation Details

- MGCT is trained on a workstation equipped with an NVIDIA Quadro GV100 GPU for 20 epochs (about 7.5 hours).
- Adam optimization with learning rate of 2e-4 and weight decay of 1e-5.
- Batch size is 1 (due to samples having varying bag sizes) and 32 gradient accumulation steps.
- Our related models and scripts will be publicly made available ASAP at https://github.com/lmxmercy/MGCT.

# Experiments

— Baselines

- Unimodal Baselines:
  a) Genomic Only: MLP, SNN, DeepSurv, CoxRegression
  b) Pathology Only: Deep Sets, Attention MIL, CLAM, DeepAttnMISL, Patch-GCN
- Multimodal Baselines:
  c) Enhanced MILs with concatenation and bilinear pooling as multimodal baselines
  d) Current State-of-the-Art methods: PORPOISE, MCAT

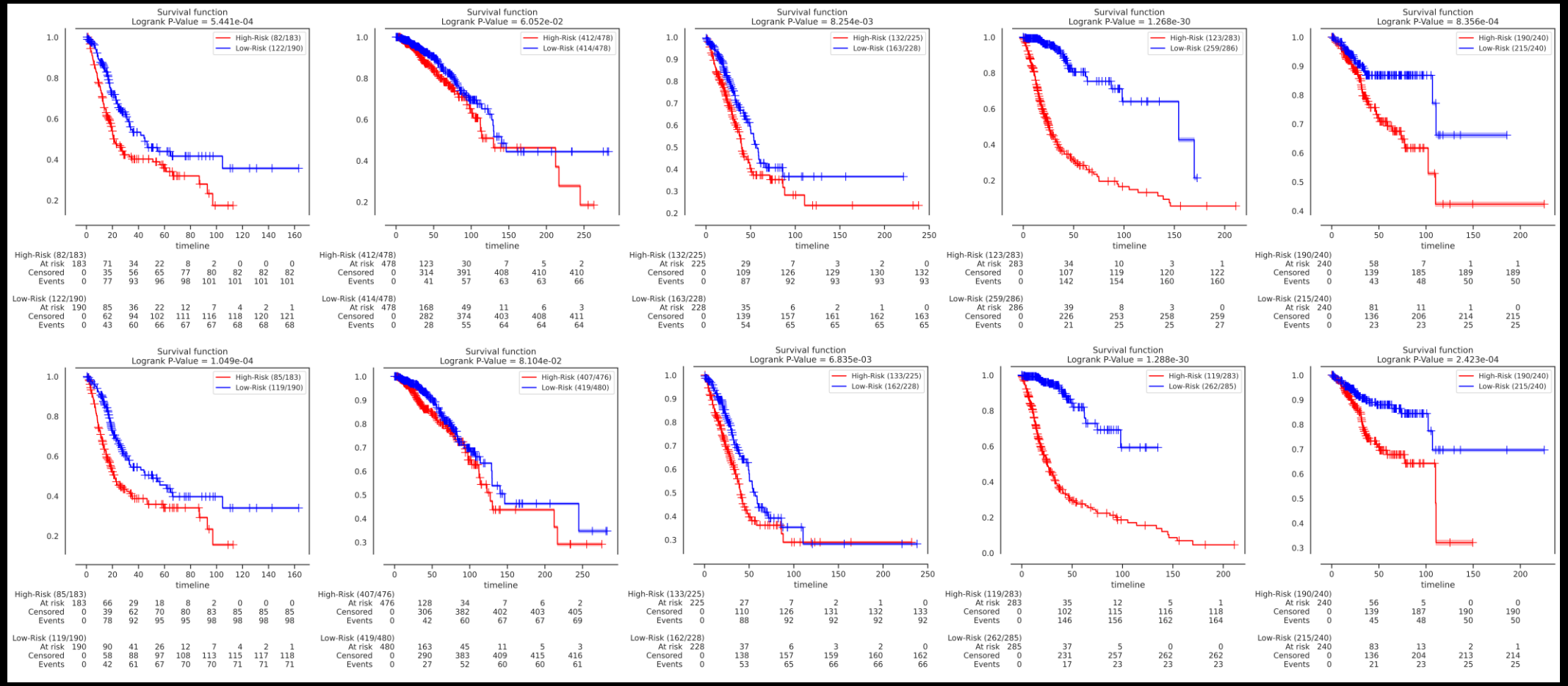
# Experiments

— Concordance index Comparison

| | Methods | BLCA | BRCA | LUAD | GBMLGG | UCEC | Overall |
|---|---|---|---|---|---|---|---|
| Genomic | SNN | 0.541 ± 0.016 | 0.466 ± 0.058 | 0.539 ± 0.069 | 0.598 ± 0.054 | 0.493 ± 0.096 | 0.527 |
| | DeepSurv | 0.567 ± 0.049 | 0.598 ± 0.054 | 0.608 ± 0.026 | 0.810 ± 0.020 | 0.577 ± 0.058 | 0.632 |
| | CoxRegression | 0.591 ± 0.041 | 0.568 ± 0.077 | 0.574 ± 0.042 | 0.705 ± 0.014 | 0.464 ± 0.099 | 0.580 |
| Pathology | Deep Sets | 0.500 ± 0.000 | 0.500 ± 0.000 | 0.496 ± 0.008 | 0.498 ± 0.014 | 0.500 ± 0.000 | 0.499 |
| | CLAM | 0.565 ± 0.027 | 0.578 ± 0.032 | 0.582 ± 0.072 | 0.776 ± 0.034 | 0.609 ± 0.082 | 0.622 |
| | DeepAttnMISL | 0.504 ± 0.042 | 0.524 ± 0.043 | 0.548 ± 0.050 | 0.734 ± 0.029 | 0.597 ± 0.059 | 0.581 |
| | Patch-GCN | 0.560 ± 0.034 | 0.580 ± 0.025 | 0.585 ± 0.012 | 0.824 ± 0.024 | 0.629 ± 0.052 | 0.636 |
| Multimodal | Attention MIL (Concat) | 0.605 ± 0.045 | 0.551 ± 0.077 | 0.563 ± 0.050 | 0.816 ± 0.011 | 0.614 ± 0.052 | 0.630 |
| | DeepAttnMISL (Concat) | 0.611 ± 0.049 | 0.545 ± 0.071 | 0.595 ± 0.061 | 0.805 ± 0.014 | 0.615 ± 0.020 | 0.634 |
| | PORPOISE | 0.613 ± 0.021 | 0.563 ± 0.056 | **0.621 ± 0.045** | 0.818 ± 0.011 | 0.622 ± 0.061 | 0.647 |
| | MCAT | 0.624 ± 0.034 | 0.580 ± 0.069 | 0.620 ± 0.032 | 0.817 ± 0.021 | 0.622 ± 0.019 | 0.653 |
| | **MGCT (Ours)** | **0.640 ± 0.039** | **0.608 ± 0.026** | 0.596 ± 0.078 | **0.827 ± 0.024** | **0.645 ± 0.039** | **0.663** |

# Experiments

— Patient Stratification: Kaplan-Meier Survival Curves



# Experiments

— Ablation study on designed components

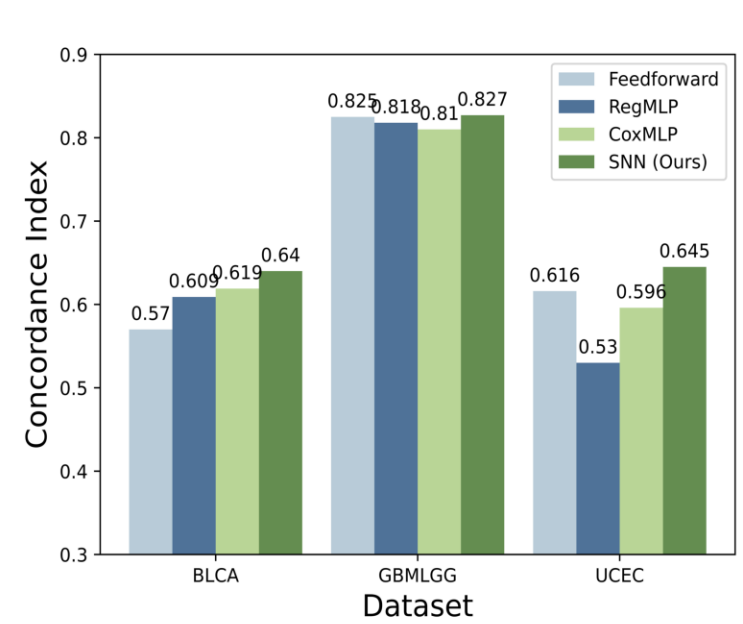
- Test on TCGA-BLCA and TCGA-UCEC two benchmarks.
- Deep Fusion: stack two parallel MGCT layers in depth.
- MGCA: mutual-guided cross-modality attention.
- GAP: gated-attention pooling operation in MGCT layer.
- Feedforward: position-wise feed-forward network in MGCT layer.

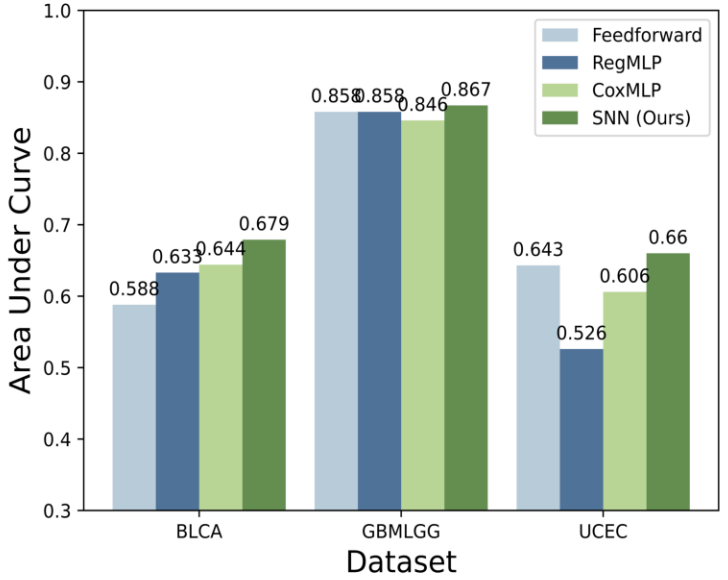
| Model | Designs in MGCT | | | | TCGA-BLCA | | TCGA-UCEC | |
|---|---|---|---|---|---|---|---|---|
| | Deep Fusion | MGCA | GAP | Feedforward | C-index ↑ | AUC ↑ | C-index ↑ | AUC ↑ |
| A | | | | | $0.499 \pm 0.002$ | $0.499 \pm 0.002$ | $0.499 \pm 0.002$ | $0.499 \pm 0.002$ |
| B | √ | | | | $0.535 \pm 0.038$ | $0.532 \pm 0.045$ | $0.541 \pm 0.063$ | $0.558 \pm 0.034$ |
| C | √ | √ | | | $0.590 \pm 0.045$ | $0.621 \pm 0.072$ | $0.608 \pm 0.062$ | $0.627 \pm 0.071$ |
| D | √ | √ | √ | | $0.601 \pm 0.047$ | $0.621 \pm 0.072$ | $0.608 \pm 0.062$ | $0.627 \pm 0.071$ |
| E | √ | √ | √ | √ | **$0.640 \pm 0.039$** | **$0.679 \pm 0.039$** | **$0.645 \pm 0.039$** | **$0.660 \pm 0.039$** |

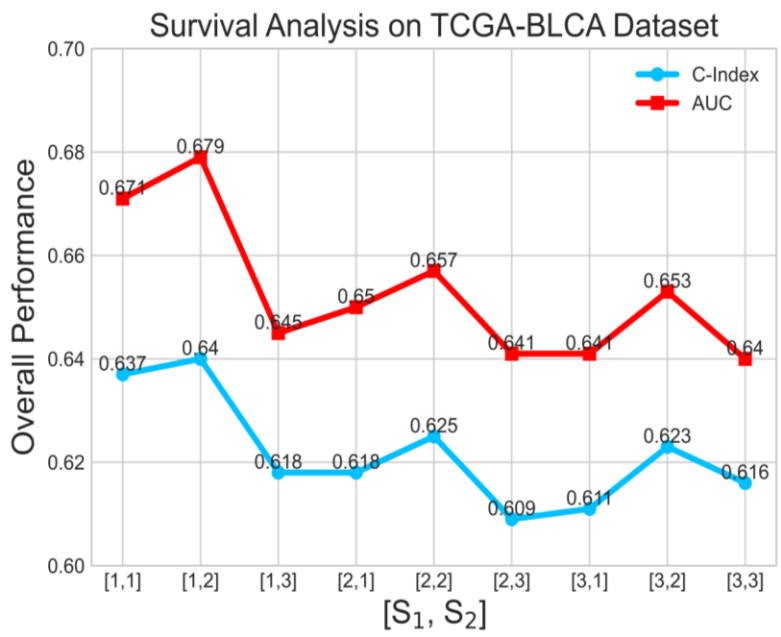
# Experiments

— Ablation study on genomic feature embedding method & #MGCT layers

- Test on TCGA-BLCA, TCGA-GBMLGG, and TCGA-UCEC three benchmarks.



# References


- Xia C, Dong X, Li H, et al. Cancer statistics in China and United States, 2022: profiles, trends, and determinants. Chinese medical journal, 2022, 135(05): 584-590.
- Bray F, Laversanne M, Weiderpass E, et al. The ever-increasing importance of cancer as a leading cause of premature death worldwide. Cancer, 2021, 127(16): 3029-3030.
- Sung H, Ferlay J, Siegel R L, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 2021, 71(3): 209-249.
- Siegel R L, Miller K D, Wagle N S, et al. Cancer statistics, 2023. CA: a cancer journal for clinicians, 2023, 73(1): 17-48.
- Zhou F, Chen H. Cross-Modal Translation and Alignment for Survival Analysis. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 21485-21494.
- Song A H, Williams M, Williamson D F K, et al. Weakly Supervised AI for Efficient Analysis of 3D Pathology Samples. arXiv preprint arXiv:2307.14907, 2023.
- Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in neural information processing systems, 2017, 30.
- Subramanian A, Tamayo P, Mootha V K, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 2005, 102(43): 15545-15550.
- Klambauer G, Unterthiner T, Mayr A, et al. Self-normalizing neural networks. Advances in neural information processing systems, 2017, 30.

# References


- Katzman J L, Shaham U, Cloninger A, et al. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC medical research methodology, 2018, 18(1): 1-12.
- Kvamme H, Borgan Ø, Scheel I. Time-to-event prediction with neural networks and Cox regression. arXiv preprint arXiv:1907.00825, 2019.
- Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. International conference on machine learning. PMLR, 2018: 2127-2136.
- Lu M Y, Williamson D F K, Chen T Y, et al. Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering, 2021, 5(6): 555-570.
- Yao J, Zhu X, Jonnagaddala J, et al. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. Medical Image Analysis, 2020, 65: 101789.
- Chen R J, Lu M Y, Shaban M, et al. Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24. Springer International Publishing, 2021: 339-349.
- Chen R J, Lu M Y, Williamson D F K, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell, 2022, 40(8): 865-878. e6.
- Chen R J, Lu M Y, Weng W H, et al. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 4015-4025.

# Thank You!

E-mail: mxliu.mercy@gmail.com