

DIMAFv2: Interpretable Multimodal Disentangled Representations for Breast Cancer Survival Prediction

Aniek Eijpe^{*1}, Soufyan Lakbir¹, Melis Erdal Cesur², Sara P. Oliveira², Sanne Abeln¹, and Wilson Silva¹

¹AI Technology for Life, Department of Information and Computing Sciences, Department of Biology, Utrecht University, Utrecht, The Netherlands

a.eijpe@uu.nl

²Computational Pathology group, Department of Pathology, The Netherlands Cancer Institute, Amsterdam, The Netherlands

Abstract

Understanding tumor biology and predicting patient survival is challenging due to the complex and heterogeneous information across multiple modalities. We propose an interpretable framework that integrates histopathology whole-slide images and transcriptomics data by explicitly separating modality-shared and modality-specific representations through disentangled attention fusion. Our approach achieves state-of-the-art breast cancer survival prediction and enhanced disentanglement. DIMAFv2's inherent interpretability revealed important factors driving prediction, such as KRAS signaling and G2M checkpoint pathways interacting with tumor morphology.

1 Introduction

Effective cancer survival prediction remains challenging due to the complex, heterogeneous nature of tumor biology [1]. While unimodal models using histopathology Whole Slide Images (WSI) or transcriptomics data have achieved strong results [2–4], multimodal approaches can capture richer biological context and improve prediction performance [5–8].

However, most multimodal approaches focus on the shared information across modalities, potentially suppressing important complementary information [8, 9]. Disentangled Representation Learning (DRL) [10, 11] offers a way to explicitly separate modality-specific and modality-shared information in disentangled representations, also enhancing interpretability. Yet, current DRL-based models rely on linear or network-based disentanglement objectives [8, 12, 13] and rarely quantify or analyze the disentanglement and interpretability achieved.

To address this, we previously proposed **Disentangled and Interpretable Multimodal Attention Fusion (DIMAF)** [9], a framework that separates intra- and inter-modal interactions between WSIs and transcriptomics and promotes disentanglement between the resulting modality-specific and modality-shared representations via a Distance Cor-

relation (DC) loss. Extending this, we introduce **DIMAFv2**, which uses refined WSI representations and an attention-based aggregation within the disentangled representations to improve interpretability and disentanglement. Moreover, we perform an interpretability analysis to study how modality-shared and modality-specific features emerge from unimodal interactions and drive survival outcomes.

2 Methods

We developed an interpretable deep learning model that predicts survival risk by integrating H&E-stained WSIs and bulk RNA-Sequencing (RNA-Seq) data. Following our prior work [9], we create 2 unimodal representations, one containing 50 RNA-seq features and the other 16 WSI features. The RNA-seq features correspond to known pathways with well-defined biological functions [14]. The WSI features represent distinct morphological prototypes, which were annotated by a pathologist, and their cardinality [4, 7]. In DIMAFv2, we modify the WSI encoding to better capture prototype cardinality by adding this directly before multimodal fusion.

The unimodal features are fused in the Disentangled Attention Fusion layer, which models the intra-modal interactions using two self-attention layers and the inter-modal interactions using two cross-attention layers, creating two modality-specific and two modality-shared representations, respectively [9]. In other words, each unimodal feature generates two multimodal features: one enriched with the other modality (modality-shared) and one with the same modality (modality-specific). The multimodal features in the obtained representations are pooled via an attention-based aggregation layer [15], forming the final four disentangled representations. Using these representations and a linear predictor, we obtain a risk score r . The overall loss objective is

$$\mathcal{L} = \mathcal{L}_{surv} + \lambda_{dis} \cdot \mathcal{L}_{dis} \quad (1)$$

where $\mathcal{L}_{surv}$ is the Cox partial log likelihood loss [16, 17], $\lambda_{dis} = 7$ and $\mathcal{L}_{dis}$ the DC-based [18, 19] loss promoting disentanglement between the modality-specific and shared representations [9].

^{*}Corresponding Author.

Table 1. DSS test results on TCGA-BRCA with **best** and **second-best** performances. The p-value corresponds to the log-rank test between the predicted risk cohorts.

	C-index	C-index-IPCW	p-value
PIBD [8]	0.608 ± 0.100	0.497 ± 0.137	$7.95e^{-2}$
MMP [7]	0.741 ± 0.056	0.519 ± 0.185	$9.04e^{-5}$
DIMAF [9]	0.759 ± 0.065	0.580 ± 0.161	$4.08e^{-5}$
DIMAFv2	0.760 ± 0.057	0.553 ± 0.150	$2.80e^{-5}$

Table 2. Disentanglement test results on TCGA-BRCA with **best** and **second-best** performances.

	Distance Correlation	Orthogonal Score
PIBD [8]	0.624 ± 0.021	0.246 ± 0.034
DIMAF [9]	0.356 ± 0.045	0.066 ± 0.009
DIMAFv2	0.333 ± 0.041	0.060 ± 0.016

We trained the model to predict disease-specific survival (DSS) risk using 5-fold cross-validation on the TCGA-BRCA (Breast Invasive Carcinoma) dataset for 30 epochs using the AdamW optimizer.

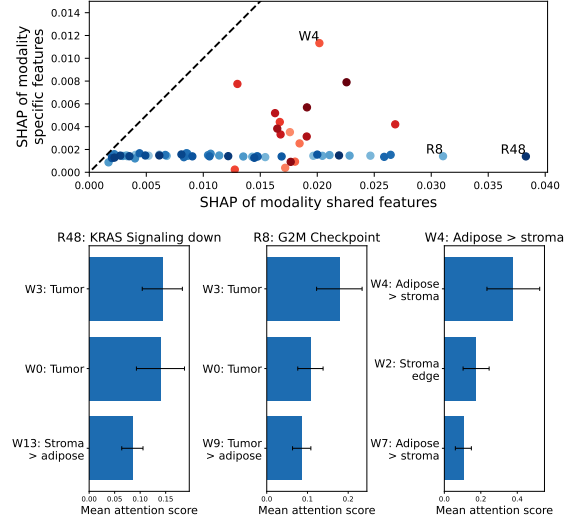
We evaluated the model’s survival prediction performance using the C-index, its weighted variant C-index-IPCW [20], and performed the log-rank test between high and low risk cohorts. Moreover, we assessed total representation disentanglement using DC and an orthogonal score, defined as the absolute cosine similarity. Lastly, we applied DeepSHAP [21] to quantify the overall contribution of modality-shared and modality-specific features, and we visualized the attention weights to reveal the inter- and intra-modal interactions driving survival prediction.

3 Results

Table 1 shows that our proposed framework, DIMAFv2, achieves the highest average c-index and the lowest log-rank p-value. For the C-index-IPCW, DIMAFv2 ranks second, behind DIMAF [9].

Table 2 shows that DIMAFv2 further improves disentanglement between modality-specific and modality-shared representations, achieving a lower DC and orthogonal score than DIMAF. Both substantially improve over PIBD, demonstrating the effectiveness of our disentanglement strategy.

Figure 1 shows how modality-shared and modality-specific features contribute to breast cancer survival prediction. For each unimodal feature (e.g., W4, the WSI prototype annotated as adipose tissue with stroma), we plot the SHAP values of its corresponding modality-shared against its modality-specific feature. Overall, modality-shared features play a more dominant role in driving survival predictions, consistent with previous observations [9]. In particular, the modality-shared KRAS signaling down and G2M checkpoint pathway features are highly pre-

**Figure 1.** Interpretability analysis of DIMAFv2. **Top:** Shows the mean absolute SHAP for the modality-shared versus modality-specific features derived from each pathway feature (blue) and morphological prototype feature (red). Top features in either category are annotated. **Bottom:** Top mean attention weights contributing to the highlighted features, with error bars representing the standard deviation across test samples.

dictive, primarily through interactions with tumor morphological prototypes. The top modality-specific features are non-tumor prototypes, such as W4, arising from interactions with non-tumor prototypes.

4 Discussion

In this work, we introduced DIMAFv2, which integrates WSIs and transcriptomics data through disentangled attention fusion. We demonstrated improved survival prediction performance while enhancing disentanglement between modality-shared and modality-specific representations. The interpretable nature of DIMAFv2 allows us to identify which multimodal features (modality-specific or modality-shared) and interactions drive breast cancer survival prediction. We found that the modality-shared KRAS signaling and G2M checkpoint features, primarily enriched via interactions with tumor prototypes, are important for survival prediction in DIMAFv2. Both pathways are associated with breast cancer subtypes and survival outcomes [22, 23]. Moreover, we found that the modality-specific prototype of adipose tissue with stroma was also important, which is consistent with previous studies suggesting that breast adipose tissue plays a major role in breast cancer risk [24]. Overall, these findings demonstrate that DIMAFv2 can both predict clinical outcomes and provide meaningful insights into multimodal cancer biology.

References

- [1] S. Steyaert, M. Pizurica, D. Nagaraj, P. Khandelwal, T. Hernandez-Boussard, A. J. Gentles, and O. Gevaert. “Multimodal data fusion for cancer biomarker discovery with deep learning”. In: *Nature machine intelligence* 5.4 (2023), pp. 351–362. DOI: [10.1038/s42256-023-00633-5](https://doi.org/10.1038/s42256-023-00633-5).
- [2] J. Yao, X. Zhu, J. Jonnagaddala, N. Hawkins, and J. Huang. “Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks”. In: *Medical image analysis* 65 (2020), p. 101789. DOI: [10.1016/j.media.2020.101789](https://doi.org/10.1016/j.media.2020.101789).
- [3] Á. Nagy, G. Munkácsy, and B. Györfy. “Pan-cancer survival analysis of cancer hallmark genes”. In: *Scientific reports* 11.1 (2021), p. 6047. DOI: [10.1038/s41598-021-84787-5](https://doi.org/10.1038/s41598-021-84787-5).
- [4] A. H. Song, R. J. Chen, T. Ding, D. F. Williamson, G. Jaume, and F. Mahmood. “Morphological prototyping for unsupervised slide representation learning in computational pathology”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 11566–11578. DOI: [10.1109/CVPR52733.2024.01099](https://doi.org/10.1109/CVPR52733.2024.01099).
- [5] R. J. Chen, M. Y. Lu, J. Wang, D. F. Williamson, S. J. Rodig, N. I. Lindeman, and F. Mahmood. “Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis”. In: *IEEE Transactions on Medical Imaging* 41.4 (2020), pp. 757–770. DOI: [10.1109/TMI.2020.3021387](https://doi.org/10.1109/TMI.2020.3021387).
- [6] G. Jaume, A. Vaidya, R. J. Chen, D. F. Williamson, P. P. Liang, and F. Mahmood. “Modeling dense multimodal interactions between biological pathways and histology for survival prediction”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 11579–11590. DOI: [10.1109/CVPR52733.2024.01100](https://doi.org/10.1109/CVPR52733.2024.01100).
- [7] A. H. Song, R. J. Chen, G. Jaume, A. J. Vaidya, A. Baras, and F. Mahmood. “Multimodal Prototyping for cancer survival prediction”. In: *Proceedings of the 41st International Conference on Machine Learning*. PMLR, 2024, pp. 46050–46073. URL: <https://proceedings.mlr.press/v235/song24b.html>.
- [8] Y. Zhang, Y. Xu, J. Chen, F. Xie, and H. Chen. “Prototypical Information Bottlenecking and Disentangling for Multimodal Cancer Survival Prediction”. In: *The Twelfth International Conference on Learning Representations*. 2024. URL: <https://openreview.net/forum?id=otHZ8JA1gh>.
- [9] A. Eijpe, S. Lakbir, M. Erdal Cesur, S. P. Oliveira, S. Abeln, and W. Silva. “Disentangled and Interpretable Multimodal Attention Fusion for Cancer Survival Prediction”. In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2025*. Cham: Springer Nature Switzerland, 2026, pp. 117–127. DOI: [10.1007/978-3-032-05185-1_12](https://doi.org/10.1007/978-3-032-05185-1_12).
- [10] Y. Bengio, A. Courville, and P. Vincent. “Representation learning: A review and new perspectives”. In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828. DOI: [10.1109/TPAMI.2013.50](https://doi.org/10.1109/TPAMI.2013.50).
- [11] I. Higgins, D. Amos, D. Pfau, S. Racaniere, L. Matthey, D. Rezende, and A. Lerchner. “Towards a definition of disentangled representations”. In: *arXiv preprint arXiv:1812.02230* (2018). DOI: [10.48550/arXiv.1812.02230](https://doi.org/10.48550/arXiv.1812.02230).
- [12] X. Wu, Y. Shi, M. Wang, and A. Li. “CAMR: cross-aligned multimodal representation learning for cancer survival prediction”. In: *Bioinformatics* 39.1 (2023). DOI: [10.1093/bioinformatics/btad025](https://doi.org/10.1093/bioinformatics/btad025).
- [13] L. Robinet, A. Berjaoui, Z. Kheil, and E. Cohen-Jonathan Moyal. “DRIM: Learning Disentangled Representations from Incomplete Multimodal Healthcare Data”. In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2024*. Cham: Springer Nature Switzerland, 2024, pp. 163–173. DOI: [10.1007/978-3-031-72384-1_16](https://doi.org/10.1007/978-3-031-72384-1_16).
- [14] A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, and P. Tamayo. “The molecular signatures database hallmark gene set collection”. In: *Cell systems* 1.6 (2015), pp. 417–425. DOI: [10.1016/j.cels.2015.12.004](https://doi.org/10.1016/j.cels.2015.12.004).
- [15] M. Ilse, J. Tomczak, and M. Welling. “Attention-based Deep Multiple Instance Learning”. In: *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 2018, pp. 2127–2136. URL: <https://proceedings.mlr.press/v80/ilse18a.html>.
- [16] D. R. Cox. “Regression models and life-tables”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2 (1972), pp. 187–202. DOI: [10.1111/j.2517-6161.1972.tb00899.x](https://doi.org/10.1111/j.2517-6161.1972.tb00899.x).
- [17] W. H. Wong. “Theory of partial likelihood”. In: *The Annals of statistics* 14.1 (1986), pp. 88–123. DOI: [10.1214/aos/1176349844](https://doi.org/10.1214/aos/1176349844).

- [18] G. J. Székely, M. L. Rizzo, and N. K. Bakirov. “Measuring and testing dependence by correlation of distances”. In: *The Annals of Statistics* 35.6 (2007), pp. 2769–2794. DOI: [10.1214/009053607000000505](https://doi.org/10.1214/009053607000000505).
- [19] X. Liu, S. Thermos, G. Valvano, A. Chartias, A. O’Neil, and S. A. Tsaftaris. “Measuring the Biases and Effectiveness of Content-Style Disentanglement”. In: *Proc. of the British Machine Vision Conference (BMVC)*. 2021. URL: <https://www.bmvc2021-virtualconference.com/assets/papers/0094.pdf>.
- [20] H. Uno, T. Cai, M. J. Pencina, R. B. D’Agostino, and L.-J. Wei. “On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data”. In: *Statistics in medicine* 30.10 (2011), pp. 1105–1117. DOI: [10.1002/sim.4154](https://doi.org/10.1002/sim.4154).
- [21] S. M. Lundberg and S.-I. Lee. “A Unified Approach to Interpreting Model Predictions”. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017, pp. 4765–4774. URL: <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>.
- [22] J. J. Bower, L. D. Vance, M. Psioda, S. L. Smith-Roe, D. A. Simpson, J. G. Ibrahim, K. A. Hoadley, C. M. Perou, and W. K. Kaufmann. “Patterns of cell cycle checkpoint deregulation associated with intrinsic molecular subtypes of human breast cancer cells”. In: *NPJ breast cancer* 3.1 (2017), p. 9. DOI: [10.1038/s41523-017-0009-7](https://doi.org/10.1038/s41523-017-0009-7).
- [23] Y. Tokumaru, M. Oshi, E. Katsuta, L. Yan, V. Satyananda, N. Matsuhashi, M. Futamura, Y. Akao, K. Yoshida, and K. Takabe. “KRAS signaling enriched triple negative breast cancer is associated with favorable tumor immune microenvironment and better survival”. In: *American journal of cancer research* 10.3 (2020), p. 897. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC7136911/>.
- [24] Y.-Y. Wang, C. Lehuédé, V. Laurent, B. Dirat, S. Dauvillier, L. Bochet, S. Le Gonidec, G. Escourrou, P. Valet, and C. Muller. “Adipose tissue and breast epithelial cells: a dangerous dynamic duo in breast cancer”. In: *Cancer letters* 324.2 (2012), pp. 142–151. DOI: [10.1016/j.canlet.2012.05.019](https://doi.org/10.1016/j.canlet.2012.05.019).