MULTIMODAL REPRESENTATION LEARNING FOR HEALTHCARE INTEGRATING IMAGING AND GENOMIC DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate survival prediction in brain tumor patients remains a critical challenge in precision oncology. Multimodal deep learning models that integrate histopathological images and genomic data have shown promise in improving prognostic accuracy. In this study, we propose a cross-attention-based fusion architecture that enhances feature interaction between Whole Slide Images (WSIs) and RNA sequencing data. Unlike traditional averaging-based fusion methods, our model uses RNA profiles to dynamically reweight image patch features, enabling more context-aware aggregation. Using a subset of the TCGA glioma dataset, we evaluate performance using the Concordance Index and compare our results with the baseline joint fusion model by Steyaert et al.(5). While our model did not exceed the baseline performance, it demonstrates stable results and introduces architectural flexibility. Future directions include hyperparameter optimization, attention visualization, and the integration of additional clinical modalities. Our work lays the groundwork for more adaptive and interpretable multimodal prognostic models. Codebase:
`https://github.com/revacholiere/MultiModalBrainSurvival_coms672`

1 INTRODUCTION

Brain tumors, both in adult and pediatric populations, present significant clinical challenges due to their heterogeneity, aggressiveness, and complex biological behavior(5). Accurate prognosis prediction is essential for guiding treatment decisions and improving patient outcomes. In recent years, the integration of high-dimensional biomedical data such as histopathology whole slide images (WSIs) and genomic profiles has opened new avenues for enhancing survival prediction models. Multimodal deep learning has emerged as a powerful paradigm to combine these complementary modalities, leveraging the strengths of each to extract more informative representations than unimodal models alone.

The work by Steyaert et al. (5) demonstrated a promising step forward by employing a multimodal deep learning framework to predict prognosis in brain tumor patients, using joint fusion of WSIs and RNA sequencing data. Their study highlighted the benefits of combining imaging and molecular data, using feature aggregation and late fusion strategies to predict patient survival. However, their approach relied on relatively simple averaging techniques during fusion and used fixed dropout strategies that may limit learning efficiency and model adaptability across diverse patient data.

In our project, we aim to build upon this foundation by enhancing the fusion methodology used in multimodal survival prediction. Specifically, we introduce a **cross-attention** (6) mechanism that allows dynamic weighting of features extracted from image patches, conditioned on genomic information. This approach enables more expressive interactions between modalities and facilitates more informed decision-making in the fusion process. Our method aligns more closely with early fusion principles, offering a flexible and learnable fusion layer that improves upon the static averaging used in the baseline.

We also address key limitations of the original model, including excessive dropout and loss inconsistency caused by variable bag sizes. By redesigning the feature aggregation and fusion steps, our

goal is to produce a model that not only improves prediction performance but also provides more interpretable insights into the relative importance of each modality for survival prediction.

2 LITERATURE REVIEW

The integration of multimodal biomedical data such as histopathological images and gene expression profiles has been increasingly recognized as a critical strategy for improving prognosis prediction in oncology. Traditional survival analysis methods, including Cox proportional hazards models (3), often fall short when dealing with high-dimensional, nonlinear, and heterogeneous data. To overcome these limitations, deep learning has been widely adopted, providing the ability to learn complex patterns from large-scale datasets.

A major advancement in this space was introduced by Steyaert et al. (5), who proposed a joint fusion model combining whole slide imaging (WSI) and RNA sequencing data for brain tumor prognosis prediction. Their architecture used either feature-level or prediction-level fusion to integrate the modalities, with performance evaluated through the concordance index (CI). Despite achieving respectable CI scores (WSI: 0.656, Case: 0.621), the model's fusion strategies relied on simple averaging operations, potentially underutilizing the rich interactions between image and genomic features.

Earlier work on multimodal learning in medical prognosis includes models such as Patch-GCN and CLAM (2), which focus on slide-level representation learning using attention mechanisms, but often omit genomic data. Others like Mobadersany et al. (4) combined histology images with clinical features using CNNs and survival models, demonstrating the effectiveness of multimodal inputs but limited by rigid fusion strategies and lack of interaction modeling.

Cross-attention mechanisms have recently emerged as a promising tool in multimodal learning, particularly for tasks that require alignment or context-sensitive weighting of different modalities (8). Such mechanisms are especially beneficial in scenarios like survival prediction, where different modalities may contribute unequally to prognosis at various stages (1).

In addition to model architecture, other studies have emphasized the importance of robust training strategies and interpretability. High dropout rates, while useful for regularization, can sometimes hinder learning, especially when combined with small sample sizes or multiple forward passes per patient. There is also growing interest in interpretable deep learning models in healthcare, including techniques like attention heatmaps and feature attribution scores, which provide insight into the biological or clinical significance of learned representations.

In light of these developments, our work addresses several key limitations in prior studies. By incorporating a cross-attention module that uses gene expression data to query and reweight patch-level image features, our model allows more nuanced fusion and dynamic feature interactions. This architecture represents a hybrid between early and joint fusion, offering better expressivity and potential gains in performance and interoperability.

3 MATERIALS AND METHODS

3.1 DATASET DESCRIPTION

For this study, we utilized data from The Cancer Genome Atlas (TCGA) adult glioma cohort(7), which includes multimodal patient information consisting of histopathological Whole Slide Images (WSIs) and corresponding RNA sequencing data. The original dataset comprises:

- **844 WSIs** from **507 patients**
- **Gene expression data** for all 507 patients

Due to computational constraints, we randomly selected **250 WSIs**, ensuring balanced representation across patient cases. Each selected WSI was preprocessed into **2000 fixed-size patches**, which serve as the visual input to the model. Patients with multiple WSIs had their slide-level predictions aggregated during evaluation.

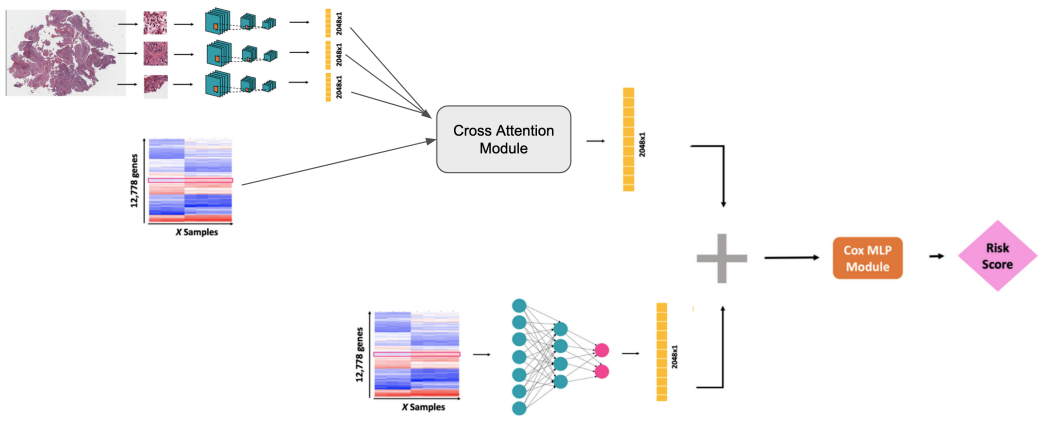

Figure 1: Our proposed architecture that makes use of gene expression data while weighing patch feature vectors.

3.2 Data Preprocessing

The preprocessing pipeline included the following steps:

- **WSI Patching:** Each slide was divided into 2000 non-overlapping image patches of equal size.
- **Feature Extraction:** A pretrained CNN was used to extract feature embeddings from each patch.
- **RNA Sequencing:** Gene expression vectors were standardized using z-score normalization.

The dataset was split into training (70%), validation (10%), and testing (20%) sets, ensuring that data from the same patient did not appear in more than one split to prevent data leakage.

3.3 Model Architecture

We adopted a modified version of the architecture introduced by Steyaert et al. (2023), augmenting the joint-fusion mechanism with a **cross-attention fusion module**. Our approach allows the genomic profile of each patient to influence the aggregation of visual features. Specifically:

- **Input:** Each WSI is represented as a bag of patch-level features (keys and values), and the associated RNA expression vector serves as the query.
- **Fusion:** A scaled dot-product cross-attention mechanism is applied to compute attention weights over the patch features, conditioned on the genomic query.
- **Aggregation:** Weighted average of patch features is computed based on attention scores and passed to the final survival prediction layer.

This mechanism enables the model to selectively emphasize image regions that are more informative in the context of a patient's gene expression profile.

3.4 Training Strategy

We conducted multiple training experiments with varying hyperparameters to assess model robustness. Key training configurations include:

- **Batch sizes:** 128 and 256
- **Bag sizes:** 16 and 32 patches per WSI

- **Dropout:** Set to 0 (unlike the baseline which used 0.8)
- **Learning rates:** 0.001 and 0.01
- **Epochs:** 20 per run

The loss function used was the negative partial log-likelihood of the Cox proportional hazards model. We trained all models using PyTorch on a GPU-enabled environment.

3.5 Evaluation Metric

We employed the **Concordance Index (CI)** to measure predictive performance. CI evaluates the concordance between predicted risk scores and observed survival times:

- CI of 1.0 indicates perfect predictive agreement.
- CI greater than 0.5 indicates better-than-random prediction.
- CI less than 0.5 indicates poor model performance.

CI was calculated at both the **WSI level** (per slide) and the **Case level** (aggregated per patient). This allows evaluation of performance consistency across different granularity levels.

4 Results

To evaluate the effectiveness of our proposed cross-attention-based fusion strategy for brain tumor survival prediction, we conducted extensive experiments using a subset of the TCGA adult glioma dataset. The dataset includes 250 Whole Slide Images (WSIs) and corresponding RNA sequencing data. Each WSI was divided into 2000 patches, and models were trained using various hyperparameter configurations. Performance was measured using the **Concordance Index (CI)**, which reflects the agreement between predicted risk scores and actual survival outcomes.

4.1 Baseline Comparison

We benchmarked our model against the baseline joint fusion approach proposed by Steyaert et al. (2023), which uses simple averaging to fuse image and genomic features. As shown in Table 1, the baseline model achieved a CI of **0.659 (WSI-level)** and **0.621 (Case-level)** with a batch size of 128 and a dropout rate of 0.8.

Table 1: Baseline model performance (Steyaert et al., 2023)

Batch Size	Dropout	CI (WSI)	CI (Case)
128	0.8	0.659	0.621
256	0.8	0.656	0.610

4.2 Proposed Model Performance

In our model, we replaced the averaging mechanism with a **cross-attention module**, where RNA sequence features were used as query vectors to dynamically reweight image patch features (keys and values). We experimented with different learning rates, dropout settings, and bag sizes.

Key results are summarized in Table 2:

Table 2: Proposed model performance with cross-attention

Batch Size	Bag Size	Dropout	Learning Rate	CI (WSI)	CI (Case)
128	32	0	0.001	0.62	0.5806
256	16	0	0.001	0.61	0.56
128	32	0	0.01	0.55	0.51

4.3 Analysis

Although our model did not outperform the baseline in terms of CI scores, it demonstrated stable learning behavior and promising potential for future improvements. The performance drop can be attributed to the following challenges:

- The removal of heavy dropout used in the baseline may have increased overfitting. We had removed it because the dropout layer was preventing our model from consistently learning, leading to random loss values.
- Resource limitations prevented exploration of larger bag sizes and comprehensive grid search.
- Variability introduced by multiple forward passes per sample affected the consistency of the loss function. This is the most likely cause since the original implementation had single forward pass per sample. The ResNet performing feature extraction multiple times, then only being backpropagated through one label may have prevented it from learning properly (because of multiple features being averaged just for 1 prediction).

Despite these constraints, our model introduces significant architectural improvements. The cross-attention mechanism enables **modality-specific feature weighting**, better aligning with early fusion principles and laying the groundwork for future improvements in accuracy and interpretability. Since early fusion separates the feature extractor and fusion models' trainings, this would help us avoid having issues with the ResNet learning incorrectly.

5 Conclusion and Future Work

This study proposed a cross-attention-based multimodal fusion model for brain tumor prognosis by integrating histopathological images and RNA sequencing data. Unlike the baseline by Steyaert et al.(5), which used single patch-based predictions, our model takes in bags of patches and dynamically reweights image features based on genomic context, offering a more expressive fusion mechanism. While the performance, measured by Concordance Index, did not surpass the baseline, the model demonstrated stable results and introduced architectural flexibility that can support future improvements.

Future work will focus on hyperparameter tuning, increasing bag size, and implementing loss normalization for training stability. We also plan to visualize attention maps for interpretability and explore early fusion strategies.

References

[1] Lianghong Chen, Zi Huai Huang, Yan Sun, Mike Domaratzki, Qian Liu, and Pingzhao Hu. Conditional probabilistic diffusion model driven synthetic radiogenomic applications in breast cancer. *PLOS Computational Biology*, 20(10):e1012490, 2024.

[2] Richard J Chen, Ming Y Lu, Muhammad Shaban, Chengkuan Chen, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 339–349. Springer International Publishing, 2021.

[3] Dhananjay Kumar and Bengt Klefsjö. Proportional hazards model: a review. *Reliability Engineering & System Safety*, 44(2):177–188, 1994.

[4] Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A Gutman, Jill S Barnholtz-Sloan, José E Velázquez Vega, Daniel J Brat, and Lee AD Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13):E2970–E2979, 2018.

[5] Sandra Steyaert, Yeping Lina Qiu, Yuanning Zheng, Pritam Mukherjee, Hannes Vogel, and Olivier Gevaert. Multimodal deep learning to predict prognosis in adult and pediatric brain tumors. *Communications Medicine*, 3(1):44, 2023.

[6] Xinyu Wang, Le Sun, Chuhan Lu, and Baozhu Li. A novel transformer network with a cnn-enhanced cross-attention mechanism for hyperspectral image classification. *Remote Sensing*, 16(7):1180, 2024.

[7] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R M Shaw, Beth A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013.

[8] Jin Zhang, Xiaohai He, Yan Liu, Qingyan Cai, Honggang Chen, and Linbo Qing. Multi-modal cross-attention network for alzheimer's disease diagnosis with multi-modality data. *Computers in Biology and Medicine*, 162:107050, 2023.