CS2952G, 2023, pp. 1–5

doi: N/A

SWEEM: Multi-Omics Transformer for Cancer Survival Analysis

Siming Feng, William Guo, Eric Han, Evan Lu and Mason Zhang

Computer Science Dept., Brown University, 69 Brown Street, 02912, Rhode Island, United States

Abstract

Since biological processes associated with cancer are complex and multifaceted, accurately diagnosing cancer progression status is a critical step toward developing strategies for treatment and prevention. To achieve this goal, a multi-omics approach shows great promise for more accurate and biologically-conscious in-silico cancer discovery. Here we trained a multimodal transformer encoder model called SWEEM on an integrated dataset that combines data across the transcriptome, proteome, and epigenome. We observe that the transformer model, compared with machine learning and deep learning baselines, outperforms previous methods by a comfortable margin. Moreover, we applied various interpretative methods to understand the model's decision-making and detected biological significance. Overall, SWEEM successfully identifies relevant biological features associated with cancer and can be used to accurately assess the cancer survival of patients from multi-omics datasets.

Key words: multiomics, transformer, interpretability, cancer survival, deep learning

1. Introduction

Deep Learning is a burgeoning method utilized for multiomics research. The integration of datasets from various modalities including but notwithstanding the transcriptome, proteome, and epigenome better equip models to understand underlying biological problems. The etiology of complex diseases, especially cancer, is not contingent on a single determining factor, and thus, considering different modalities of data provides more holistic and diverse insight (Hasin et al., 2017). In this paper, we focus on using a multi-omics approach for cancer survival analysis through self-attention. By better understanding cancer survival patterns and factors that influence them, we hope to inform and improve treatment strategies, resulting in better survival rates as well as identify early warning signs, risk factors, and biomarkers associated with specific cancer types. Even now, machine learning techniques are still used in this domain due to interpretability, a crucial factor in the clinical domain and especially the task of cancer survival analysis. However, recent advances in deep learning, especially relating to interpretability, have enabled multi-omics to be leveraged for cancer-related tasks with numerous model architectures such as transformers, graph convolutional networks, and bidirectional LSTMs (Osseni et al., 2022; Li et al., 2021; Bichindaritz & Liu, 2022).

To address the lack of existing models that integrate and process multi-omic data with feasible interpretability, we introduce SWEEM, a multi-omic transformer model architecture. Originally introduced to address seq2seq natural language processing tasks, the transformer encoder architecture has become increasingly popular and applicable in various other fields, such as but not limited to computational biology. The transformer architecture's popularity is rooted in the implementation of the attention heads which can be used to better interpret the model's decision process. To our knowledge, SWEEM is the first model that integrates multi-omics data and the transformer encoder model architecture to address cancer survival analysis.

2. Related Works

2.1 Multimodality in Genomics

As seen in Singh et al. 2019, DIABLO was introduced for identifying key molecular drivers for different tasks including cancer discovery. Although deep learning methods were widely used by the computational biology community at that point, DIABLO was still based on CCA/LDA. Using this technique, DIABLO is able to identify relevant features for cancer detection with almost 80% accuracy. On a similar note, Giang et al. 2020 employ a fast-multiple kernel learning framework, achieving accuracies ranging from 72-94% for the detection of different cancer types.

Newer papers in this space that utilize deep learning for multi-omics cancer discovery include Leng et al. 2022, where they evaluate 16 representative deep learning methods on simulated, single-cell, and cancer multi-omics datasets. They focus on evaluating mostly architectures consisting of variations with VAEs and CNNs, and identify the need for addressing class imbalance issues and developing explainable deep learning methods, potentially using knowledge-embedded algorithms. A 2021 paper by Yang et al. introduces a model where an encoder takes in one kind of data and maps it to a shared space while a decoder for a different kind of data can translate the previous type of data into that new modality. This provides key insights into how to informatively embed multi-omic data for downstream use.

2.2 Multimodality in Transformers

There is much recent work with transformers that perform well at utilizing multimodal inputs. Summarized by Xu et al. 2023, the authors note that this may be due to the architecture's inherent tendency to represent its input as graphs, a property well suited to represent multiple modalities. It's this intrinsic advantage in conjunction with their scalability that makes transformers so suited to the task. While we will not be including data with multiple kinds of modalities, but multiple kinds of omics datasets instead, we believe that the inherent benefits that the architecture brings when addressing multiple modalities will carry over when addressing multi-omics datasets. Evidence of this can already be found in Osseni et al. 2022, where a transformer was fed multi-omics information for tumor type classification. The authors found that each omic type was informative to the task and essential to achieving the highest accuracy. Such a good performance in a similar task bodes well for our approach.

2.3 Cancer Survival Analysis

The task that we will be addressing is cancer survival analysis, a departure from the transformer models previously mentioned. This formulation instead focuses on the survival time of a patient given that they have already been diagnosed with cancer. We draw inspiration for this approach from DeepOmix (Zhao et al. 2021), a multilayer perceptron network that encodes flexibility, scalability, and interpretability into its architecture by using prior user-fed biological information to determine how inputs are combined. They determine that this approach not only achieves accurate predictions on survival time but also provides interpretability to the biological underpinnings of the prediction.

Overall, there is promising potential in leveraging a multi-omics approach with a transformer architecture to enhance the accuracy of traditional tasks while also gaining further biological insight into the complex processes associated with cancer and other key diseases.

2.4 Interpretability in Transformers

Transformer methods heavily involve attention operators and skip connections, notably residual layers. In doing so, this introduces a combination of activation functions. Consequently, though attention layers serve as a method to assist with model interpretability, simply relying on attention would be a naive representation of input-output relevancy relations.

However, recent works in the field of deep learning have proposed novel methods to address these issues. Bach et al. propose layer-wise relevance propagation (LRP), a prominent technique used in explainability for machine learning that involves the decomposition of nonlinear classifiers (Bach et al., 2016). Though not originally intended for transformers, LRP has inspired related works such as XAI, which refactors the method to be more appropriate for the transformer architecture (Ali et al. 2022).

Interpretability of transformers is a promising field that is still developing. SWEEM implements and applies these novel techniques to the task of cancer survival analysis for a computationally modern and interpretable model.

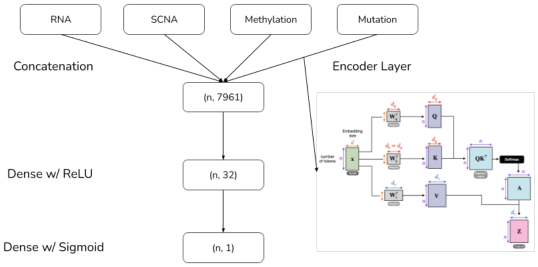


Fig. 1. Model Architecture. The model consisted of a separate encoder applied to each of the four data modalities (the diagram only displays the encoder applied over mutation data). After this, the inputs are concatenated and fed into two linear layers, the first with hidden size 32 with ReLU activation. The output is gated into [0, 1].

3. Method

While previous methods utilized multi-layered perceptrons, SWEEM is a transformer-based encoder model that predicts the rate of cancer survival provided a multi-omic input. We drop the decoder part of a full transformer to better fit the nature of the task (regression instead of seq-to-seq). We also draw inspiration from visualization techniques in the aforementioned related works sections and use them to evaluate the interpretability of our model compared to related models. Our model is open-source for additional reference

3.1 Data acquisition and pre-processing

SWEEM was trained on patients' multi-omics data and their clinical survival time data, the same data that DeepOmix utilized for training. Multi-omics data and clinical survival time data were received from the TCGA pan-cancer dataset, from the UCSC Xena data portal. There were a total of 33 tumor types in the dataset, and for each of the tumor types, there were four types of omics data: mRNA (RNA-seq), somatic mutation, DNA methylation, and somatic copy number alteration (SCNA).

For feature selection on mRNA and DNA methylation data, we used mean absolute deviation to choose the top 2000 features for both sources. After feature selection was performed, we integrated all the data together into one matrix with n rows (# of samples) and m columns (all the omics features). For preprocessing, the distribution of the class labels was evenly split among training and testing sets.

3.2 Model Construction

SWEEM is constructed as a multi-layered self-attention-based encoder architecture. The input layer accepts normalized data from four different omics sources (i.e. RNA, SCNA, Methylation, and Mutation). For the i-th sample, $x_{i,g,k}$ represents the k^{th} omics data of gene g ($k = 4$ in our analysis). The attention mechanism allows the model to weigh the importance of each gene-pathway relationship dynamically.

We introduce Transformer encoder blocks to properly embed the information. Each block consists of self-attention, followed by dropout layers for regularization, and then feed-forward neural networks. The outputs from the Transformer encoder blocks are then passed through two dense layers, akin to the hidden layers in the original architecture. These layers further process the features before they are used to predict the survival data, which consists of survival time and status.

To optimize SWEEM, we pre-train it with binary cross-entropy on the event and fine-tune it on the negative log partial likelihood of the Cox Proportional Hazards model, the standard for cancer survival analysis, which is given by:

$$L(\theta) = -\frac{1}{N} \sum_{i:\delta_i=1} \left(h_i - \log \left(\sum_{j:t_j \geq t_i} e^{h_j} \right) \right) + \lambda \|\theta\|_2^2 \tag{1}$$

The model had a batch size of 16 and a hidden dimension of 32, running for 1000 epochs. The model also used L2 normalization, 80% dropout for input layers, and 50% for intermediate layers to prevent overfitting. There were several important observations during training that informed model hyperparameters: 1) a lower learning rate of 0.0001 was crucial in enabling the model to actually learn and 2) the hidden size for the linear layers after the attention layers could not exceed 64 or else the model would begin to overfit significantly.

Binary cross entropy loss was chosen as the model loss as it was 1) more stable to train in batches and 2) the combined loss using both c-index and Brier score, metrics suited for cancer survival analysis, was not viable since backpropagation became too costly, squaring the amount of computations needed.

Attention was applied across all input features. When testing cross-attention, it performed similarly compared to self-attention but had substantially longer training time and used more than three times as many parameters (252,587,449 vs. 84,502,365 parameters in the respective attention matrices). Thus, cross-attention was not ultimately chosen as it was more expensive space and time-wise compared to self-attention and did not result in substantial improvement.

It should also be noted that using up to three data modalities as features led to overfitting, which indicated that there were not enough samples for training. With the current dataset, there were only 400 samples for 15000 features, which is likely to restrict model performance.

4. Experiments

SWEEM performed competitively compared to the traditional DeepOmix model and better compared to other benchmark methods. The rationale behind this is rooted in the inherent advantages of the transformer architecture, which is designed to capture long-range dependencies and intricate relationships in the data. This is especially crucial for multi-omics data, where interactions between different omics layers can provide valuable insights into the underlying biological processes.

4.1 Baselines

We compared our model results with 4 other popular baseline models that utilize multi-omics data for survival analysis: DeepOmix (MLP), Gradient Boosting (Ensemble), LASSO (L1 prior as regularizer), and Random Forest (Decision trees).

These baseline models, besides DeepOmix, were chosen from the comprehensive evaluation review by Herrmann et al for their model performance. DeepOmix was selected as the most recent model that addresses the same problem as SWEEM, utilizes the same data, and outperforms previous deep learning methods.

5. Results

5.1 Performance comparison with other methods

We compared SWEEM with the other state-of-the-art methods in the baseline above models: Gradient Boosting, LASSO, and Random Forest. Two baselines, DeepSurv and DeepHit were omitted due to their similar model architecture structure to DeepOmix (variations of feed-forward neural networks) and their incompatibilities with certain data modalities. We also compared SWEEM with the DeepOmix model with various input modalities, namely with the Methylation modality channel and with all multiomics data modalities as per the DeepOmix final methodology. All baselines were trained, selected, and tested on all modalities (RNA, SCNA, Mutation, Methylation) other than the distinct cases for DeepOmix. Since the proportion of surviving patients can be unbalanced in training and validation sets, we used the area under the receiver operating characteristics curve (auROC) to quantify the predictive performance of the models. We have also considered and included two other relevant metrics, including the aforementioned concordance index (C-Index) and the Brier Score metric, relevant for survival analysis (Table. 1).

SWEEM performed better than any of the alternate methods on our given datasets and performed competitively with DeepOmix (Table. 1). These results indicate that the transformer encoder architecture with self-attention can capture the relationships within the multi-omics datasets. Additionally, we tested adding self-attention to DeepOmix. Using the same model architecture as DeepOmix, and adding several self-attention layers, we were able to obtain better performance metrics than DeepOmix originally. For this new experiment, we were able to obtain an AUC-ROC of 0.819, a

Table 1. Comparison of Baselines with SWEEM.

Models	AUC-ROC	C-index	Brier Score
Gradient Boosting	0.714	0.462	0.270
LASSO	0.587	0.506	0.292
Random Forest	0.672	0.540	0.250
DeepOmix (Methylation Only)	0.820	0.851	0.282
DeepOmix (All Omics)	0.844	0.876	0.313
SWEEM (simplified)	0.819	0.836	0.211

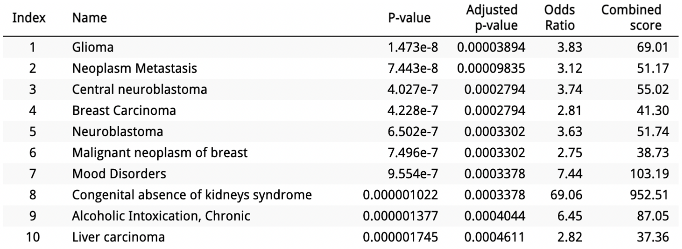
Table 1. An analysis of different baseline functions' auROC, C-index, and Brier score metrics. The range of auROC and C-index is measured from a continuous scale of 0-1, with 1 representing perfect accuracy. The range of the Brier score is measured from a continuous scale of 0-1, with 0 representing no error (i.e. lower is better). Note that SWEEM performs competitively with other baselines across all performance metrics.

C-index of 0.836, and a Brier score of 0.211. Having comparable results to DeepOmix due to self-attention without requiring human-inserted information (i.e. gene interaction pathways) implies that, as mentioned earlier, self-attention can effectively map from multi-omics data to both tasks of survival time analysis as well as patient outcome.

5.2 Interpretability

Integrated gradients were chosen as the model's main method for interpretation as with perturbation analysis being very expensive computationally, it remained unscalable for models with many features. After performing Integrated Gradients to analyze the importance our model places on input genes, we performed gene set enrichment analysis (GSEA) on the top gradients to enhance model interpretability.

Based on promising results from GSEA and examining the highest magnitude gradients themselves, we conclude that SWEEM is capable of capturing biological significance in the data, such as the key genes associated with glioma identification and prognosis.



Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Glioma	1.473e-8	0.00003894	3.83	69.01
2	Neoplasm Metastasis	7.443e-8	0.00009835	3.12	51.17
3	Central neuroblastoma	4.027e-7	0.0002794	3.74	55.02
4	Breast Carcinoma	4.228e-7	0.0002794	2.81	41.30
5	Neuroblastoma	6.502e-7	0.0003302	3.63	51.74
6	Malignant neoplasm of breast	7.496e-7	0.0003302	2.75	38.73
7	Mood Disorders	9.554e-7	0.0003378	7.44	103.19
8	Congenital absence of kidneys syndrome	0.000001022	0.0003378	69.06	952.51
9	Alcoholic Intoxication, Chronic	0.000001377	0.0004044	6.45	87.05
10	Liver carcinoma	0.000001745	0.0004611	2.82	37.36

Fig. 2. Enrichment analysis results for the top 100 magnitude gradients to known diseases in DisGeNET, one of the largest gene-disease associations databases. Using the open source tool Enrichr, glioma was the top enriched term with the lowest p-value (3.984e-5).

We used Enrichr to perform GSEA on the mean top 100 genes, obtained by finding the integrated gradients with the highest absolute magnitudes, those that contribute most to the model's predictions. Glioma was elicited as the top enriched term with the lowest p-value (Fig 2.). This is in line with the DeepOmix dataset, whose clinical information data (including the survival status) is based on brain lower-grade glioma (LGG). SWEEM is significant in that the gene association is extracted from the attention layer itself. In comparison, DeepOmix utilizes pathway modules, which require pre-existing biological knowledge to properly use (lest risk overfitting on noise).

Another method to understand model predictions was examining the genes relevant to the predictions themselves and evaluating their interactions with glioma cancer. (Fig. 3.) SWEEM effectively captures the underlying biological significance of glioma-associated genes based on the integrated gradient values with the highest magnitudes, denoting their effect on the predicted survival outcome. Notably, the gene with the highest positive gradient, FGFR3, is a very significant gene for glioma identification, as supported by literature (Bielle, 2018; Georgescu, 2021). SSTR1, a somatostatin receptor gene, is also highly expressed in glioma samples and identified as a valuable prognostic factor for glioma (He et al, 2021). SWEEM's ability to highlight individual genes allows for not only the validation of pre-existing biological knowledge and context but also the potential discovery of new gene-disease interactions.

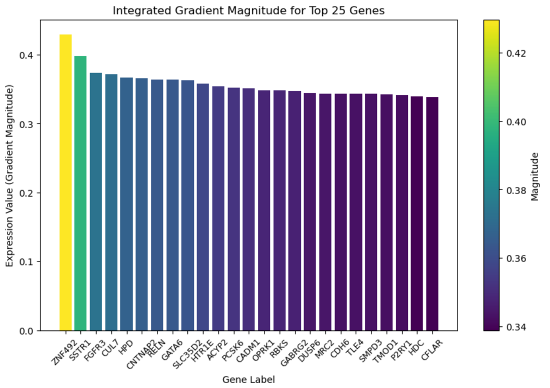


Fig. 3. A list of averaged magnitudes of integrated gradients for SWEEM run with methylation data over the test set. A higher magnitude for a given gene denotes a strong influence on the predicted survival output. Top genes (e.g. ZNF492, SSTR1, FGFR3) were traced to their association with glioma cancer.

6. Discussion / Conclusion

Here we present SWEEM, the first model that integrates multi-omics data and the transformer encoder architecture to address the task of interpretable cancer survival analysis. SWEEM performs competitively against previous baselines in successfully identifying relevant biological aspects associated with cancer and accurately assessing cancer survival. We systemically applied the model to the multi-omics data provided by the DeepOmix dataset. In all metrics provided above, our model matches or outperforms previous baseline models.

Though our model outperforms predecessor models, transformer-based models are fundamentally more complex and computationally expensive; despite this, our model is still relatively stable within the lens of similar deep learning transformer models and the training time should not be considered lengthy by the standards of other models in the multi-omics field.

The task of creating models for cancer survival analysis is fundamentally roadblocked by its interpretability; though a more complex model's performance may improve in comparison to previous works and our model, it is ultimately less valuable to healthcare professionals if the model cannot provide a rationale for its predictions. In our choice of a transformer-based architecture, the interpretability of our model becomes an intrinsically harder task in comparison to models such as random forests or DeepOmix with a multi-layer perception.

Powered by its self-attention architecture SWEEM is able to attend to biologically significant genes to inform its decision on patient cancer survival. The removal of a need for hand-curated biological information to be fed into the model while still achieving significant and interpretable results shows the potential of using the transformer model in multi-omics datasets.

Within the scope of model interpretability, our future steps will primarily entail applying novel transformer-based visualization techniques to improve our understanding of input relevancy and expanding our model to include additional layers if supported by said discoveries. Another potential action would be to include additional data from other sources and electronic medical records to improve model accuracy. In conclusion, as an effective and interpretable tool, SWEEM facilitates promising results for multi-omic cancer survival analysis and offers strong future potential as an avenue to address interpretability issues.

Author contributions statement

All authors contributed equally.

Acknowledgments

The authors thank the anonymous reviewers for their valuable suggestions. This work is supported by the amazing students, TA, and Professor of CS2952G.

References

1. I. Bichindaritz and G. Liu. Adaptive Multi-omics Survival Analysis in Cancer. In: YW. Chen, S. Tanaka, RJ. Howlett, LC. Jain (eds) *Innovation in Medicine and Healthcare. Smart Innovation, Systems and Technologies*, vol 308. Springer, Singapore, 2022.
2. T. T. Giang, T. P. Nguyen, and D. H. Tran. Stratifying patients using fast multiple kernel learning framework: Case studies of Alzheimer's disease and cancers. *BMC Medical Informatics and Decision Making*, 20:108, 2020.
3. Y. Hasin, M. Seldin, and A. Lusis. Multi-omics approaches to disease. *Genome Biology*, 18:83, 2017.
4. M. Herrmann, P. Probst, R. Hornung, V. Jurinovic, A. Boulesteix Large-scale benchmark study of survival prediction methods using multi-omics data *Briefings in Bioinformatics*, 22:3, 2021.
5. D. Leng, L. Zheng, Y. Wen, et al. A benchmark study of deep learning-based multi-omics data fusion methods for cancer. *Genome Biology*, 23:171, 2022.
6. B. Li, T. Wang, and S. Nabavi. Cancer Molecular Subtype Classification by Graph Convolutional Networks on Multi-Omics Data. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB '21)*, 50, 2021.
7. M. A. Osseni, P. Tossou, F. Laviolette, and J. Corbeil. MOT: A Multi-Omics Transformer for Multiclass Classification Tumour Types Predictions. *BioRxiv*, 2022.11.14.516459, 2022.
8. A. Singh, C. P. Shannon, B. Gautier, F. Rohart, M. Vacher, S. J. Tebbutt, and K.-A. Lê Cao. DIABLO: An integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*, 35(17):3055–3062, 2019.
9. P. Xu, X. Zhu, and D. A. Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
10. K. D. Yang, A. Belyaeva, S. Venkatachalapathy, et al. Multi-domain translation between single-cell imaging and sequencing data using autoencoders. *Nature Communications*, 12:31, 2021.
11. L. Zhao, Q. Dong, C. Luo, Y. Wu, D. Bu, X. Qi, Y. Luo, and Y. Zhao. DeepOmix: A scalable and interpretable multi-omics deep learning framework and application in cancer survival analysis. *Computational and Structural Biotechnology Journal*, 19:2719-2725, 2021.
12. Lamszus, K. and Meyerhof, W. and Westphal, M. Somatostatin and somatostatin receptors in the diagnosis and treatment of gliomas, J Neurooncol, 353:364, 1997.
13. He, J. H., Wang, J., Yang, Y. Z., Chen, Q. X., Liu, L. L., Sun, L., Hu, W. M., Zeng, J., SSTR2 is a prognostic factor and a promising therapeutic target in glioma, Am J Transl Res, 11223:11234, 2021
14. Bielle, F., Di Stefano, A. L., Meyronet, D., Picca, A., Villa, C., Bernier, M., Schmitt, Y., Giry, M., Rousseau, A., Figarella-Branger, D., Maurage, C. A., Uro-Coste, E., Lasorella, A., Iavarone, A., Sanson, M., Mokhtari, K., Diffuse gliomas with FGFR3-TACC3 fusion have characteristic histopathological and molecular features, Brain Pathol, 674:683, 2018.
15. Georgescu, M., Islam, M., Li, Y, Traylor, J, Nanda, A, Novel targetable FGFR2 and FGFR3 alterations in glioblastoma associate with aggressive phenotype and distinct gene expression programs, Acta Neuropathologica Communications, 219:226, 2021.