# Hybrid Early Fusion for Multi-Modal Biomedical Representations

**Konstantin Hemker**
Department of Computer Science & Technology
Cambridge University
Cambridge, United Kingdom
konstantin.hemker@cl.cam.ac.uk

**Nikola Simidjievski**
Department Oncology
Cambridge University
Cambridge, United Kingdom
ns779@cam.ac.uk

**Mateja Jamnik**
Department of Computer Science & Technology
Cambridge University
Cambridge, United Kingdom
mateja.jamnik@cl.cam.ac.uk

## Abstract

Technological advances in medical data collection such as high-resolution histopathology and high-throughput genomic sequencing have contributed to the rising requirement for multi-modal biomedical modelling, specifically for image, tabular, and graph data. Most multi-modal deep learning approaches use modality-specific architectures that are trained separately and cannot capture the crucial cross-modal information that motivates the integration of different data sources. This paper presents the Hybrid Early-fusion Attention Learning Network (HEALNet) – a flexible multi-modal fusion architecture, which: a) preserves modality-specific structural information, b) captures the cross-modal interactions and structural information in a shared latent space, c) can effectively handle missing modalities during training and inference, and d) enables intuitive model inspection by learning on the raw data input instead of opaque embeddings. We conduct multi-modal survival analysis on Whole Slide Images and Multi-omic data on four cancer cohorts of The Cancer Genome Atlas (TCGA). HEALNet achieves state-of-the-art performance, substantially improving over both uni-modal and recent multi-modal baselines, whilst being robust in scenarios with missing modalities.

## 1 Introduction

A key challenge in Multi-Modal Machine Learning (MMML) is *multi-modal fusion* – the integration of heterogeneous data into a unified and informative representation [5] that leads to improved downstream performance, whilst reducing the dimensionality of the data. Especially considering the complex and multi-causal nature of cancer [7], there is an increasing requirement for ML approaches to model different scales within a biological system simultaneously to capture important information about the tumour microenvironment (TME). The utility of multi-modal fusion has

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

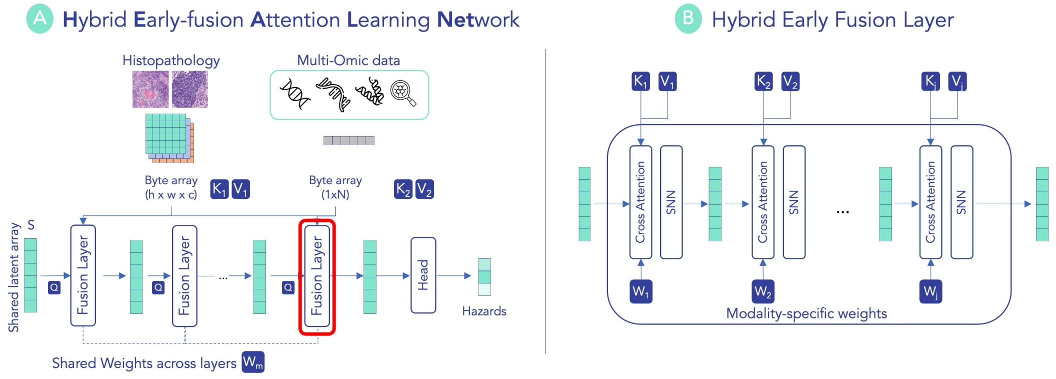

Figure 1: Overview of HEALNet (**H**ybrid **E**arly-fusion **A**ttention **L**earning **Net**work) using both a shared and modality-specific parameter space to learn from structurally heterogeneous data sources in the same model. The shared space is a query array $S$ that is iteratively passed as the query through attention-based fusion layers and captures the shared information between data sources. The hybrid early fusion layer learns the modality-specific attention weights $W_m$, which are shared between layers, and captures structural information of each modality before encoding them and updating the shared space.

also been demonstrated on a variety of cancer data analysis tasks at different scales [10, 6], that commonly rely on combining image (histopathology and/or radiology), tabular data (multi-omics, EHRs) and/or graphs (molecular data). Multi-modal fusion approaches differ in how and when the data is combined, which also determines the capabilities and properties of the resulting model. One common approach is late fusion, which constructs separate models for each modality before combining their output into an ensemble. This allows for capturing salient structural information through modality-specific architectures but prevents the resulting model from learning interactions between modalities [5]. Early fusion methods, on the other hand, tend to train a single model from combined (raw) data (e.g., through concatenation), which incurs the cost of dismissing structural information (spatial, morphological, etc.). More sophisticated multi-modal fusion approaches rely on intermediate fusion, which attempts to overcome this trade-off by learning a low-level representation (embedding) to pick up complex interactions whilst taking advantage of the internal data structure. However, the problem with many intermediate fusion approaches is that the latent representations are not interpretable and struggle to handle missing modalities, yet both of these aspects are a necessity in most biomedical applications. Therefore, we posit that there is a need for more sophisticated early fusion representation learning approaches that: a) preserve structural information of the image, b) learn cross-modal interactions, and c) work on the raw data to preserve meaningful features for improved explainability. We introduce HEALNet to address all of these aspects (Figure 1).

## 2 HEALNet

**Preliminaries.** Let $X^m$ represent data from modality $m = 1, ..., j \in \mathbb{N}$. Let $X^m \in \mathbb{R}^{p \times n}$ be either a tabular dataset with $p$ features and $n$ samples; or an image dataset $X^m \in \mathbb{R}^{h \times w \times c \times n}$ with $n$ images with height $h$, width $w$ and channels $c$. The goal of a multi-modal fusion approach is to learn a fusion function $f()$ such that $y = f(X^1, ..., X^j; \theta)$ where $\theta$ denotes the set of hyperparameters. A conventional design of such a system is to first learn a modality-specific function $g_m()$ which learns an intermediate representation $h^m = g_m(X^m; \phi^m)$ for intermediate hyperparameters $\phi$ and then apply a fusion function $f()$ for predicting the target variable $\hat{y} = f(h^1, ..., h^j; \theta)$.

**Architecture.** Instead of computing $h_m$ and applying a single fusion function $f()$, HEALNet uses an iterative learning setup. Let $t$ denote a step, where the total number of steps $T = d \times m$ for the number of layers $d \in \theta$. Let $S_t$ represent a latent array shared across modalities, initialised at $S_0$ where $S \in \mathbb{R}^{a \times b}$ and $a, b \in \phi$ which is updated at each step. First, instead of learning an intermediate representation $h^m$ as encoded inputs for $X^m$, we compute the attention weights:

$$a_t^m = \alpha(X^m, S_t; \phi^{a_m}) \tag{1}$$

for each modality $m$ at each step $t$. Second, we learn an update function $\psi()$ to be applied at each step. The update of $S$ with modality $m$ is given by $S_{t+1,m} = \psi(S_t, a_t^m; \rho)$ where $\rho$ denotes the shared hyperparameters across $T$. For parameter efficiency, the final implementation uses weight sharing between layers. Across modalities, each early-fusion layer becomes an update function of the form:

$$S_{t+j} = \psi(S_t, a^1, ..., a^j; \rho) \tag{2}$$

The final function for generating a prediction only takes the final state of the shared array and returns the predictions of the target variable:

$$\hat{y} = f(S_T; \theta) \tag{3}$$

Figure 1 depicts a high-level visual representation of this approach, showing: (a) Hybrid Early-fusion Attention Learning Network, and (b) its key component, the early fusion layer (as given in Equation 2). We start by randomly initialising a latent bottleneck array, which is iteratively used as a query into each of the fusion layers and is updated with information from the different modalities at each layer pass. Passing the modalities through the shared latent bottleneck array helps to significantly reduce the dimensionality whilst learning important structural information through the cross-attention layers.

**Preserving structural information.** To handle heterogeneous modalities, we use modality-specific cross-attention layers $\alpha()$ (Figure 1b) and their associated attention weights $a_t^m$, whilst having the latent array $S$ shared between all modalities. We structure the early fusion model as an attention network due to its ability to be generally applicable in different settings, making fewer assumptions about the input data (e.g., compared to a standard convolutional network). Sharing the latent array between modalities allows the model to learn from information across modalities, which is repeatedly passed through the model. Meanwhile, the modality-specific weights between the cross-attention layers focus on learning from inputs of different dimensions as well as learning the implicit structural assumptions of each modality. Specifically, in this work, we use cross-attention as outlined in [8], using the latent array $S$ as the query and the input matrix $X^m$ as the keys and values for each modality. As such, we define the query for each sample as $q^{(n)} = W_q^m S$ and the keys and values as $k^{(n)} = W_k^m x^{(n)}$ and $v^{(n)} = W_v^m x^{(n)}$for all $n \in [1, N]$.

**Handling missing modalities.** Another common challenge in clinical practice is missing data modalities during inference. While models may have been trained on multiple modalities, there is a great chance that only a subset of modalities for a patient is available in practice. Therefore, multi-modal approaches must be robust in such scenarios. Typical intermediate fusion approaches would need to randomly initialise a tensor of the same shape or sample the latent space for a semantically similar replacement to pass into the fusion function $f(h^1, ..., h^j; \theta)$ at inference, which is likely to introduce noise. In contrast, HEALNet overcomes this issue by design: the iterative paradigm can simply skip a modality update step (Equation 2) at inference time in a noise-free manner. Note that these practical benefits also extend to training scenarios, where a (typically small) number of samples is missing some modalities. Rather than imputing this data or completely omitting the samples, HEALNet can train and utilise all the available data using the same update principle.

**High-dimensioanl biomedical data.** One problem with attention-based architectures is their high number of trainable parameters, which we reduced by implementing weight sharing between the layers. Another challenge is that attention-based architectures are commonly trained on very large datasets, while biomedical data is typically high-dimensional with only a few samples. This leads to two problems – computational complexity and training instabilities [2]. To handle the gigapixel scale of whole slide images (WSIs) within computational constraints, we use non-overlapping 224x224 pixel patches on the 20x magnified whole-slide image for preprocessing. To ensure comparability with our baselines [1, 2], we extract a 1024-dimensional feature vector for each patch using a standard ResNet 50 pre-trained on ImageNet-1k V2. While the HEALNet architecture can also achieve competitive performance on the raw patch data, we found the training process to be very resource-intensive due to the high resolution (up to 150,000 x 150,000 pixels).

Table 1: Mean and standard deviation of the concordance Index on four survival risk categories. We report the performance on the hold-out test set across five cross-validation folds. HEALNet outperforms all of its multi-modal baselines and three out of four uni-modal baselines in absolute c-Index performance.

| **Model** | **BLCA** | **BRCA** | **KIRP** | **UCEC** |
|---|---|---|---|---|
| Unimodal (Omics) | 0.606 ± 0.019 | 0.580 ± 0.027 | 0.780 ± 0.035 | 0.550 ± 0.026 |
| Unimodal (WSI) | 0.556 ± 0.039 | 0.550 ± 0.037 | 0.533 ± 0.099 | **0.630** ± 0.028 |
| Porpoise (Late) | 0.620 ± 0.048 | 0.630 ± 0.040 | 0.790 ± 0.041 | 0.590 ± 0.034 |
| MCAT (Interm.) | 0.620 ± 0.040 | 0.589 ± 0.073 | 0.789 ± 0.087 | 0.589 ± 0.062 |
| Perceiver (Early) | 0.565 ± 0.042 | 0.566 ± 0.068 | 0.783 ± 0.135 | 0.623 ± 0.107 |
| HEALNet (ours) | **0.668 ± 0.036** | **0.638 ± 0.073** | **0.812 ± 0.055** | 0.626 ± 0.037 |

## 3 Experiments

This study focuses on survival analysis on The Cancer Genome Atlas (TCGA) data. Concretely, we train a multi-modal model from tissue WSIs, combing them with gene expressions (whole-genome sequencing) and mutations (RNAseq) data, on cohorts from Muscle-Invasive Bladder Cancer (BLCA, n=436), Breast Invasive Carcinoma (BRCA, n=1021), Cervical Kidney Renal Papillary Cell Carcinoma (KIRP, n=284), and Uterine Corpus Endometrial Carcinoma (UCEC, n=538). We compare the results of HEALNet to state-of-the-art late [2], intermediate [1], and early fusion baselines. In line with our benchmark, we use the same survival hazard calculation and survival loss (negative log-likelihood loss). To calculate the patient hazard, we are given the censorship status $c$ and the survival months $T_{cont}$, which are divided into 4 non-overlapping bins for censored patients, and apply the bin cut-offs onto uncensored patients.

## 4 Results & Discussion

The results of the survival analysis are summarised in Table 1, showing the mean and standard c-Index across the 5 cross-validation folds. Across all tested cancer sites, HEALNet learns a relevant unified representation $S$ which allows the model to outperform all multi-modal baselines, achieving state-of-the-art performance in three (out of four) cancer sites. This corresponds to an improvement over multi-modal baselines of approximately 7%, 1%, 3% and 6% on the BLCA, BRCA, KIRP, and UCEC tasks, respectively. Note that the UCEC dataset is an example of *modality dominance*, where all informative signals stem from one modality (in this case WSI), while the other modality is mostly noise with respect to the task.

To put this into context, we compare our results to existing data fusion approaches that focus on image and tabular data for biomedical tasks. Our Porpoise baseline [2] uses a late fusion approach, which trains a modality-specific model for both images (attention-based multiple instance learning (MIL)) and multi-omic data (self-normalising network) before passing the modality representations through an attention gating mechanism. More recently, the Multi-modal Co-Attention Transformer (MCAT) uses two encoders – one "genomic-guided" co-attention followed by a set-based MIL Transformer. The resulting embeddings are then concatenated and passed into a simple classifier [1]. Finally, the Perceiver [4] uses an iterative attention paradigm and achieves highly competitive performance on a range of uni-modal tasks. In line with its original paper, we use concatenation of the input tensors and modality-specific positional encoding to be our early fusion baseline. The problem with Porpoise is that both modalities are entirely learned in isolation, leaving little room for the genomic data to contextualise the imaging modality, which is reflected in the overall c-Index performance. The MCAT baseline does learn a shared representation between both modalities but struggles to handle missing or noisy modalities during training and inference. This can be seen in the performance on the UCEC dataset, where we know that one modality is mostly noise. Since MCAT adds noisy context to its co-attention unit if a modality is missing, this can lead to worse performance than uni-modal baselines.

In contrast, HEALNet overcomes these shortcomings by design. Its end-to-end training allows for a shared latent space that encodes cross-modal interactions while learning modality-specific attention weights that encode structural information. Additionally, the iterative modality-specific updates of the shared representation allow us to easily scale to more than two modalities, and simply skip an update if a modality for a sample is missing without introducing noise (see ablation in Appendix B). These design benefits make HEALNet suitable to handle tasks with high dimensionality (high-resolution whole slide images and 20k+ multi-omic features) but few samples, as is typical in many medical scenarios.

# References

[1] Richard J. Chen, Ming Y. Lu, Wei-Hung Weng, Tiffany Y. Chen, Drew Fk. Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal Co-Attention Transformer for Survival Prediction in Gigapixel Whole Slide Images. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3995–4005, Montreal, QC, Canada, October 2021. IEEE.

[2] Richard J. Chen, Ming Y. Lu, Drew F.K. Williamson, Tiffany Y. Chen, Jana Lipkova, Zahra Noor, Muhammad Shaban, Maha Shady, Mane Williams, Bumjin Joo, and Faisal Mahmood. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40(8):865–878.e6, August 2022. Publisher: Cell Press.

[3] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, 58:101563, December 2019.

[4] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General Perception with Iterative Attention. *ICML*, March 2021. arXiv: 2103.03206.

[5] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions, February 2023. arXiv:2209.03430 [cs].

[6] Evangelina López de Maturana, Lola Alonso, Pablo Alarcón, Isabel Adoración Martín-Antoniano, Silvia Pineda, Lucas Piorno, M. Luz Calle, and Núria Malats. Challenges in the integration of omics and non-omics data. *Genes*, 10(3), 2019.

[7] Sandra Steyaert, Marija Pizurica, Divya Nagaraj, Priya Khandelwal, Tina Hernandez-Boussard, and Olivier Gevaert. Multimodal data fusion for cancer biomarker discovery with deep learning. *Nature Machine Intelligence*, 5(4):351–362, April 2023.

[8] Ashish Vaswani, Google Brain, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. *Advances in Neural Information Processing Systems*, 30, 2017.

[9] Ruchika Verma, Neeraj Kumar, Abhijeet Patil, Nikhil Cherian Kurian, Swapnil Rane, Simon Graham, Quoc Dang Vu, Mieke Zwager, Shan E. Ahmed Raza, Nasir Rajpoot, Xiyi Wu, Huai Chen, Yijie Huang, Lisheng Wang, Hyun Jung, G. Thomas Brown, Yanling Liu, Shuolin Liu, Seyed Alireza Fatemi Jahromi, Ali Asghar Khani, Ehsan Montahaei, Mahdieh Soleymani Baghshah, Hamid Behroozi, Pavel Semkin, Alexandr Rassadin, Prasad Dutande, Romil Lodaya, Ujjwal Baid, Bhakti Baheti, Sanjay Talbar, Amirreza Mahbod, Rupert Ecker, Isabella Ellinger, Zhipeng Luo, Bin Dong, Zhengyu Xu, Yuehan Yao, Shuai Lv, Ming Feng, Kele Xu, Hasib Zunair, Abdessamad Ben Hamza, Steven Smiley, Tang-Kai Yin, Qi-Rui Fang, Shikhar Srivastava, Dwarikanath Mahapatra, Lubomira Trnavska, Hanyun Zhang, Priya Lakshmi Narayanan, Justin Law, Yinyin Yuan, Abhiroop Tejomay, Aditya Mitkari, Dinesh Koka, Vikas Ramachandra, Lata Kini, and Amit Sethi. MoNuSAC2020: A Multi-Organ Nuclei Segmentation and Classification Challenge. *IEEE transactions on medical imaging*, 40(12):3413–3423, December 2021.

[10] Marinka Žitnik, Francis Nguyen, Bo Wang, Jure Leskovec, Anna Goldenberg, and Michael M. Hoffman. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*, 50:71 – 91, 2019.

## A Cross-modal learning

The motivation of using a hybrid-early fusion over a late fusion approach is to enable the model to learn cross-modal interactions that are unavailable to modal-specific models trained in isolation. We can see the effect of this in Figure 2, showing HEALNEt's substantially higher uplift compared to the late fusion benchmark. We note, however, that using a multi-modal model is not always a requirement, especially in the presence of *modality dominance* which we see on the UCEC dataset. However, HEALNet is robust to such cases, achieving comparable performance to the best uni-modal model. Upon further inspection of the HEALNet's omic attention weights on the UCEC task, we found that they barely changed since their initialisation. As such, HEALNet was able to (correctly) inhibit this signal, which is not the case for the other multi-modal baselines where it leads to a loss in performance.

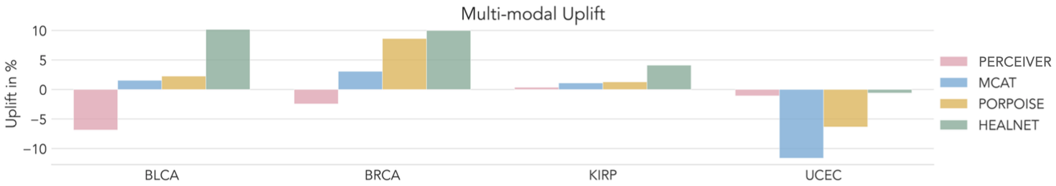

Figure 2: Mean percentage uplift of all multi-modal models compared to the best uni-modal baseline. Across all tested TCGA cancer sites, the *hybrid early fusion* paradigm that HEALNet uses outperforms early, intermediate, and late fusion methods.

## B Missing modality handling

One benefit of using iterative attention is that we can skip updates if modalities are missing at inference time without adding additional noise. For many intermediate fusion methods, missing modalities introduce noise since the fusion function $f()$ expects an intermediate representation $h^m$ for all modalities. This requires initialising a random array or doing a latent search for a similar array to impute the missing portion. A practical approach to this challenge is a late fusion approach, which requires training and keeping several uni-modal alternatives, that can act as a substitution. This, however, can be computationally intensive. HEALNet, on the other hand, overcomes this challenge by design. We believe that this underlines another key benefit of *hybrid early-fusion* – handling mixed missing modalities, at inference time, which takes advantage of multi-modal training, without introducing additional noise.

Table 2: Analysis of the performance of HEALNet in scenarios with missing modalities at inference, compared to uni-modal baselines. Each test sample contains only one of the two modalities. The HEALNet's *hybrid early-fusion* generally achieves a higher average c-Index across all datasets.

| *Test data* | *50% Omic + 50% WSI* | | *100%* |
|---|---|---|---|
| **Dataset** | **XOR (baseline)** | **HEALNet** | **HEALNet** |
| BLCA | 0.547 | 0.612 | 0.668 |
| BRCA | 0.543 | 0.541 | 0.638 |
| KIRP | 0.644 | 0.714 | 0.812 |
| UCEC | 0.533 | 0.580 | 0.626 |

# C Inspections and explanations

Another design benefit of using attention on the raw input data is that it allows for instance-level insights into the model's behaviour, without the need for additional post-hoc explanation methods. Figure 3 shows what parts of the sample the model attends to on average across layers. For images, one can create a high-level heatmap of the cell tissue to highlight relevant regions for more detailed insights on the tumour microenvironment and disease progression. In turn, these regions can be further analysed in post-hoc, such as via nucleus segmentation. To showcase this capability, in Figure 3, we take the highest attention patches and perform nucleus segmentation into epithelial cells, lymphocytes, macrophages, and neutrophils using a HoverNet [3] pre-trained on the MoNuSAC dataset [9]. We acknowledge that attention alone does not provide the entire view of the HEALNet's behaviour, but nevertheless is a helpful capability for model inspection during development.

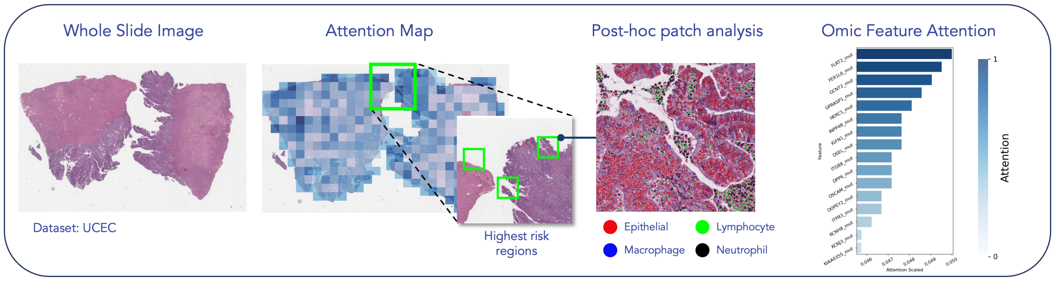

Figure 3: Illustration of model's inspection capabilities using HEALNet on a high-risk patient of the KIRP study. We use the mean modality-specific attention weights across layers to highlight high-risk regions and inspect high-attention omic features. Individual patches can be used for further clinical or computational post-hoc analysis such as nucleus segmentation.