# Disentangled and Interpretable Multimodal Attention Fusion for Cancer Survival Prediction

Aniek Eijpe[1][0009−0009−7785−8885], Soufyan Lakbir[1][0000−0002−8521−4408], Melis Erdal Cesur[2][0000−0001−8841−1768], Sara P. Oliveira[2][0000−0002−6586−9079], Sanne Abeln[1][0000−0002−2779−7174], and Wilson Silva[1][0000−0002−4080−9328]

[1] AI Technology for Life, Department of Information and Computing Sciences, Department of Biology, Utrecht University, Utrecht, The Netherlands
a.eijpe@uu.nl
[2] Computational Pathology group, Department of Pathology, The Netherlands Cancer Institute, Amsterdam, The Netherlands

**Abstract.** To improve the prediction of cancer survival using whole-slide images and transcriptomics data, it is crucial to capture both modality-shared and modality-specific information. However, multimodal frameworks often entangle these representations, limiting interpretability and potentially suppressing discriminative features. To address this, we propose Disentangled and Interpretable Multimodal Attention Fusion (DIMAF), a multimodal framework that separates the intra- and inter-modal interactions within an attention-based fusion mechanism to learn distinct modality-specific and modality-shared representations. We introduce a loss based on Distance Correlation to promote disentanglement between these representations and integrate Shapley additive explanations to assess their relative contributions to survival prediction. We evaluate DIMAF on four public cancer survival datasets, achieving a relative average improvement of 1.85% in performance and 23.7% in disentanglement compared to current state-of-the-art multimodal models. Beyond improved performance, our interpretable framework enables a deeper exploration of the underlying interactions between and within modalities in cancer biology. Code and checkpoints are publicly available at: https://github.com/Trustworthy-AI-UU-NKI/DIMAF.

**Keywords:** Cancer survival prediction · Disentangled representation learning · Multimodal fusion · Interpretability in AI.

## 1 Introduction

Integrating multimodal cancer data offers a promising approach for enhancing deep learning-based survival prediction and personalized cancer care. Each modality can provide complementary insights into tumor biology, and a single data modality might therefore not be sufficient to fully capture the heterogeneity and complexity of cancer [26]. For example, genomics can reveal genetic mutations driving tumor development, while medical images can capture tissue

alterations. By combining modalities, we can leverage the strengths of each data source to improve predictive accuracy [2, 26], aligning with clinical practice.

Among multimodal strategies, integrating omics data with Whole Slide Images (WSIs) has gained increasing attention [21, 5, 6, 13]. Early multimodal approaches typically opted for simple mechanisms, like concatenation [21], or a gated Kronecker product [5]. However, these mechanisms generally fail to capture complex interactions between omics data and WSIs [6]. Recent advances address this through multilevel fusion, as seen in HBFSurv [15], or attention-based methods [6, 30]. Among the latest attention-based frameworks, SurvPath [13] processes bulk transcriptomics data with biological pathways [16] to generate semantically meaningful representations and fuses them with WSI patch features using a memory-efficient transformer. Building on this, MMP [25] summarizes the WSIs with Gaussian Mixture Models (GMMs) [24], paving the way for full transformer-based fusion and post-hoc interpretability analysis.

Despite significant advances, multimodal methods focus primarily on shared information across modalities [13, 25, 30, 6], leading to the suppression of modality-specific information [31]. While shared information can enhance robustness [26], information specific to each modality can provide additional discriminative power. To address this, Disentangled Representation Learning (DRL) [1] can be used to decompose the input into modality-shared and modality-specific representations. Recent works have explored this, promoting disentanglement with minimizing an orthogonal [29] or a contrastive loss [23], and using adversarial training [29, 23]. Similarly, PIBD [31] promotes disentanglement by minimizing a contrastive log-ratio upper bound of mutual information [7]. However, none of these methods have explicitly measured disentanglement or assessed the contribution of the modality-specific and modality-shared representations to survival prediction.

In this work, we introduce **Disentangled and Interpretable Multimodal Attention Fusion (DIMAF)**, an inherently interpretable framework that disentangles the fusion between WSIs and bulk transcriptomics for cancer survival prediction. Specifically, we use two self-attention layers to capture the intra-modal interactions and two cross-attention layers to model the inter-modal interactions, resulting in separate modality-specific and modality-shared representations. Unlike MMP [25], we disentangle the multimodal interactions to provide a structured, disentangled view of multimodal information, enhancing interpretability and eliminating the need for post-attention FNNs. Measuring and enforcing disentanglement between latent representations without the ground truth is challenging and remains underexplored [3]. To address this, we employ Distance Correlation (DC) [27, 18], a statistical metric measuring the dependence between two representations. DC does not require pre-defined kernels, as the Hilbert-Schmidt independence criterion [11] does, and captures both linear and non-linear dependencies, unlike orthogonality constraints. We assess the interpretability of our framework by having a pathologist evaluate the learned WSI representations and use SHapley Additive exPlanations (SHAP) [20], a feature attribution method that estimates the marginal contribution of each feature to

the model's output, to quantify the relative contributions of the modality-specific and modality-shared representations.

We evaluate DIMAF on four public cancer survival datasets, demonstrating improved performance and disentanglement over state-of-the-art multimodal models. Beyond improved performance, our framework provides a strong foundation for deeper exploration of multimodal cancer biology. In summary, our work makes the following key contributions:

- We introduce **DIMAF**, an interpretable multimodal cancer survival model that explicitly disentangles the intra- and inter-modal interactions between WSIs and transcriptomics data within an attention-based fusion mechanism.
- To promote disentanglement and learn distinct modality-specific and modality-shared representations, we incorporate a disentanglement loss based on DC.
- We employ SHAP to gain insights into the importance of modality-shared and modality-specific representations.
- We evaluate DIMAF on four public cancer survival datasets, demonstrating enhanced performance, disentanglement, and interpretability.

## 2 Methodology

We propose a disentangled approach for survival analysis that leverages multimodal data. Specifically, for each patient $i$, the objective is to obtain a risk prediction $r_i$, given their WSI $\mathbf{H_i}$ and transcriptomics data $\mathbf{g_i}$. Figure 1 provides an overview of our framework, consisting of three main components: unimodal feature extraction, disentangled attention fusion, and survival prediction.

### 2.1 Unimodal feature extraction

We process the transcriptomics data and WSIs separately to obtain interpretable unimodal representations. Following prior methods [31, 25], we tokenize the transcriptomics data $\mathbf{g_i} \in \mathbb{R}^{D_g}$, representing expression levels of $D_g$ different genes, by using $N_g$ biological pathways as interpretable features. Each pathway represents a group of genes active in specific biological processes in cancer [16]. The tokenized gene expressions are processed through separate Self-Normalizing Networks (SNNs) [14] ($f_{g,\{1,..,N_g\}}$) and $N_g$ random-initialized and learnable pathway encodings [25] are concatenated to each of the obtained pathway embeddings. Lastly, these pathway embeddings with the concatenated encodings are stacked (S) to obtain the final pathway embedding $\mathbf{Z_{g,i}} \in \mathbb{R}^{N_g \times D_{gh}}$.

The WSI $\mathbf{H_i}$ is split into $N_{hi}$ non-overlapping patches, from which features are extracted using a frozen encoder $f_{hp}$. To obtain a slide representation from these patch features, we use PANTHER [24, 25]. Specifically, all the patch features $\{\mathbf{z_{i,j}}\}_{j=1}^{N_{hi}}$ are fitted to a GMM, where $\{\pi_c, \boldsymbol{\mu_c}, \Sigma_c\}_{c=1}^{N_h}$ denote the learned importance, mean, and variance of each Gaussian component $c$, respectively. The importance $\pi_c$ and mean $\boldsymbol{\mu_c}$ of each component $c$ are concatenated, passed through an MLP, concatenated with a learnable prototype encoding and stacked

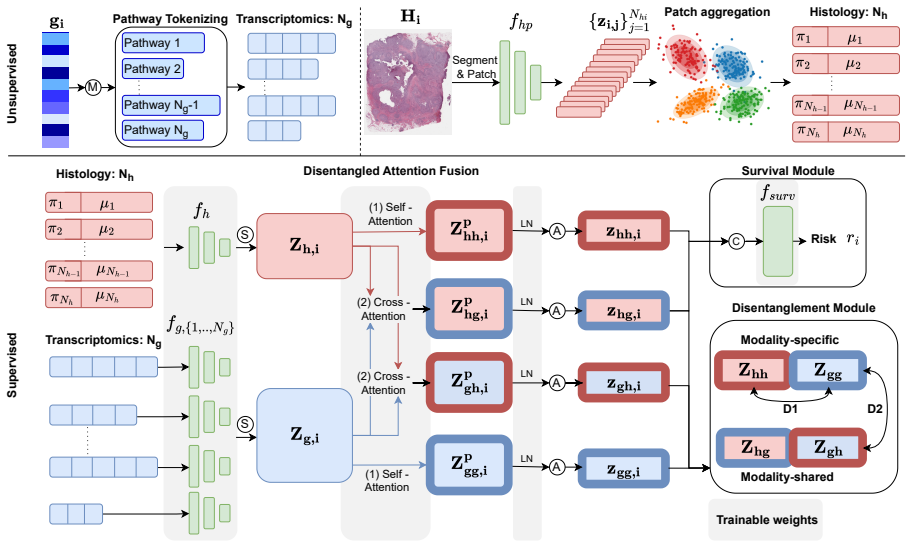


**Fig. 1.** Overview of the proposed framework. First, the transcriptomics data $\mathbf{g_i}$ is processed through pathway tokenizing and $f_{g,\{1,..,N_g\}}$ to obtain the final pathway embedding $\mathbf{Z_{g,i}}$. From the WSI $\mathbf{H_i}$, we obtain patch features $\{\mathbf{z_{i,j}}\}_{j=1}^{N_{hi}}$, which are aggregated and processed through $f_h$ to obtain the slide embedding $\mathbf{Z_{h,i}}$. Then, $\mathbf{Z_{g,i}}$ and $\mathbf{Z_{h,i}}$ are processed through the Disentangled Attention Fusion module, obtaining $\mathbf{z_{hh,i}}, \mathbf{z_{hg,i}}, \mathbf{z_{gh,i}}$, and $\mathbf{z_{gg,i}}$. These representations are further disentangled in the Disentanglement Module and used to compute the survival risk $r_i$ in the Survival Module.

to obtain the WSI embedding $\mathbf{Z_{h,i}} \in \mathbb{R}^{N_h \times D_{gh}}$. Intuitively, the mean $\boldsymbol{\mu_c}$ of each Gaussian represents a morphological prototype, while the importance $\pi_c$ reflects the proportion of patches in the WSI associated with that prototype. To interpret the prototypes, we can visualize the patches $\mathbf{z_{i,j}}$ for which component $c$ has the highest probability, determined by the posterior distribution $q(c|\mathbf{z_{i,j}})$ [24, 25].

### 2.2 Disentangled attention fusion

With the obtained unimodal representations $\mathbf{Z_{h,i}}$ and $\mathbf{Z_{g,i}}$, we perform multimodal fusion using a disentangled attention mechanism. Within this mechanism, we distinctly model the intra-modal interactions by using two self-attention layers to encode the modality-specific information of the transcriptomics data and WSI data within $\mathbf{Z^p_{gg,i}} \in \mathbb{R}^{N_g \times D_z}$ and $\mathbf{Z^p_{hh,i}} \in \mathbb{R}^{N_h \times D_z}$, respectively (Eq. 1).

$$\mathbf{Z^p_{\{gg,hh\},i}} = \sigma\left(\frac{(\mathbf{Z_{\{g,h\},i}}\mathbf{W^{\{g,h\}}_{Q1}})(\mathbf{Z_{\{g,h\},i}}\mathbf{W^{\{g,h\}}_{K1}})^T}{\sqrt{d}}\right)(\mathbf{Z_{\{g,h\},i}}\mathbf{W^{\{g,h\}}_{V1}}) \quad (1)$$

Here, d denotes a scaling parameter stabilizing the attention scores. We model the inter-modal interactions to encode the modality-shared information within $\mathbf{Z^p_{hg,i}} \in \mathbb{R}^{N_g \times D_z}$ and $\mathbf{Z^p_{gh,i}} \in \mathbb{R}^{N_h \times D_z}$, by using two cross-attention layers (Eq. 2).

$$\mathbf{Z^p_{\{gh,hg\},i}} = \sigma\left(\frac{(\mathbf{Z_{\{h,g\},i}W^{\{h,g\}}_{Q2}})(\mathbf{Z_{\{g,h\},i}W^{\{g,h\}}_{K2}})^T}{\sqrt{d}}\right)(\mathbf{Z_{\{g,h\},i}W^{\{g,h\}}_{V2}}) \quad (2)$$

On the obtained representations, we apply Layer Normalization (LN) and average over the pathways or GMM components per representation, obtaining the final disentangled multimodal representation $[\mathbf{z_{gg,i}}, \mathbf{z_{hh,i}}, \mathbf{z_{hg,i}}, \mathbf{z_{gh,i}}] \in \mathbb{R}^{4 \cdot D_z}$.

By using different learnable Query, Key, and Value weight matrices ($\mathbf{W_Q}$, $\mathbf{W_K}$, $\mathbf{W_V}$) for the different types of connections (Eq. 1 and 2) and modalities, the model is encouraged to focus on different aspects of the data, facilitating effective disentanglement. However, this capacity alone does not guarantee disentanglement, necessitating an explicit disentanglement loss. We use DC to promote disentanglement between the modality-specific representations (D1 in Figure 1) and between the modality-specific and modality-shared representations (D2).

$$\mathcal{L}_{dis} = DC(\mathbf{Z_{gg}}, \mathbf{Z_{hh}}) + DC([\mathbf{Z_{gg}}, \mathbf{Z_{hh}}], [\mathbf{Z_{hg}}, \mathbf{Z_{gh}}]) \quad (3)$$

$$\mathrm{DC}(\mathbf{Z_1}, \mathbf{Z_2}) = \frac{dCov(\mathbf{Z_1}, \mathbf{Z_2})}{\sqrt{dCov(\mathbf{Z_1}, \mathbf{Z_1})dCov(\mathbf{Z_2}, \mathbf{Z_2})}} \quad (4)$$

where $\mathbf{Z_{gg}}, \mathbf{Z_{hh}}, \mathbf{Z_{hg}}$, and $\mathbf{Z_{gh}} \in \mathbb{R}^{B \times D_z}$ are the $B$ stacked disentangled representations. $dCov(\mathbf{X}, \mathbf{Y})$ denotes the (Euclidean) distance covariance between $\mathbf{X}$ and $\mathbf{Y}$ [27]. Intuitively, $dCov(\mathbf{X}, \mathbf{Y})$ measures dependence by comparing pairwise distances within $\mathbf{X}$ and $\mathbf{Y}$, assessing how variations in one correspond to variations in the other. If two representations are independent, their variations are unrelated, resulting in a small $dCov$. DC normalizes $dCov$ such that $\mathrm{DC} \in [0, 1]$.

### 2.3 Survival Analysis

Finally, for the survival prediction, we utilize the Cox proportional hazards model [8] on the concatenated disentangled representations.

$$h(t|\mathbf{z_{mm,i}}) = h_0(t) \cdot e^{r_i} = h_0(t) \cdot e^{f_{surv}([\mathbf{z_{gg,i}}, \mathbf{z_{hh,i}}, \mathbf{z_{hg,i}}, \mathbf{z_{gh,i}}])} \quad (5)$$

$h_0(t)$ denotes the baseline hazard, $f_{surv}$ is a linear predictor, and $r_i$ is the risk score of patient $i$, which is optimized with the Cox partial log likelihood [28].

$$\mathcal{L}_{surv} = -\sum_{i \in U}\left(r_i - log(\sum_{j \in K_i} e^{r_j})\right) \quad (6)$$

Here, $U$ is the set of uncensored patients and $K_i$ is the set of patients whose time of death or last known follow-up time is after $i$.

# 3 Experiments

## 3.1 Experimental settings

**Dataset** We evaluate our approach on four public cancer datasets from The Cancer Genome Atlas[3] (TCGA): Breast Invasive Carcinoma (BRCA, $n = 868$), Bladder Urothelial Carcinoma (BLCA, $n = 359$), Lung Adenocarcinoma (LUAD, $n = 412$), and Kidney Renal Clear Cell Carcinoma (KIRC, $n = 340$). The transcriptomics data is obtained via the XENA database [10] and includes RNA-seq expression levels that are normalized across all TCGA cohorts. Moreover, Disease-Specific Survival (DSS) is added to the clinical data as survival endpoint [17], as a better estimate of disease status than Overall Survival (OS).

**Baselines** First, we compare our method against multivariate linear Cox regression [9] as a clinical baseline using grade, age, and sex (if available). Moreover, we employ three state-of-the-art multimodal cancer survival models using WSIs and transcriptomics data, which demonstrated superior performance over unimodal models. SurvPath [13] tokenizes the transcriptomics data via biological pathways and approximates transformer-based fusion. MMP [25] summarizes the WSIs with morphological prototypes for full transformer-based fusion with the same pathway data. PIBD [31] uses prototypical information bottlenecking and tries to disentangle modality-specific and modality-shared representations.

**Implementation details** For all multimodal models, WSIs were split into $256 \times 256$ px patches with a resolution of $0.5\mu m$ per pixel using CLAM [19] and the patch features were extracted with UNI [4, 22]. Additionally, the same $N_g = 50$ Hallmark pathways were used [16]. The models were trained for 30 epochs with the AdamW optimizer, a learning rate of $1e^{-4}$, a cosine learning scheduler, and $1e^{-5}$ weight decay. A batch size of 64 was used, except for SurvPath, which required a batch size of 1. The number of mixture distributions $N_h$ was set to 16 for both MMP and DIMAF. The total loss function used for optimization is defined as $\mathcal{L} = \lambda_{surv} \cdot \mathcal{L}_{surv} + \lambda_{dis} \cdot \mathcal{L}_{dis}$, with $\lambda_{\text{surv}} = 1$ and $\lambda_{\text{dis}} = 7$. We performed 5-fold cross-validation to predict DSS, evaluating model performance with the concordance-index (c-index) [12]. To mitigate potential batch effects, train-test splits were stratified by sample site [13]. Additionally, the disentanglement in PIBD and DIMAF was assessed using DC.

## 3.2 Results

**Survival analysis** Table 1 presents the DSS test results across four cancer types. Our proposed DIMAF model achieves the highest average c-index with a relative improvement of 1.85% compared to the best-performing baseline MMP and 12.2% compared to the DRL baseline PIBD. The ablation variant DIMAF$_{\neg \text{dis}}$,

[3] https://portal.gdc.cancer.gov

**Table 1.** DSS test results (c-index) over the different folds (mean ± std). The best and second-best performances are denoted by **bold** and underlined, respectively.

| **Model** | **BRCA** | **BLCA** | **LUAD** | **KIRC** | **Avg.** (↑) |
|---|---|---|---|---|---|
| Clinical | $0.491 \pm 0.113$ | $0.519 \pm 0.067$ | $0.461 \pm 0.057$ | $0.533 \pm 0.083$ | 0.501 |
| PIBD [31] | $0.671 \pm 0.036$ | $0.573 \pm 0.029$ | $0.578 \pm 0.027$ | $0.726 \pm 0.114$ | 0.637 |
| SurvPath [13] | $0.701 \pm 0.017$ | $0.610 \pm 0.055$ | $0.611 \pm 0.060$ | $0.737 \pm 0.100$ | 0.665 |
| MMP [25] | $0.750 \pm 0.052$ | $0.656 \pm 0.046$ | $\mathbf{0.674 \pm 0.038}$ | $0.728 \pm 0.106$ | 0.702 |
| DIMAF$_{\neg \text{dis}}$ | $\mathbf{0.769 \pm 0.044}$ | $\underline{0.675 \pm 0.034}$ | $0.651 \pm 0.043$ | $\underline{0.747 \pm 0.093}$ | $\underline{0.711}$ |
| DIMAF | $\underline{0.759 \pm 0.067}$ | $\mathbf{0.679 \pm 0.043}$ | $\underline{0.669 \pm 0.062}$ | $\mathbf{0.752 \pm 0.092}$ | **0.715** |

**Table 2.** Disentanglement test results (DC) over the different folds (mean ± std). The best total performances are denoted by **bold**.

| **Model** | **Type** | **BRCA** | **BLCA** | **LUAD** | **KIRC** | **Avg.** (↓) |
|---|---|---|---|---|---|---|
| PIBD [31] | D1 | $0.346 \pm 0.011$ | $0.454 \pm 0.033$ | $0.455 \pm 0.038$ | $0.534 \pm 0.021$ | 0.447 |
| | D2 | $0.922 \pm 0.049$ | $0.920 \pm 0.054$ | $0.957 \pm 0.025$ | $0.908 \pm 0.019$ | 0.927 |
| | Total | $0.634 \pm 0.026$ | $0.687 \pm 0.025$ | $0.706 \pm 0.028$ | $\mathbf{0.721 \pm 0.007}$ | 0.687 |
| DIMAF$_{\neg \text{dis}}$ | D1 | $0.505 \pm 0.018$ | $0.641 \pm 0.04$ | $0.602 \pm 0.07$ | $0.600 \pm 0.084$ | 0.587 |
| | D2 | $0.952 \pm 0.017$ | $0.977 \pm 0.003$ | $0.973 \pm 0.005$ | $0.985 \pm 0.006$ | 0.972 |
| | Total | $0.728 \pm 0.015$ | $0.809 \pm 0.02$ | $0.787 \pm 0.036$ | $0.792 \pm 0.043$ | 0.779 |
| DIMAF | D1 | $0.245 \pm 0.042$ | $0.401 \pm 0.029$ | $0.339 \pm 0.046$ | $0.525 \pm 0.089$ | 0.378 |
| | D2 | $0.468 \pm 0.049$ | $0.672 \pm 0.068$ | $0.609 \pm 0.029$ | $0.931 \pm 0.031$ | 0.670 |
| | Total | $\mathbf{0.356 \pm 0.044}$ | $\mathbf{0.537 \pm 0.048}$ | $\mathbf{0.474 \pm 0.018}$ | $0.728 \pm 0.053$ | **0.524** |

from which the disentanglement objective (Eq. 3) is removed, performs comparably in terms of survival prediction. Its impact varies by dataset, indicating that disentanglement affects cancer types in distinct ways. Overall, these results highlight DIMAF's improved performance in integrating multimodal data for survival prediction.

**Disentanglement analysis** Table 2 presents the test results of the disentanglement between the two modality-specific representations (D1 in Figure 1), the disentanglement between the modality-specific and modality-shared representations (D2 in Figure 1), and the total disentanglement (see Eq. 3). DIMAF achieves the lowest average DC for all three types, indicating the highest degree of disentanglement. It substantially outperforms PIBD, the only baseline designed for disentanglement, by a relative average improvement of 23.7%. Notably, DIMAF increases D2 disentanglement by 27.7% over PIBD and 31.1% over DIMAF$_{\neg \text{dis}}$, further supporting DIMAF's effectiveness in obtaining disentangled multimodal representations and highlighting the importance of the disentanglement loss. While some correlation remains, DIMAF shows substantial progress in disentangling modality-specific and modality-shared representations.

**Interpretability** We evaluated the BRCA slide representations by analyzing the morphological patterns captured by the learned prototypes (Section 2.1). A pathologist annotated the patch clusters, confirming that the prototypes capture

**Table 3.** Normalized SHAP values (%) of the disentangled representations (**Repr.**) over the different folds (mean ± std).

| Model | Repr. | BRCA | BLCA | LUAD | KIRC | Avg. |
|---|---|---|---|---|---|---|
| DIMAF$_{\neg \text{dis}}$ | Specific | $0.513 \pm 0.033$ | $0.548 \pm 0.028$ | $0.538 \pm 0.028$ | $0.517 \pm 0.024$ | 0.529 |
| | Shared | $0.487 \pm 0.033$ | $0.452 \pm 0.028$ | $0.462 \pm 0.028$ | $0.483 \pm 0.024$ | 0.471 |
| | $\mathbf{Z^P_{hh}}$ | $0.315 \pm 0.024$ | $0.325 \pm 0.047$ | $0.291 \pm 0.034$ | $0.294 \pm 0.027$ | 0.306 |
| | $\mathbf{Z^P_{gg}}$ | $0.207 \pm 0.040$ | $0.223 \pm 0.028$ | $0.237 \pm 0.017$ | $0.232 \pm 0.032$ | 0.225 |
| | $\mathbf{Z^P_{gh}}$ | $0.188 \pm 0.022$ | $0.209 \pm 0.033$ | $0.218 \pm 0.022$ | $0.201 \pm 0.032$ | 0.204 |
| | $\mathbf{Z^P_{hg}}$ | $0.290 \pm 0.043$ | $0.243 \pm 0.020$ | $0.253 \pm 0.027$ | $0.273 \pm 0.043$ | 0.265 |
| DIMAF | Specific | $0.180 \pm 0.037$ | $0.273 \pm 0.038$ | $0.257 \pm 0.027$ | $0.302 \pm 0.043$ | 0.253 |
| | Shared | $0.820 \pm 0.037$ | $0.727 \pm 0.038$ | $0.743 \pm 0.027$ | $0.698 \pm 0.043$ | 0.747 |
| | $\mathbf{Z^P_{hh}}$ | $0.083 \pm 0.03$ | $0.108 \pm 0.022$ | $0.071 \pm 0.015$ | $0.073 \pm 0.033$ | 0.084 |
| | $\mathbf{Z^P_{gg}}$ | $0.111 \pm 0.024$ | $0.176 \pm 0.028$ | $0.183 \pm 0.023$ | $0.232 \pm 0.031$ | 0.176 |
| | $\mathbf{Z^P_{gh}}$ | $0.300 \pm 0.046$ | $0.322 \pm 0.042$ | $0.357 \pm 0.023$ | $0.298 \pm 0.039$ | 0.319 |
| | $\mathbf{Z^P_{hg}}$ | $0.506 \pm 0.078$ | $0.394 \pm 0.064$ | $0.389 \pm 0.026$ | $0.397 \pm 0.092$ | 0.421 |

distinct morphological structures, including tumor and immune cells, adipose tissue, and stroma, with some overlap within and between prototypes. Furthermore, the structures seen in the clusters were consistent across train and test sets.

Next, we investigated the relative contributions of the modality-specific ($\mathbf{Z^P_{hh}}$ and $\mathbf{Z^P_{gg}}$) and modality-shared ($\mathbf{Z^P_{gh}}$ and $\mathbf{Z^P_{hg}}$) representations obtained by Deep SHAP [20] over all four datasets (see Table 3). Without the disentanglement loss (DIMAF$_{\neg \text{dis}}$), the model relies almost equally on the shared and specific representations. In contrast, DIMAF shifts the focus towards the shared representations, suggesting that many discriminative features are inherently shared but are incorrectly captured in modality-specific representations without explicit disentanglement. However, the observed contribution of the modality-specific representations highlights the importance of explicitly modeling these features to preserve important information. Notably, the contribution of the WSI-specific representation ($\mathbf{Z^P_{hh}}$) decreases by 72.5% when disentanglement is promoted, suggesting that without disentanglement, $\mathbf{Z^P_{hh}}$ contains a substantial amount of shared features. The contribution of the transcriptomics-specific representation ($\mathbf{Z^P_{gg}}$) also decreases with disentanglement, but only by 21.8%. It is important to note here that the dimensions of $\mathbf{Z^P_{gg}}$ and $\mathbf{Z^P_{hh}}$ differ, as the number of pathways does not match the number of GMM elements (i.e., $N_g \neq N_h$). Future analyses into the SHAP values of the various morphological prototypes and pathways could offer deeper insights into the contributions of modality-specific features. Additionally, when comparing with Table 2, a potential correlation emerges between the degree of disentanglement and the shift in SHAP values, i.e., higher disentanglement appears to increase the SHAP values for shared representations. Overall, DIMAF's disentanglement strategy highlights shared representations as the most informative for survival prediction. However, the presence of important features in the specific representations shows the value of explicitly modeling the modality-specific information.

## 4 Conclusion

In this work, we introduced **DIMAF**, an interpretable framework for cancer survival prediction that explicitly disentangles intra- and inter-modal interactions within the fusion of WSIs and transcriptomics data. We demonstrated that DIMAF improves state-of-the-art performance and disentanglement across four public cancer survival datasets. The SHAP-based interpretability analysis revealed the importance of modality-shared representations while confirming that modality-specific features still contribute to survival prediction.

Our framework provides a strong foundation for further interpretability analysis, enabling a more comprehensive understanding of multimodal cancer biology. Future work can explore the inherent interpretability of DIMAF by combining the attention weights of the intra- and inter-modal interactions with SHAP values before and after fusion to capture both feature importance and multimodal interactions. This dual approach will offer more robust insights into how different data modalities interact in DIMAF for accurate survival prediction.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence **35**(8), 1798–1828 (2013)
2. Boehm, K.M., Khosravi, P., Vanguri, R., Gao, J., Shah, S.P.: Harnessing multimodal data integration to advance precision oncology. Nature Reviews Cancer **22**(2), 114–126 (2022)
3. Carbonneau, M.A., Zaidi, J., Boilard, J., Gagnon, G.: Measuring disentanglement: A review of metrics. IEEE transactions on neural networks and learning systems (2022)
4. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (2024)
5. Chen, R.J., Lu, M.Y., Wang, J., Williamson, D.F., Rodig, S.J., Lindeman, N.I., Mahmood, F.: Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. IEEE Transactions on Medical Imaging **41**(4), 757–770 (2020)
6. Chen, R.J., Lu, M.Y., Weng, W.H., Chen, T.Y., Williamson, D.F., Manz, T., Shady, M., Mahmood, F.: Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4025 (2021)
7. Cheng, P., Hao, W., Dai, S., Liu, J., Gan, Z., Carin, L.: Club: A contrastive log-ratio upper bound of mutual information. In: International conference on machine learning. pp. 1779–1788. PMLR (2020)
8. Cox, D.R.: Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) **34**(2), 187–202 (1972)

9. Davidson-Pilon, C.: lifelines: survival analysis in python. Journal of Open Source Software **4**(40), 1317 (2019)
10. Goldman, M.J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., Banerjee, A., Luo, Y., Rogers, D., Brooks, A.N., et al.: Visualizing and interpreting cancer genomics data via the xena platform. Nature biotechnology **38**(6), 675–678 (2020)
11. Gretton, A., Bousquet, O., Smola, A., Schölkopf, B.: Measuring statistical dependence with hilbert-schmidt norms. In: International conference on algorithmic learning theory. pp. 63–77. Springer (2005)
12. Harrell Jr, F.E., Lee, K.L., Mark, D.B.: Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in medicine **15**(4), 361–387 (1996)
13. Jaume, G., Vaidya, A., Chen, R.J., Williamson, D.F., Liang, P.P., Mahmood, F.: 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. pp. 11579–11590 (2024)
14. Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S.: Self-normalizing neural networks. Advances in neural information processing systems **30** (2017)
15. Li, R., Wu, X., Li, A., Wang, M.: Hfbsurv: hierarchical multimodal fusion with factorized bilinear models for cancer survival prediction. Bioinformatics **38**(9), 2587–2594 (2022)
16. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., Tamayo, P.: The molecular signatures database hallmark gene set collection. Cell systems **1**(6), 417–425 (2015)
17. Liu, J., Lichtenberg, T., Hoadley, K.A., Poisson, L.M., Lazar, A.J., Cherniack, A.D., Kovatich, A.J., Benz, C.C., Levine, D.A., Lee, A.V., et al.: An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell **173**(2), 400–416 (2018)
18. Liu, X., Thermos, S., Valvano, G., Chartsias, A., O'Neil, A., Tsaftaris, S.A.: Measuring the biases and effectiveness of content-style disentanglement. In: Proc. of the British Machine Vision Conference (BMVC) (2021)
19. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**(6), 555–570 (2021)
20. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc. (2017)
21. Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D.A., Barnholtz-Sloan, J.S., Velázquez Vega, J.E., Brat, D.J., Cooper, L.A.: Predicting cancer outcomes from histology and genomics using convolutional networks. Proceedings of the National Academy of Sciences **115**(13), E2970–E2979 (2018)
22. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023)
23. Robinet, L., Berjaoui, A., Kheil, Z., Cohen-Jonathan Moyal, E.: Drim: Learning disentangled representations from incomplete multimodal healthcare data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 163–173. Springer (2024)

24. Song, A.H., Chen, R.J., Ding, T., Williamson, D.F., Jaume, G., Mahmood, F.: Morphological prototyping for unsupervised slide representation learning in computational pathology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11566–11578 (2024)
25. Song, A.H., Chen, R.J., Jaume, G., Vaidya, A.J., Baras, A., Mahmood, F.: Multimodal prototyping for cancer survival prediction. In: Forty-first International Conference on Machine Learning (2024)
26. Steyaert, S., Pizurica, M., Nagaraj, D., Khandelwal, P., Hernandez-Boussard, T., Gentles, A.J., Gevaert, O.: Multimodal data fusion for cancer biomarker discovery with deep learning. Nature machine intelligence **5**(4), 351–362 (2023)
27. Székely, G.J., Rizzo, M.L., Bakirov, N.K.: Measuring and testing dependence by correlation of distances (2007)
28. Wong, W.H.: Theory of partial likelihood. The Annals of statistics pp. 88–123 (1986)
29. Wu, X., Shi, Y., Wang, M., Li, A.: Camr: cross-aligned multimodal representation learning for cancer survival prediction. Bioinformatics **39**(1), btad025 (2023)
30. Xu, Y., Chen, H.: Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21241–21251 (2023)
31. Zhang, Y., Xu, Y., Chen, J., Xie, F., Chen, H.: Prototypical information bottlenecking and disentangling for multimodal cancer survival prediction. In: The Twelfth International Conference on Learning Representations (2024)