Department of Physics

UNIVERSITY OF CAMBRIDGE
Cavendish Laboratory

Pathomic Fusion Final Report

by

Deyi Zeng

June 2024

Supervised by
Dr Mireia Crispin-Ortuzar, Dr Zeyu Gao

Contents

1 Introduction

Currently, medical diagnosis relies heavily on the experience and the knowledge of physicians, introducing subjectivity and leading to incorrect prediction. With the development of artificial intelligence and machine learning, different types of neural networks are extensively used in making cancer and tumor prognosis and diagnosis, based on the genomic profile and histologic tissues of patients. However, there is limited research related to making predictions by combining the genomic profile and the histologic tissues.

In this report, we reproduce the results in [1] and then compare the performance of uni-modal neural networks and multi-modal fusion networks in making predictions of the survival hazards and classification of WHO grade paradigm. Also, the performance of networks in stratifying the patients survival time is discussed. The Integrated Gradients (IG) [2] and Grad-CAM [3] are utilized to interpret the models.

As an improvement on the original network, instead of splitting patients into multiple samples for predictions and aggregating the prediction results, pre-aggregation of the cell graphs or the histology images by attention is utilized. This can not only capture the relationships between the different histology information of the same patients but also prevent the models from making several predictions for the same patients.

In conclusion, the trimodal fusion network achieves better performance compared to the unimodal network. Moreover, aggregating features before multi-feature fusion further enhances model performance.

2 Related Work

Assigning patients into different risk levels to tailor specific treatment strategies is a common practice in medical research. Survival prediction is widely applied in clinical prognosis [4]. Each patient is assigned a hazard level, allowing for the stratification of patients [5] into different groups that receive varying levels of treatment based on their risk. The WHO glioma grading system [6] also classifies the patients into different levels of risk. Both survival prediction and WHO grades classification are methods to stratifying the patients so that they can receive different treatment strategies based on their prognosis and severity of their condition.

2.1 Unimodal Learning

Three unimodal learning: histology-based convolutional neural network, gene-based feed-forward network, and graph-based graph neural network are discussed.

Recent studies have highlighted the potential of predictive modeling using DNA and RNA sequences [7]. Making predictions based on patients' genotypes is one of the most classical methods. For example, Copy Number Variation (CNV) is related to changes in the number of copies of specific DNA segments in the genome, mutation status is related to alterations in the DNA sequence, Isocitrate Dehydrogenase (IDH) is related to certain types of gliomas, which are significant molecular markers for gliomas. However, traditional biological researches, which focus on studying different combinations of one or a few genes to understand how genes control protein generations, are relatively inefficient in modern science as there are millions of genes which potentially control the expressiveness of diseases in human-beings. Furthermore, traditional statistical models do not perform well in predictive accuracy as they cannot establish complicated correlations between genes. Therefore, using deep neural networks to analyze the statistical relationships between genes and diseases can offer significant guidance to doctors and biologists. Interpretation analysis of models allows biologists to concentrate only on the genes that are statistically significant. The Cox Proportional Hazards Model [8] is one of the most classical models for predicting outcomes using genetic and RNA sequencing data. The deep learning frameworks [9, 10, 11] have been utilized to analyze DNA and RNA sequences to predict survival outcomes.

Recent advances in image analysis techniques have facilitated the use of pathological slides for predictions, harnessing the power of deep learning to improve prediction accuracy [12, 13, 14]. Deep learning frameworks have also found extensive application in lung cancer research [15, 16]. The convolutional neural network (CNN) can implicitly capture the features of pathological slides. It extracts coarse features such as boundaries between diseased and healthy areas, as well as fine details including cellular morphology and nuclear characteristics.

Mohammed Adnan et al. propose a two-stage framework for whole slide image representation using graph neural network [17]. Özen et al. propose transforming histologies into graph data, allowing for the input of histologies with various shapes and sizes [18]. Shen et al. designed a fusion framework that enhances global image-level representations captured by convolutional neural networks with geometric information learned by graph neural networks [19]. Pati et al. proposed Hierarchical Cell-to-Tissue graph representation along with a message-passing graph neural network to make classification on breast dataset [20].

2.2 Multimodal Learning

The previous models primarily explored genomic profiles or pathological slides independently, without integrating different types of data and models to cooperatively make predictions. The reproduction task is based on the [1] which integrates genomic and pathological data by Kronecker Product. The late-fusion technique [21] is a widely used method for integrating different types of data. Researchers typically follow this approach: first, they train separate uni-modal networks using individual data types (e.g., genomic data, imaging data) and extract the latent layers or features from each network. Another method is to train a feature extractor directly. These extracted features capture the low-dimensional representations of each data type. Subsequently, the extracted features from all modalities are combined in a certain way and then used to train a multi-modal network. The genomic details, histopathology images, and clinical details are used to make predictions with gated attentive deep models [22]. Wang et al. addressed the challenge of ROI annotation by designing a multi-modality fusion mechanism applied directly to the entire pathological slides [23]. Shao et al. focus on the feature selections, combining different types of features and making predictions using classical models [24].

3 Dataset and Pre-processing

The dataset used for training and testing the performance of our model are The Cancer Genome Atlas - Glioblastoma Multiforme Lower Grade Glioma (TCGA-GBMLGG) dataset and -Kidney Renal Clear Cell Carcinoma (TCGA-KIRC).

TCGA-GBMLGG (Table 6) contains 769 patients, each with 320 recorded genes and RNA-sequences. Each patient has a varying number of pathological slices, documented as $3 \times 1024 \times 1024$ RGB images. There are totally 1505 pathological slices, also termed as large patches regions of interest (ROIs). To eliminate the randomness when loading the ROIs in fusion step, we generate 9 overlapping small patches ROIs by cropping the large ROIs patches into 9 overlapping small patches with size $3 \times 512 \times 512$. The 2D top-leftmost pixel of cropped small patches are: $(0,0),(0,256),(0,512),(256,0),(256,256),(256,512),(512,0),(512,256),(512,512)$ respectively, as indicated in Figure 1.

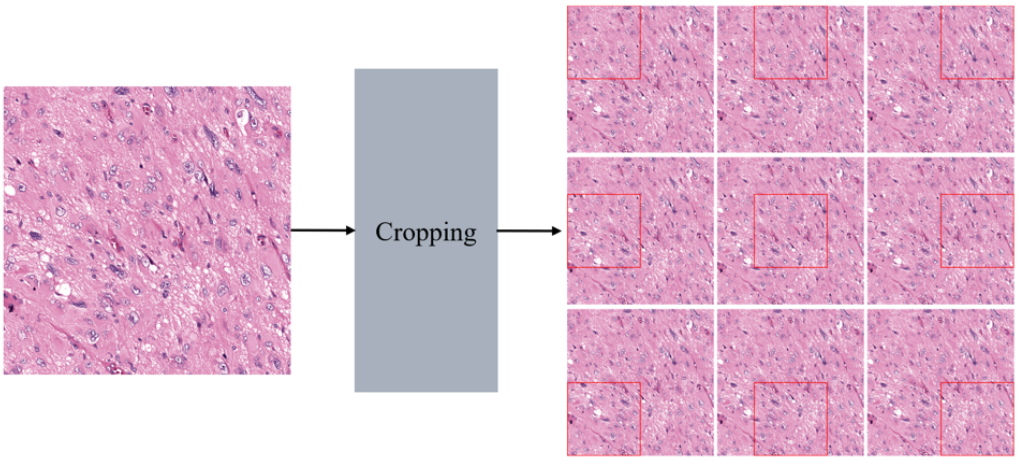

Figure 1: Cropping ROIs for Test and Fusion

TCGA-KIRC (Table 7) contains 414 patients where each patient has 3 ROIs, and another 3 patients TCGA-T7-A92I, TCGA-A3-3365, TCGA-MM-A564 who have 6 ROIs. In total 417 patients and 1260 ROIs and the ROIs are $3 \times 512 \times 521$ RGB images.

For large-patch ROIs in TCGA-GBMLGG and ROIs in TCGA-KIRC, we want to create a cell graph for each ROI (the reason is explained in Section Methodology). We implement nuclei segmentation by using the conditional Generative Adversarial Network (cGAN) [25, 26] to identify cells in the images, which creates the vertices of the cell graph. Then we use K-Nearest Neighbors (KNN) [27] to connect the nearest 5 vertices of each vertex by edges. Then we compute the contour and texture features of each cell to obtain 12-dimensional manual features, followed by utilizing unsupervised contrastive predictive coding (CPC) [28, 29, 30] to extract 1024-dimensional features shared by the surrounding environment of cells. The cell graph explicitly describes the attributes of cells and connections, relationships between cells, and therefore are added into the datasets.

The dataset is split into 15-fold the same as [14] and we use cross-validation for validating the performance of the models.

4 Task

The survival outcome prediction tasks are conducted on the TCGA-GBMLGG and TCGA-KIRC datasets, whereas the WHO grade classification is performed exclusively on the TCGA-GBMLGG dataset.

In clinical scenarios, unlike general machine learning tasks, we do not always predict the target variables directly. Instead, we aim to predict virtual variables that are correlated with the target variables. To evaluate the performance of predictions and models, we measure the coherence between these virtual variables and the target variables.

4.1 Survival Outcome Prediction

4.1.1 Target Variables / Labels

In the dataset, we have two target variables: ***censor*** and ***survival months***.

censor (δ): This variable indicates whether the patient is still alive or not at the time of the last follow-up:

A value of 0 for the ***censor*** means that the patient is still alive (the survival time is censored). A value of 1 for the ***censor*** means that the patient has passed away (the event of interest, which is death, has occurred).

4.1.2 Predicted Variable

We want to design a virtual variable called ***hazard***, which can be coherent with the real ***censor*** and ***survival months*** with the patients.

hazard (S): This variable indicates the hazard of patient to die. Prediction should be in $[-3,3]$.

4.1.3 Evaluation Metrics

The Harrell's Concordance Index (Harrell's C-Index), is adopted to evaluate the prediction of models. Generally speaking, if a patient survives less time than another patient, then this patient is considered to be more hazardous than another patient, therefore has a higher ***hazard***. Therefore, we compare the prediction of hazards between each pair of patients and want to ensure that the patients surviving longer can have lower hazards. Then the definition of C-Index is defined as follows:

$$C\text{-Index} = \frac{\sum_{i<j}\left[\mathbb{I}(T_i > T_j, \delta_j = 1)\mathbb{I}(S_i < S_j) + \mathbb{I}(T_j > T_i, \delta_i = 1)\mathbb{I}(S_j < S_i) + \frac{1}{2}\mathbb{I}(S_i = S_j)\right]}{\sum_{i<j}\left[\mathbb{I}(T_i > T_j, \delta_j = 1) + \mathbb{I}(T_j > T_i, \delta_i = 1)\right]} \tag{1}$$

where:

- T_i and T_j are the ***survival months*** for patients i and j.
- S_i and S_j are the predicted ***hazard*** for patients i and j.
- δ_i and δ_j are the ***censor*** indicators for patients i and j
- $\mathbb{I}(\cdot)$ is the indicator function that equals 1 if the condition inside is true and 0 otherwise.

A higher C-Index implies that the better our model performs.

4.2 WHO Glioma Grade Classification

4.2.1 Target Variable/Label

In the dataset, we have a target variable: ***grade***. The World Health Organization (WHO) classifies gliomas into three grades. Grade II gliomas are characterized by slow growth and relatively normal appearance under the microscope, making them low-risk tumors. Grade III gliomas are aggressive, exhibiting rapid growth and appearing abnormal under microscopic examination. Grade IV gliomas are the most aggressive, developing rapidly and invading neighboring tissues, leading to high mortality rates.

grade (g): A value of 2 for the Grade II glioma, 3 for the Grade III glioma, 4 for Grade IV glioma.

4.2.2 Predicted Variable

We want to design a model which makes an accurate prediction of the grade of each patient. We need to predict the values of three variables ***grade II***, ***grade III***, ***grade IV*** which are the predicted probability of three types of gliomas.

***grade* II** (g_2): Predicted probability of the Grade II glioma.
***grade* III** (g_3): Predicted probability of the Grade III glioma.
***grade* IV** (g_4): Predicted probability of the Grade IV glioma.

4.2.3 Evaluation Metrics

The evaluation metrics includes: micro-average Area Under Curve (AUC), micro-average precision (AP), F1 score (F1), and F1 score of grade IV gliomas only (F1 Grade IV).

5 Methodology: Unimodal Learning

The final objective is to train a multi-modal neural network capable of predicting survival outcomes and WHO glioma grades using genomic profiles, pathological slides (ROIs), and cell graphs as input.

Training from scratch can be time-consuming and inefficient, potentially preventing our large network from finding optimal parameters and leading to suboptimal performance. The first step is to separately train three uni-modal networks: the Self-Normalizing Network (SNN) using genomic data, the Graph Convolutional Neural Network (GCN) using cell graphs, and the Convolutional Neural Network (CNN) using pathological slides (ROIs).

Following this, a gating-based attention mechanism and Kronecker Product are employed to fuse the outputs from the three uni-modal networks. This fusion process facilitates the integration of information across multiple modalities, and therefore enhancing the performance of predictions.

5.1 Convolutional Neural Network (CNN)

The convolutional neural network (CNN) has been widely employed in supervised learning to implicitly extract coarse and fine features within images. However, training a CNN from scratch poses significant challenges for several reasons. Firstly, the network architecture can be highly intricate, comprising numerous parameters, leading to slow forward and backpropagation processes. Secondly, deep neural networks often encounter the issue of vanishing, exploding, or fluctuating gradients, exacerbating the training difficulty. Thirdly, the dataset size is relatively small in comparison to the complexity of the network structure, thus increasing the risk of overfitting.

To mitigate these problems, transfer learning is leveraged. Specifically, pre-trained Visual Geometry Group Network (VGG), which have been well-trained and frozen, are utilized to extract features from histology images. It is important to note that these features are distinct from those used for fusion. Subsequently, a classifier is appended after the VGG network for classification tasks.

The training and testing processes are very special in CNN, and therefore we split the training and testing into two sections.

5.1.1 Training

In terms of TCGA-GBMLGG, the entire network structure for training is displayed on the left hand side in Figure 2. The large-patched ROIs are the input of this procedure, followed by a random cropping which crops the ROIs into a random small patch which is fourth as large as the original one which is subsequently fed into the frozen VGG network to extract features of size 25088. Following this, a feed-forward network (MLP) is employed to further refine these features into 32-length vectors $\boldsymbol{h}_{\text{path}}$, as the features of the ROI. Finally, a classifier is utilized to predict the outcome based on these extracted features. A special design in this network is the use of random cropping. Random cropping of large-patched images allows the network to enhance its robustness and generalization capabilities.

In terms of TCGA-KIRC, the entire network structure for training is depicted on the left-hand side in Figure 3. The structure mirrors that of TCGA-GBMGLL, albeit without the random cropping as the ROIs are already of shape $3 \times 512 \times 512$.

5.1.2 Inference and Testing

In terms of TCGA-GBMLGG, the inference and testing part for CNN is slightly different from the training part. In the training part, large-patched ROIs are randomly cropped into small patches, while this randomness is removed

in predicting and testing part because the predictions are usually not random. Recall that when creating the dataset of TCGA-GBMLGG, non-random small patches of ROIs are created by cropping 9 overlapping patches from the large-patched ROIs Figure 1. These small patches are the input when making predictions and the following network structure is completely the same as the training ones (right hand side of Figure 2).

In terms of TCGA-KIRC, the network used for making predictions (right hand side of Figure 3) is identical to the training part, but all parameters are frozen during prediction.

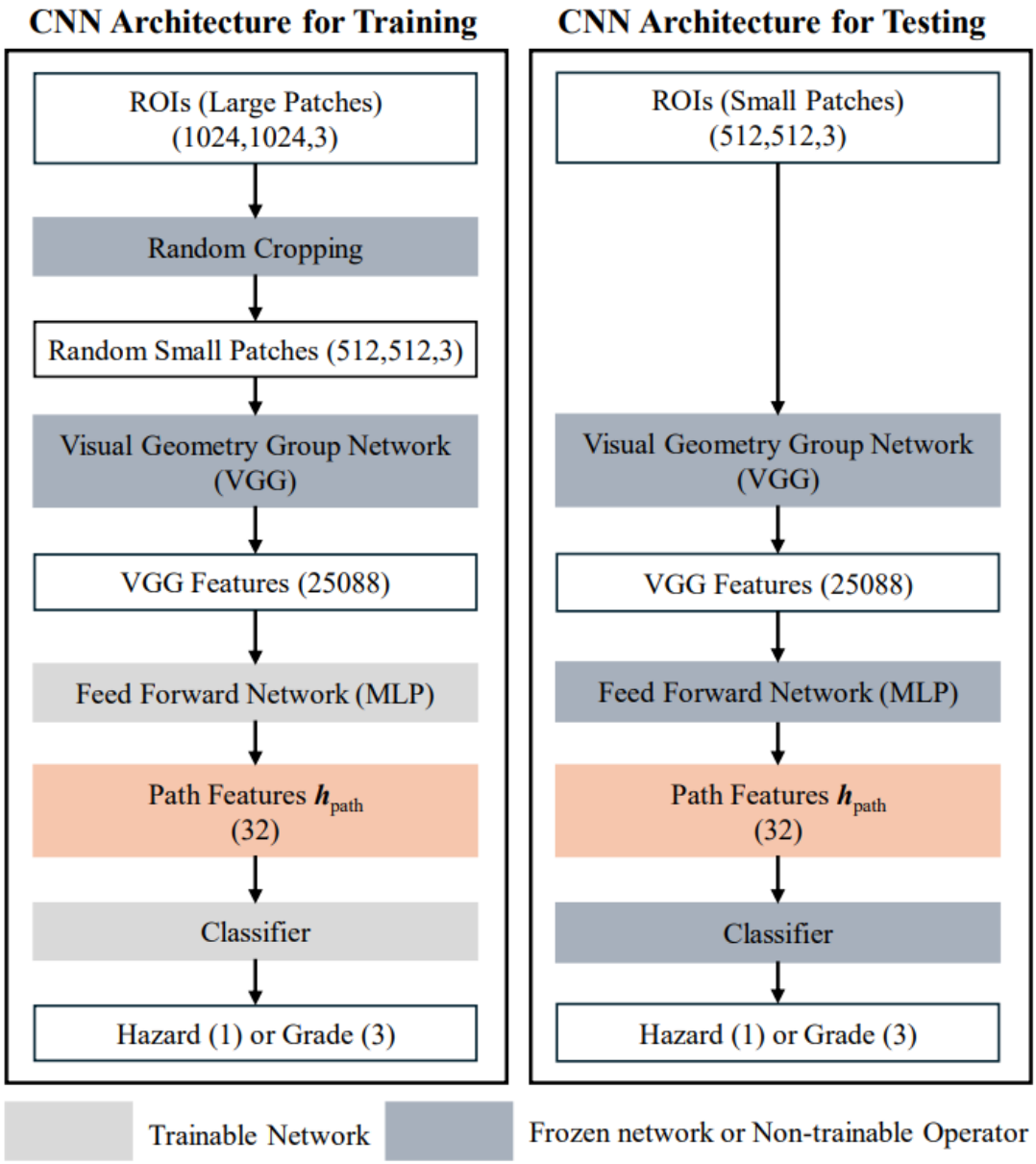

Figure 2: CNN Architecture for TCGA-GBMLGG

5.2 Graph Convolutional Neural Network (GCN)

Although the CNN is employed for predicting hazards and grades using histology images, it is essential to also utilize cell graphs derived from these images. The Pathology-based CNN is an explicit method, which potentially lacks the ability to capture intricate features of individual cells and their complex communities. Compared to CNN, GCN (Figure 4) is an explicit method which leverages prior knowledge such as shape, size, texture, density, and homogeneity of cells and their interactions, which are important markers in glioma and tumor progression.

In terms of feature engineering, valuable features are manually and explicitly extracted, including characteristics of individual cells and local features surrounding them during data preprocessing. Subsequently, a Graph Convolutional Neural Network (GCN) applies these explicit features for predictions.

Considering cells exhibit invasive behaviors and tend to form groups, cluster and communities, leveraging a hierarchical graph structure is beneficial. Therefore, Sample Aggregation Graph Embedding (SAGE) convolutional layers and Self-Attention Graph (SAG) pooling are employed. SAGE layers integrate information from nodes and their neighbors, simulating iterative message passing where node knowledge includes information from all connected nodes. SAG pooling assesses node importance and discards insignificant nodes to accelerate the convergence. Skip connection techniques are integrated into the network architecture by incorporating global mean

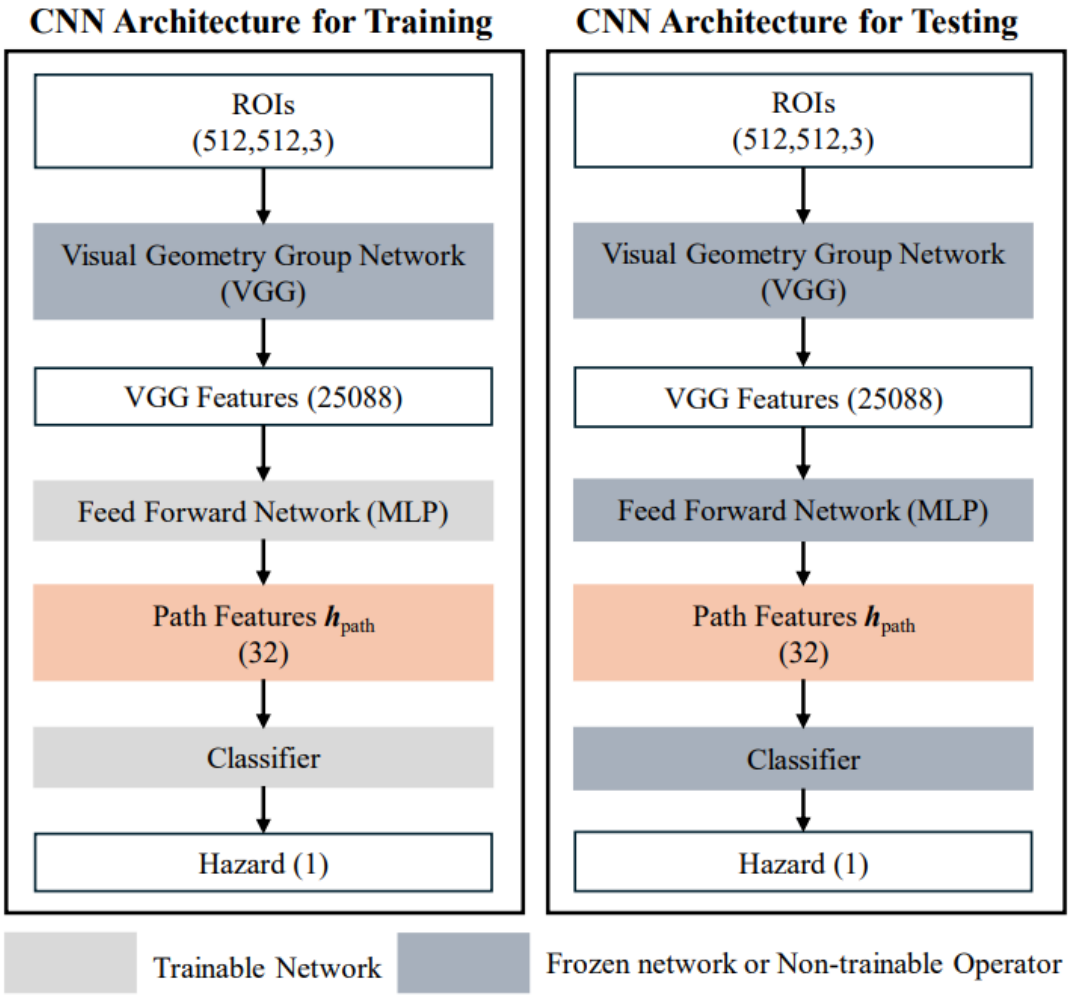

Figure 3: CNN Architecture for TCGA-KIRC

pooling and max pooling after each combination of SAGE and SAG pooling layers, followed by concatenating the global pooling results at the end. This strategy effectively mitigates issues related to unstable gradients during training.

At the end of GCN, graph features are extracted as $\boldsymbol{h}_{\text{graph}}$, followed by a simple and trainable classifier which is used to make predictions.

The whole network is trainable and there is nothing special in training the GCN. The well-trained and frozen network from the training part is used to make inferences (predictions).

5.3 Self-Normalizing Network (SNN)

The self-normalizing network in Figure 5 is applied for predicting patient outcomes based on genomic profiles. Given that genomic profiles are vectors, feed-forward networks are naturally utilized for making predictions. However, during training, overparametrization of feed-forward networks (multi-layer perceptrons) can result in model overfitting, while underparametrization can lead to underfitting. Striking a balance between these extremes is hard. Therefore, employing techniques to mitigate overfitting becomes crucial. In this problem, the self-normalizing neural network is chosen.

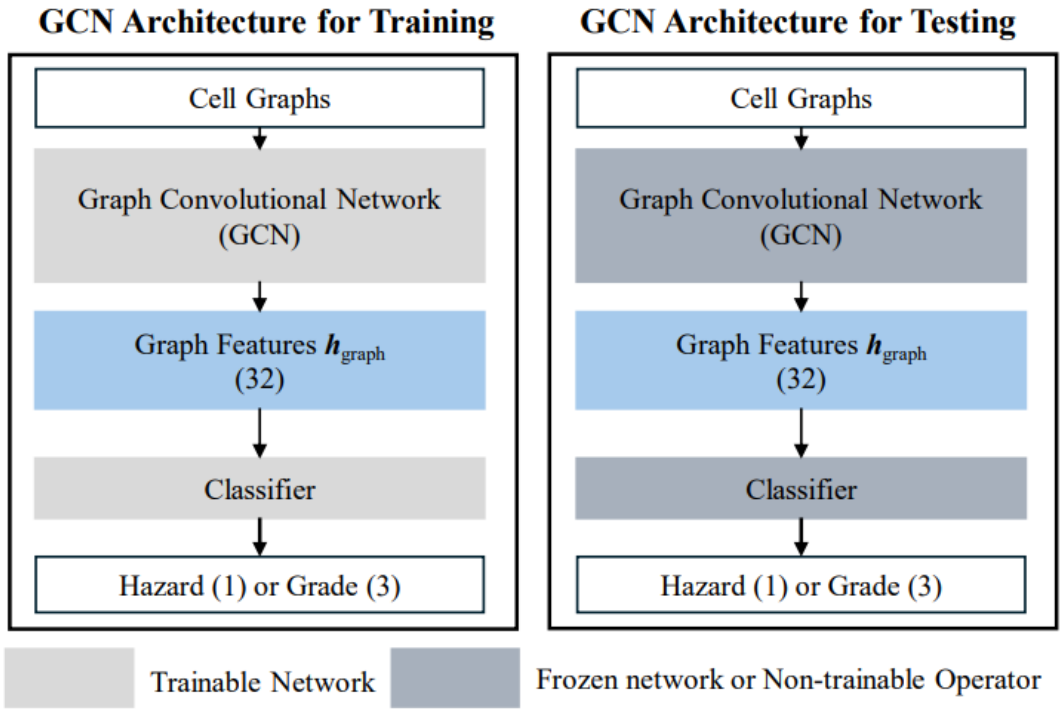

Figure 4: GCN Architecture for TCGA-GBMLGG and TCGA-KIRC

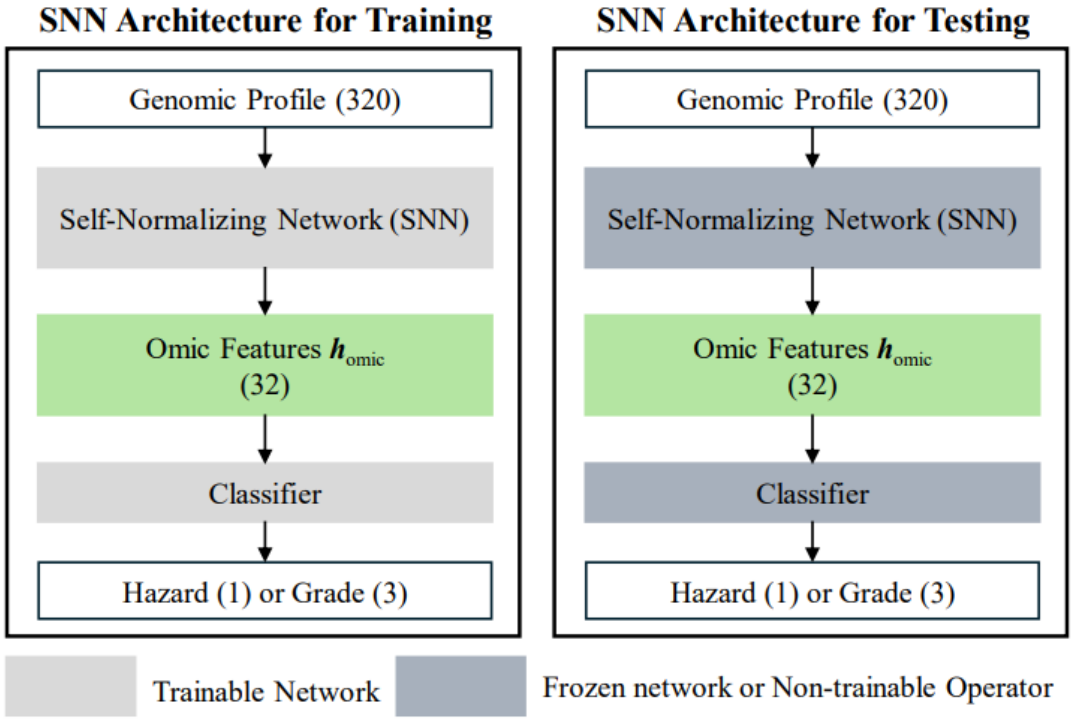

Figure 5: SNN Architecture TCGA-GBMLGG and TCGA-KIRC

The neural network is composed of self-normalizing blocks, each consisting of a linear layer, a scaled exponential linear unit (SeLU) (2), and an Alpha Dropout function. These blocks are concatenated to extract genomic features $\boldsymbol{h}_{\text{omic}}$ from genomic data. Then, a classifier consists of linear and the activation layer is used to make different predictions based on different tasks.

$$\text{ELU}(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha(e^x - 1) & \text{if } x \leq 0 \end{cases} \tag{2}$$

where $\alpha = 1$.

The whole network is trainable and there is nothing special in training the SNN. The well-trained and frozen network from the training part is used to make inferences (predictions).

6 Methodology: Multimodal Learning (Trimodal Fusion Network)

To fully leverage the information embedded in three types of data—histology images, cell graphs, and genomic profiles—we aim to design a network capable of integrating information from these diverse sources. However, we cannot simply average all the data due to the heterogeneity gap among them: pathology ROI images are of shape $3 \times 512 \times 512$, cell graphs comprise node attributes, vertices, and edge attributes, and genomic data have a shape of 320. To bridge the data heterogeneity gap, we implement a gating-based attention mechanism [31], followed by a Kronecker product. This sophisticated approach enables us to explore and exploit the relationships inherent in the multi-modal data, thereby enhancing the performance of model.

6.1 Gate-based Attention Mechanism

We use the gating-based attention mechanism for two reasons: the attention mechanism allows one vector (should be a feature here) to attend to another vector, learning which component of the features is the most informative. Also, the gating mechanism plays a crucial role in reducing the model's sensitivity to noisy or irrelevant data.

$$\boldsymbol{v}_{\text{path}} = \text{ReLU}(\boldsymbol{W}_{\text{path}}\boldsymbol{h}_{\text{path}} + \boldsymbol{b}_{\text{path}}) \tag{3}$$
$$\boldsymbol{v}_{\text{graph}} = \text{ReLU}(\boldsymbol{W}_{\text{graph}}\boldsymbol{h}_{\text{graph}} + \boldsymbol{b}_{\text{graph}}) \tag{4}$$
$$\boldsymbol{a}_{\text{path,omic}} = \text{Sigmoid}(\boldsymbol{h}_{\text{path}}\boldsymbol{W}_{\text{path,omic}}\boldsymbol{h}_{\text{omic}} + \boldsymbol{b}_{\text{path,omic}}) \tag{5}$$
$$\boldsymbol{a}_{\text{graph,omic}} = \text{Sigmoid}(\boldsymbol{h}_{\text{graph}}\boldsymbol{W}_{\text{graph,omic}}\boldsymbol{h}_{\text{omic}} + \boldsymbol{b}_{\text{graph,omic}}) \tag{6}$$
$$\boldsymbol{h}_{\text{path,gated}} = \text{ReLU}(\boldsymbol{W}_{\text{path,gated}}(\boldsymbol{a}_{\text{path,omic}} \odot \boldsymbol{v}_{\text{path}}) + \boldsymbol{b}_{\text{path}}) \tag{7}$$
$$\boldsymbol{h}_{\text{graph,gated}} = \text{ReLU}(\boldsymbol{W}_{\text{graph,gated}}(\boldsymbol{a}_{\text{graph,omic}} \odot \boldsymbol{v}_{\text{graph}}) + \boldsymbol{b}_{\text{graph}}) \tag{8}$$
$$\boldsymbol{h}_{\text{omic,gated}} = \text{ReLU}(\boldsymbol{W}_{\text{omic,gated}}\boldsymbol{h}_{\text{omic}} + \boldsymbol{b}_{\text{omic}}) \tag{9}$$

In gating-based attention, we use the genomic feature to attend the other two features. The equations (3) (4) represent the computation of ***Values***, and (5) (6) represent the computation of ***Attention***, which can be interpreted as the attention weight vector. Instead of calculating ***Queries*** and ***Keys*** respectively, we use the bilinear combination of two vectors, which simplifies the structure but achieves the same function. (7) (8) (9) represent the gating-attended features.

6.2 Kronecker Product Fusion

Instead of constructing a trainable feed-forward neural network and concatenating three types of gated features as input, we explicitly define the relationship between each feature by employing the Kronecker Product.

$$\boldsymbol{h}_{\text{fusion}} = \begin{bmatrix} \boldsymbol{h}_{\text{path,gated}} \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \boldsymbol{h}_{\text{graph,gated}} \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \boldsymbol{h}_{\text{omic,gated}} \\ 1 \end{bmatrix} \tag{10}$$

where $\otimes$ is the outer product, and $\boldsymbol{h}_{\text{fusion}} \in \mathbb{R}^{33\times33\times33}$

A single outer product between the gated pathology features and the gated graph features allows each component in the pathology features to interact (multiply) with each component in the graph features. The resulting combined matrix can then be multiplied with the gated genomic features vector, creating a tri-multiplication that involves each component from the three features vector. To preserve both the unimodal features and the bimodal interaction between features, '1's are appended behind feature vectors.

The primary reason for using Kronecker Product Fusion instead of simply flattening and concatenating features is that Kronecker Product Fusion explicitly captures interactions between features. Flattening and concatenation, on the other hand, rely on the network to implicitly learn these relationships, which can be challenging and may not effectively capture interactions between features.

6.3 Network

The architecture of trimodal fusion network is displayed in Figure 6. The network is built upon three unimodal networks. The CNN network (a combination of VGG and MLP) is frozen for the entire training steps, while GCN and SNN networks are frozen for the first 5 epochs, after which they will be unfrozen for model fine-tuning.

The omic, graph, and path features are input into a gating-based attention network, followed by the application of the Kronecker Product. Subsequently, an encoder flattens and transforms these high-dimensional features into length-64 vectors. Finally, a classifier is employed.

7 Sample Aggregation

As introduced in the dataset section, since a single patient may contain several samples in the dataset and it is insensible to make several predictions on the same patients. An aggregation for predictions of different samples from the same patient is made.

In survival hazard prediction, after making predictions for all samples, the predicted hazards for the same patient is aggregated by mean.

In grade classification, after performing predictions for all samples, the predicted probabilities for each grade class of the same patient are aggregated using the maximum value. Suppose a patient has m predictions:

$\left[g_2^{(1)}, g_3^{(1)}, g_4^{(1)}\right]^T, \left[g_2^{(2)}, g_3^{(2)}, g_4^{(2)}\right]^T, \cdots, \left[g_2^{(m)}, g_3^{(m)}, g_4^{(m)}\right]^T,$

then the aggregation is: $\left[\max_{i=1,2,\cdots,m} g_2^{(i)}, \max_{i=1,2,\cdots,m} g_3^{(i)}, \max_{i=1,2,\cdots,m} g_4^{(i)}\right]^T$

It is important to note that while maximum aggregation can result in a sum of probabilities greater than 1 across all grade classes, it does not affect the computation of evaluation metrics such as AUC, AP, and F1-score. These metrics are calculated based on the ranking of probabilities rather than the values of probabilities themselves.

8 Improvement by Pre-aggregation Trimodal Fusion Network

In the original paper, the Kronecker Product and Gating-based attention mechanisms are used for combining different features. However, combinations of the same features are not considered.

According to the dataset construction, a patient can have several histology ROIs and cell graphs, resulting in multiple predictions for the same patient. Clinically, it is only sensible to provide a single prediction for each patient. To do this, mean and maximum aggregation of multiple samples from the same patient is used in the original design, which is a brute-force method that ignores the correlations between different ROIs during training.

To address this, we designed a network (termed Pre-aggregation) that packs all ROIs for the same patient into one sample, rather than splitting them into several samples as in the original design.

A challenge arises because each patient may have a varying number of cell graphs and ROIs, posing a problem when inputting them into the networks, but fortunately, the attention mechanism [32] allows vectors with different shapes as input. Therefore, the attention-based pre-aggregation is implemented.

8.1 Pre-aggregation Technique

Pre-aggregation is an attention-based network. For example, if a patient has 6 ROIs, then the outputs of CNN networks are calculated to obtain 6 path features of shape 32. These 32 features can be combined into a matrix of shape 32×6 (Figure 7) and then put into the attention network.

8.2 Network

The network structure is displayed in Figure 8, which is constructed by inserting another attention network between GCN/CNN and Gating-based attention to the original trimodal fusion network in Figure 6. The omic SNN network is retained as each patient has one omic feature and no need to be packed up.

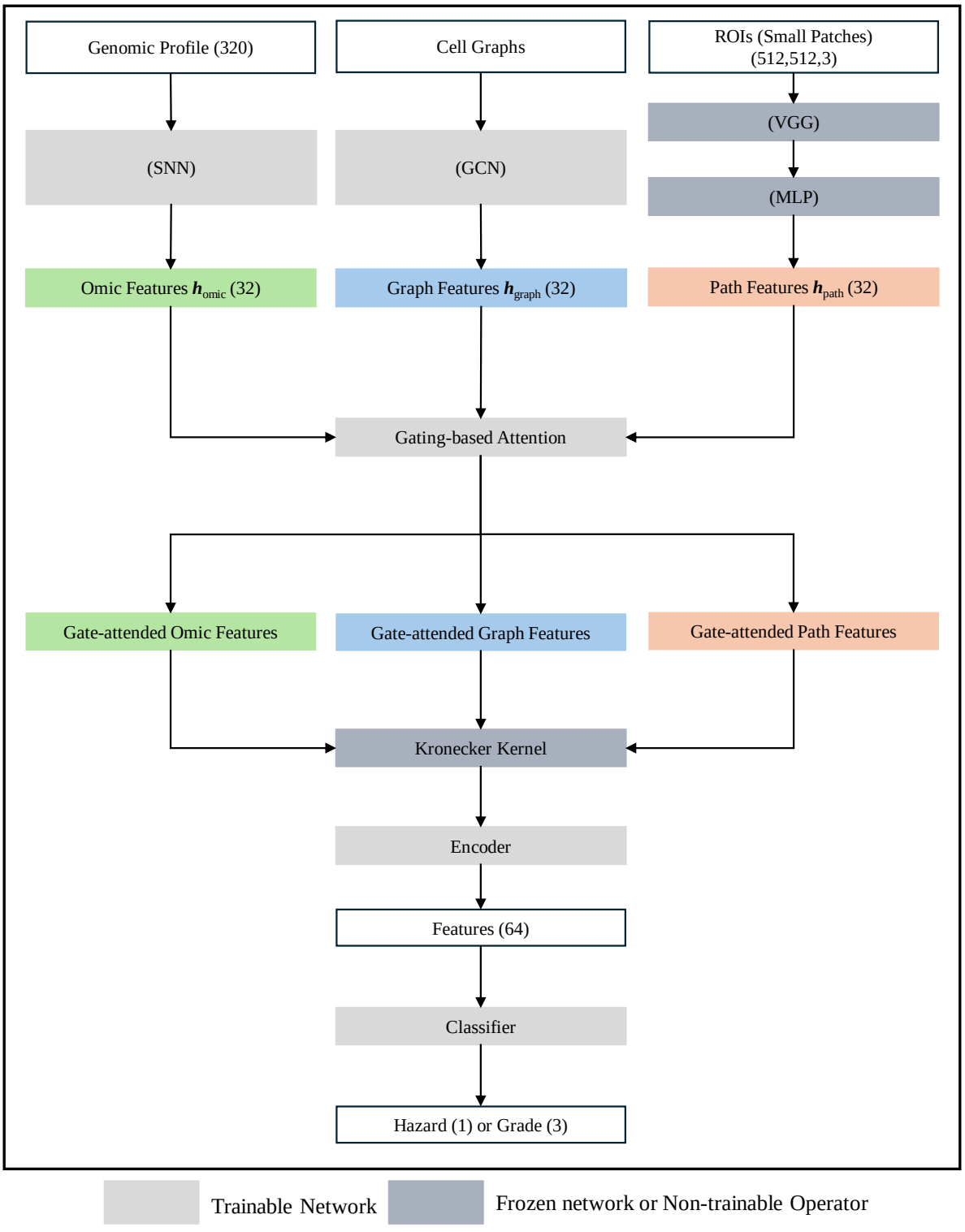

Figure 6: Trimodal Fusion Network for TCGA-GBMLGG and TCGA-KIRC (Training Version), freeze all blocks to make predictions or testing

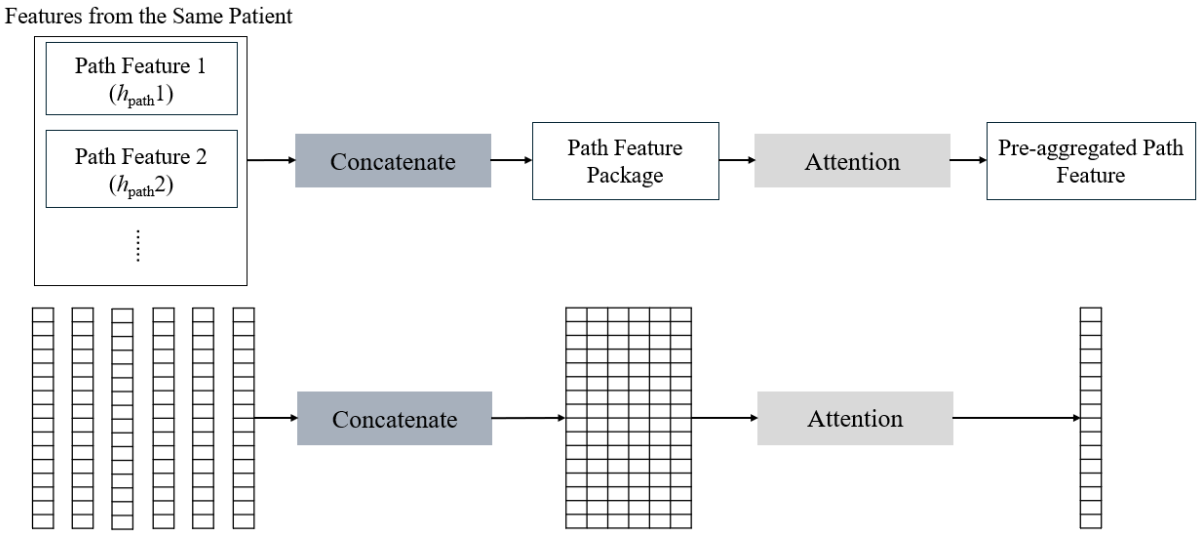

Figure 7: Pre-Aggregation

8.3 Training

Since batch training does not allow samples with different shapes, the samples can be repeatedly selected to construct the batch. The training procedure is described as follows:

- Select B patients for the batch.
- For each patient:
  - Take its genomic profile
  - Calculate the output of SNN to obtain omic features.
  - Randomly choose 5 ROIs, and 5 graph cells ***with replacement*** from this patient.
  - Calculate the output of CNN and GCN to obtain 5 path features and 5 graph features.
  - Concatenate path and graph features into two 5×32 matrices, respectively.
- Concatenate B patients' omic features to form the omic batch.
- Concatenate B patients' path and graph features matrices respectively into two $B \times 5 \times 32$ tensors to form the path and graph batches.
- Train the network with batched features.

The SNN, GCN, and CNN (combination of VGG and MLP) networks are initially frozen. The SNN network is then unfrozen after 5 epochs to allow for fine-tuning. This approach prevents the model from training an overwhelming amount of parameters simultaneously, which facilitates the convergence speed and helps prevent overfitting.

8.4 Inference and Testing

For inference and testing, the whole model is frozen. Additionally, each patient has only one prediction outcome, there is no need to aggregate the prediction by mean or maximum. The inference procedure is as follows:

- Select only 1 patient for a batch.
- Take its genomic profile
- Calculate the output of SNN to obtain omic features.
- Choose all ROIs, and all graph cells from this patient.
- Calculate the outputs of CNN and GCN to obtain all path features and all graph features.
- Concatenate path and graph features into two 5×32 matrices, respectively.
- Calculate the output of Pre-aggregation trimodal fusion network, to obtain prediction outcome.

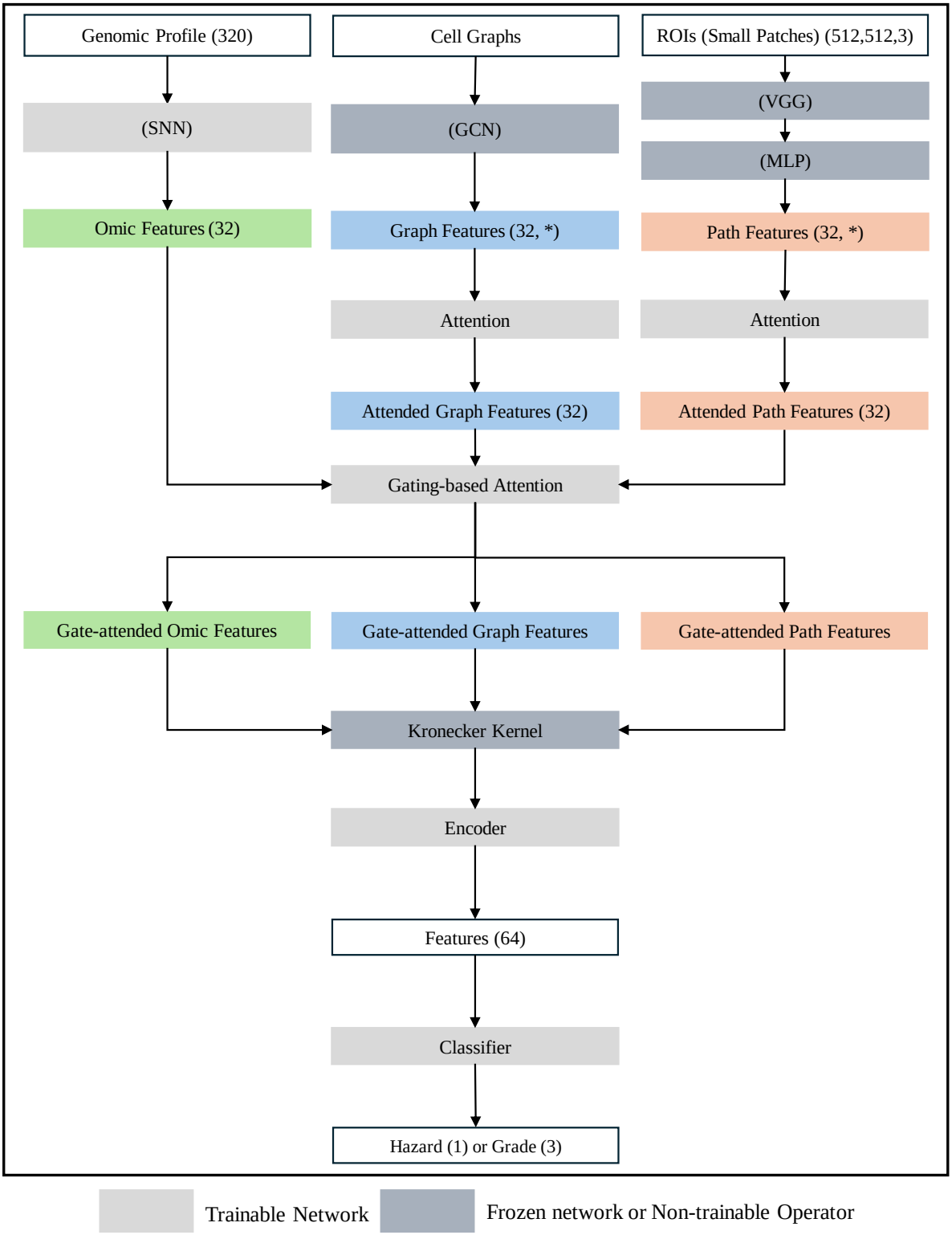

Figure 8: Pre-aggregation Tri-fusion Network (Training Version), freeze all blocks to make predictions or testing

9 Loss Function

Two distinct loss functions are selected for the two tasks. Regardless of the models we are training, we employ the following loss function with varying options of λ. To prevent the deep neural network from overfitting we apply l_1 regularization on the trainable parameters W of the neural networks.

9.1 (Hazard Prediction) Cox Loss with Regularization

$$\text{CoxLoss} = -\frac{1}{n}\sum_{i=1}^{n}\left(s^{(i)} - \log\sum_{j=1}^{n}\exp\left(s^{(j)}\right)\cdot\mathbb{I}(T^{(j)} \geq T^{(i)})\right)\cdot\delta^{(i)} + \lambda||W||_1 \tag{11}$$

where:

- $T^{(i)}$ and $T^{(j)}$ are the ***survival months*** for patients i and j.
- $S^{(i)}$ and $S^{(j)}$ are the predicted ***hazard*** for patients i and j.
- $\delta^{(i)}$ and $\delta^{(j)}$ are the ***censor*** indicators for patients i and j.
- $\mathbb{I}(\cdot)$ is the indicator function that equals 1 if the condition inside is true and 0 otherwise.

9.2 (Grade Prediction) Negative Log Likelihood Loss with Regularization

$$\text{NLLLoss} = -\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{I}\left(g^{(i)}=2\right)\log\left(g_2^{(i)}\right) + \mathbb{I}\left(g^{(i)}=3\right)\log\left(g_3^{(i)}\right) + \mathbb{I}\left(g^{(i)}=4\right)\log\left(g_4^{(i)}\right)\right) + \lambda||W||_1 \tag{12}$$

where:

- ***grade*** (g): A value of 2 for the Grade II glioma, 3 for the Grade III glioma, 4 for Grade IV glioma.
- ***grade* II** (g_2): Predicted probability of the Grade II glioma.
- ***grade* III** (g_3): Predicted probability of the Grade III glioma.
- ***grade* IV** (g_4): Predicted probability of the Grade IV glioma.

10 Result of WHO Glioma Grades Classification

10.1 TCGA-GBMLGG

Table 1 presents the performance metrics for various methods in terms of AUC (Area Under the Curve), AP (Average Precision), F1-score, and F1-score specifically for Grade IV patients.

A first glance reveals that the pathology-based (CNN) model demonstrates better overall prediction performance than the gene-based (SNN) model.

Fusion methods show significant improvement across all metrics, where the trimodal fusion network achieves the highest performance in predicting all grades. The F1-score in the prediction of Grade IV patients improves significantly to 0.9235, which is crucial in clinical settings as these patients often require more aggressive treatment strategies and closer monitoring.

A comparison between the bimodal fusion network and the trimodal fusion network indicates that the improvement in AUC and AP is less significant. This may be because the information from the cell graphs, as a feature extraction from the histology images, likely overlaps with the information provided by the pathology ROIs. The model has already learn enough information from the ROI Images, and the extra extraction from ROIs could make little improvement on predictions.

The ensembling methods (SNN $\otimes$ SNN, GCN $\otimes$ GCN, CNN $\otimes$ CNN) do not significantly improve the performance of the models. This may be due to the complexity of the unimodal network, which is sufficient to leverage the information, rendering the ensembling methods less effective in enhancing performance by increasing network complexity. The conclusion can be drawn that the improvement in Pathomic Fusion results from the abundance of information (Genomic and Histology) rather than increased network complexity.

According to the ROC curve in Figure 9, gene-based SNN network is not as predictive as the pathology-based CNN network. The potential reason is that the WHO glioma grade is obtained through the microscopic examination of histopathology ROIs obtained via biopsy or surgery. The pathologists analyze the specific cellular characteristics to evaluate the aggressiveness of gliomas and the grades [6]. The trimodal fusion network (Pathgraphomic fusion) has the highest AUC in all grade predictions, which proves the success in histology and genomic profile fusion.

Table 1: Performance Metrics of WHO Grade Classification Task in TCGA-GBMLGG

Method	AUC	AP	F1-score	F1 Grade IV
Genomic SNN	0.8522 ± 0.012	0.7287 ± 0.018	0.6503 ± 0.016	0.8541 ± 0.018
Genomic (SNN $\otimes$ SNN)	0.8529 ± 0.012	0.7295 ± 0.019	0.6503 ± 0.018	0.8552 ± 0.017
Histology GCN	0.8471 ± 0.014	0.7627 ± 0.017	0.6493 ± 0.024	0.8148 ± 0.027
Histology (GCN $\otimes$ GCN)	0.8498 ± 0.011	0.7649 ± 0.016	0.6492 ± 0.015	0.8155 ± 0.019
Histology CNN	0.8823 ± 0.007	0.7791 ± 0.020	0.7152 ± 0.022	0.8799 ± 0.017
Histology (CNN $\otimes$ CNN)	0.8875 ± 0.008	0.8083 ± 0.014	0.7176 ± 0.017	0.8734 ± 0.014
Pathomic F. (CNN $\otimes$ SNN)	0.9050 ± 0.009	0.8338 ± 0.016	0.7330 ± 0.019	0.9141 ± 0.013
Pathomic F. (GCN $\otimes$ SNN)	0.9018 ± 0.011	0.8243 ± 0.018	0.7318 ± 0.020	0.9127 ± 0.017
Pathomic F. (CNN $\otimes$ GCN $\otimes$ SNN)	0.9088 ± 0.010	0.8360 ± 0.018	0.7459 ± 0.021	0.9235 ± 0.013

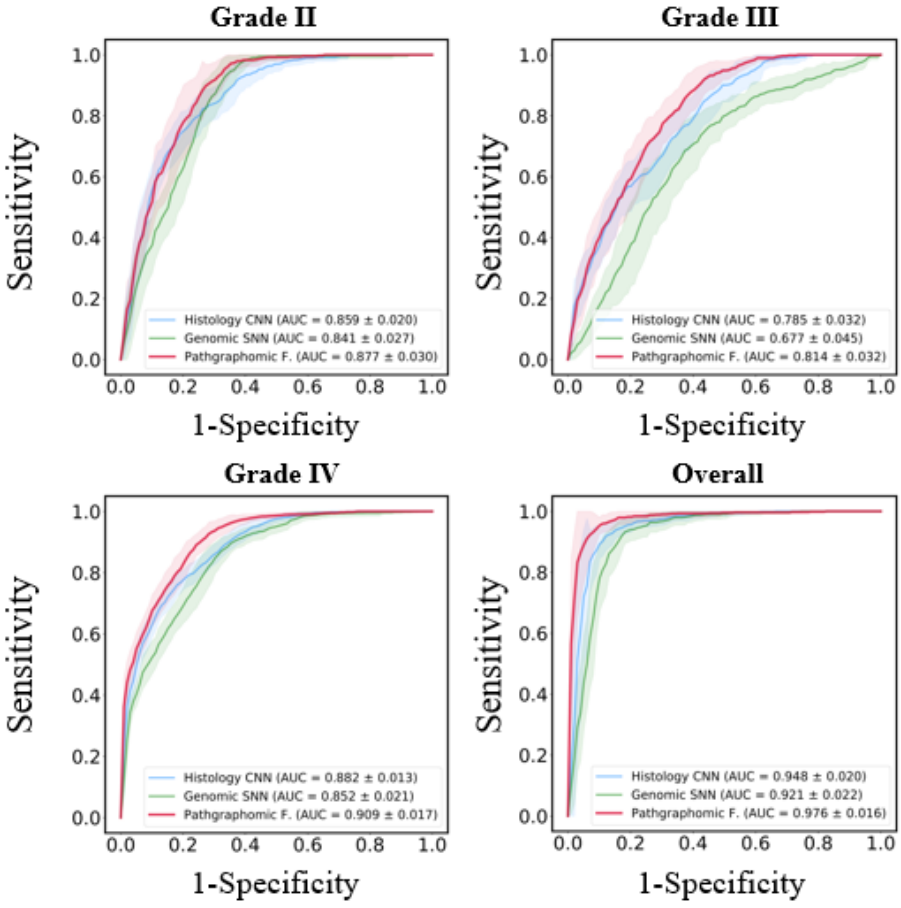

Figure 9: ROC curves

11 Result of Survival Outcome Prediction

11.1 TCGA-GBMLGG

By integrating three types of data—histology ROI images, cell graphs, and genomic profile—using a trimodal fusion network, it is observed that trimodal fusion network outperforms both Cox models and unimodal networks (Table 2).

Among all unimodal networks, the genomic SNN achieves the highest performance with a *C*-Index of 0.8024. The bimodal fusion network (CNN $\otimes$ SNN), improves the c-Index by 0.0181, resulting in a c-Index of 0.8205. The trimodal fusion model network (CNN $\otimes$ GCN $\otimes$ SNN) further increased the c-Index to 0.8254. The pre-aggregation version (($\cup$ CNN) $\otimes$ ($\cup$ GCN) $\otimes$ SNN) achieves the highest c-Index of 0.8361.

However, no significant improvement when adding the graph cell data into the (CNN $\otimes$ SNN) network is seen, potentially because the CNN network has captured the detailed information embedded in the histology images. The information comes from the cell graph data, as an explicit feature extraction from the histology images, is likely to be repetitive with the histology images.

The ensembling methods do not significantly improve the performance which demonstrates that the improvement in trimodal fusion performance is not caused by the increasing complexity of models but integration of information.

Thus, it is demonstrated that the trimodal fusion network significantly enhances the performance of the models, and the pre-aggregated improvement using self-attention mechanisms further boosts the performance to a higher level.

Table 2: Concordance Index of Survival Outcome Prediction in TCGA-GBMLGG

Model	***C*-Index**
Cox (Age+Gender)	0.7316 ± 0.012
Cox (Subtype)	0.7595 ± 0.011
Cox (Grade)	0.7379 ± 0.013
Cox (Grade+Subtype)	0.7769 ± 0.013
Genomic SNN	0.8024 ± 0.017
Genomic (SNN $\otimes$ SNN)	0.7848 ± 0.012
Histology GCN	0.7478 ± 0.019
Histology (GCN $\otimes$ GCN)	0.7374 ± 0.026
Histology CNN	0.7955 ± 0.016
Histology (CNN $\otimes$ CNN)	0.7930 ± 0.014
Pathomic F. (CNN $\otimes$ SNN)	0.8205 ± 0.011
Pathomic F. (GCN $\otimes$ SNN)	0.8126 ± 0.014
Pathomic F. (CNN $\otimes$ GCN $\otimes$ SNN)	0.8254 ± 0.013
Pathomic F. (($\cup$ CNN) $\otimes$ SNN)	0.8318 ± 0.011
Pathomic F. (($\cup$ GCN) $\otimes$ SNN)	0.8239 ± 0.016
Pathomic F. (($\cup$ CNN) $\otimes$ ($\cup$ GCN) $\otimes$ SNN)	$\mathbf{0.8361 \pm 0.015}$

11.2 TCGA-KIRC

Similarly, the trimodal fusion yields an improvement of *C*-Index from 0.6754 (Cox Grade), 0.6828 (SNN), 0.6403 (GCN), 0.6628 (CNN) to 0.7144 (Table 3). However, in TCGA-KIRC dataset the trimodal fusion does not improve the performance compared to bimodal fusion (CNN $\otimes$ SNN) but the gap between these two models is negligible. This implies that the pathology-based CNN network is complex enough and has provided enough information and refined histology, graph-based GCN network does not provide more information in this TCGA-KIRC dataset.

Table 3: Concordance Index of Survival Outcome Prediction in TCGA-KIRC

Model	***C*-Index**
Cox (Age+Gender)	0.6300 ± 0.0240
Cox (Grade)	0.6754 ± 0.0360
Genomic SNN	0.6828 ± 0.0260
Genomic (SNN $\otimes$ SNN)	0.6820 ± 0.0270
Histology GCN	0.6403 ± 0.0300
Histology (GCN $\otimes$ GCN)	0.6382 ± 0.0320
Histology CNN	0.6628 ± 0.0210
Histology (CNN $\otimes$ CNN)	0.6706 ± 0.0230
Pathomic F. (CNN $\otimes$ SNN)	$\mathbf{0.7187 \pm 0.0270}$
Pathomic F. (GCN $\otimes$ SNN)	0.6884 ± 0.0250
Pathomic F. (CNN $\otimes$ GCN $\otimes$ SNN)	0.7144 ± 0.0280

Table 4: Log-rank Test between Different Risk Groups in TCGA-GBMLGG

Model	**[0,50] vs.(50,100]**	**[0,33] vs.(33,66]**	**[33,66] vs.(66,100]**
Cox (Age+Gender)	1.9009×10^{-92}	1.4754×10^{-38}	5.9276×10^{-27}
Cox (Subtype)	2.0646×10^{-228}	3.8569×10^{-26}	1.9446×10^{-51}
Cox (Grade)	5.9983×10^{-224}	9.5682×10^{-23}	2.9416×10^{-66}
Cox (Grade+Subtype)	5.2919×10^{-215}	1.1409×10^{-40}	5.0163×10^{-52}
Genomic SNN	7.8807×10^{-53}	7.6936×10^{-1}	1.3723×10^{-83}
Genomic (SNN $\otimes$ SNN)	2.9816×10^{-51}	NA!	9.6741×10^{-140}
Histology GCN	1.0782×10^{-27}	4.4546×10^{-2}	2.0796×10^{-25}
Histology (GCN $\otimes$ GCN)	2.5976×10^{-24}	1.0016×10^{-1}	5.1262×10^{-24}
Histology CNN	1.2894×10^{-42}	2.0428×10^{-5}	2.8042×10^{-25}
Histology (CNN $\otimes$ CNN)	4.9528×10^{-44}	1.0932×10^{-6}	1.7967×10^{-26}
Pathomic F. (CNN $\otimes$ SNN)	3.0635×10^{-54}	7.5417×10^{-2}	2.6446×10^{-77}
Pathomic F. (GCN $\otimes$ SNN)	2.7414×10^{-55}	8.1649×10^{-2}	3.5469×10^{-72}
Pathomic F. (CNN $\otimes$ GCN $\otimes$ SNN)	2.6135×10^{-56}	1.4395×10^{-2}	2.5224×10^{-72}
Pathomic F. (($\cup$ CNN) $\otimes$ SNN)	2.0570×10^{-60}	1.1250×10^{-4}	6.3161×10^{-70}
Pathomic F. (($\cup$ GCN) $\otimes$ SNN)	3.1880×10^{-55}	5.7846×10^{-2}	5.3907×10^{-78}
Pathomic F. (($\cup$ CNN) $\otimes$ ($\cup$ GCN) $\otimes$ SNN)	2.8401×10^{-63}	5.1256×10^{-2}	3.0150×10^{-80}

Table 5: Log-rank Test between Different Risk Groups in TCGA-KIRC

Model	**[0,50] vs.(50,100]**	**[0,25] vs.(25,50]**	**[25,50] vs.(50,75]**	**[50,75] vs.(75,100]**
Cox (Age+Gender)	1.2681×10^{-16}	1.0818×10^{-1}	1.2000×10^{-5}	3.5955×10^{-1}
Cox (Grade)	4.4188×10^{-17}	1.2497×10^{-7}	4.5200×10^{-4}	5.1309×10^{-1}
Genomic SNN	4.9788×10^{-19}	3.2797×10^{-1}	6.0500×10^{-3}	2.5958×10^{-16}
Genomic (SNN $\otimes$ SNN)	1.0636×10^{-21}	2.2891×10^{-1}	1.2990×10^{-3}	4.3309×10^{-16}
Histology GCN	8.5770×10^{-7}	2.3327×10^{-1}	1.9146×10^{-1}	1.5224×10^{-3}
Histology (GCN $\otimes$ GCN)	1.3076×10^{-4}	8.8447×10^{-1}	1.2797×10^{-1}	1.3982×10^{-2}
Histology CNN	5.2572×10^{-16}	3.6142×10^{-1}	2.0000×10^{-6}	2.2440×10^{-3}
Histology (CNN $\otimes$ CNN)	1.6608×10^{-14}	1.0652×10^{-1}	1.9210×10^{-3}	1.0836×10^{-3}
Pathomic F. (CNN $\otimes$ SNN)	2.5786×10^{-25}	3.3537×10^{-1}	7.6100×10^{-4}	4.4696×10^{-14}
Pathomic F. (GCN $\otimes$ SNN)	5.5496×10^{-20}	2.8652×10^{-1}	1.0740×10^{-3}	8.0657×10^{-15}
Pathomic F. (CNN $\otimes$ GCN $\otimes$ SNN)	1.3312×10^{-27}	5.8778×10^{-1}	8.0000×10^{-6}	7.1031×10^{-13}

12 Patient Stratification

Although the model has achieved a better performance in terms of c-Index, it is interesting and meaningful to know whether the model can stratify the patients, which means that whether the patients of different hazards have significantly different survival time.

The patients in TCGA-GBMLGG are equally split into two groups, the lower half of groups containing patients have relatively low predicted hazards $[0\%, 50\%]$, while the upper half $(50\%, 100\%]$ containing patients with high hazards. Then a log-rank test [33] is carried on the survival time of these two groups to analyze the statistical significance between two groups. Another split is created by $[0\%, 33\%]$, $(33\%, 66\%]$, $(66\%, 100\%]$.

The log-rank test is conducted between the low-hazard and intermediate-hazard groups, and between the intermediate-hazard and high-hazard groups.

A smaller p-value of log-rank test is preferred, which implies that the survival time of two groups is unlikely to be the same.

The results for TCGA-GBMGLL are displayed in Table 4. It reveals that the low-to-intermediate groups are difficult to distinguish compared to the intermediate-to-high groups. When comparing the $[0\%, 50\%]$ group with the $(50\%, 100\%]$ group, the p-value of the SNN model is 2.06×10^{-228}, which is the lowest observed in unimodel. This p-value decreases to 2.61×10^{-56} with the trimodal fusion (CNN $\otimes$ GCN $\otimes$ SNN). The pre-aggregation trimodal fusion (($\cup$ CNN) $\otimes$ ($\cup$ GCN) $\otimes$ SNN) improves the performance by decreasing the p-value to 2.84×10^{-63}. Similarly, all fusion models have poor performance in distinguishing the low-to-intermediate hazard patients ($[0\%, 33\%]$ group with the $(33\%, 66\%]$ group). However, when it comes to stratifying the intermediate-to-high hazard patients ($[33\%, 66\%]$ group with the $(66\%, 100\%]$ group), the model performance increases.

Patients in the TCGA-GBMLGG dataset are categorized into two groups based on their survival time: the red group comprises patients with a survival time of less than 5 years, while the blue group includes patients with a survival time of more than 5 years. Their hazard predictions are visualized in Histogram Figure 14. It can be observed that CNN and GCN have large overlapping areas, indicating that they cannot stratify the patients well. When it comes to the trimodal fusion, the patients are clearly clustered into two groups where the left groups have blue color (implying for patients living a longer life), and right groups have red color (implying for patients living a shorter life). The left group has relatively small hazard, while the right group has relatively high hazard and there is a clear boundary between two groups. Figure 15 records the Kaplan-Meier Plots of all five models. Two trimodal fusions do improve the patients stratification compared to the histology CNN and GCN. Also, there is a significant stratification between the intermediate–to-high hazard patients in survival time.

Patients in TCGA-KIRC are equally split into two groups according to the hazards $[0\%, 50\%]$, $(50\%, 100\%]$. Another split is created by $[0\%, 25\%]$, $(25\%, 50\%]$, $(50\%, 75\%]$, $(75\%, 100\%]$.

The stratification effect of different models varies when comparing different groups in TCGA-KIRC. As shown in Table 5, when comparing the groups in the lower half hazards and the higher half hazards, the trimodal fusion achieves the most significant p-value of 1.33×10^{-27}. Overall, the trimodal fusion CNN $\otimes$ GCN $\otimes$ SNN, although not the best in all hazard splits, achieves relatively small p-values in all splits, indicating strong stratification effects across different groups.

In Figure 14, the real longer surviving and shorter surviving patients groups given by GCN,CNN,SNN, and trimodal fusion are almost overlapping, indicating poor stratification results, while the trimodal fusion separates the histogram into two high peaks. In Figure 16, the KM curves of trimodal fusion network have a significant difference.

In conclusion, the trimodal fusion method significantly improves patient stratification in both TCGA-GBMLGG and TCGA-KIRC, as evidenced by the p-values of the log-rank test, histogram, and swarm plots.

13 Alignment on the WHO Grades Paradigm

The hazards of the patients should be positively correlated to the WHO grades in the real sense. As described before, patients with higher WHO grades have more aggressive gliomas and therefore tend to survive shorter.

In TCGA-GBMLGG, the swarm plots in Figure 14 show that for patients with ***IDHwt ATC*** gliomas, the pre-aggregate trimodal fusion method provides the best alignment with WHO grades.

The hazard prediction from histology CNN and GCN also have better alignment with the WHO grade IV for patients with ***IDHmut ATC*** gliomas. The pink points (indicating Grade IV patients) are located at the peaks and the highest hazard regions.

When it comes to the patient with ***ODG*** gliomas, all models perform poorly in alignment with the WHO grades.

In TCGA-GBMLGG, in the overall setting of Kaplan-Meier plots (Figure 15), the most hazardous groups (red) predicted by the histology CNN and GCN models significantly deviate from the Grade IV groups. Additionally,

it can be observed that as time approaches 15 years, the curve of the intermediate-hazard group converges to the Grade III group, and the curve of the low-hazard group converges to the Grade II group.

In TCGA-KIRC, both the CNN and GCN models show poor alignment with the Grade 3 (purple and red) group. The SNN model aligns well with the Grade 3 (red) group but not with the Grade 3 (purple) group. In contrast, the trimodal fusion network performs well with both the Grade 3 (purple) and Grade 3 (red) groups (Figure 16).

14 Interpretation of Survival Outcome Prediction

14.1 Integrated Gradients (IG) to Interpret the Genomic Data in Survival Outcome Prediction

The Integrated Gradients (IG) (13) interprets the contribution of genomic profiles to predictions.

Integrated Gradients (IG) formula:

$$\mathrm{IG}_i(x) = (x_i - x_i') \int_{\alpha=0}^{\alpha=1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} \, d\alpha \tag{13}$$

Explanation:

- $\mathrm{IG}_i(x)$: Integrated Gradient for the i-th input feature x_i of the input x.
- $(x_i - x_i')$: Difference between the i-th feature of the input x and the baseline x'.
- F: Model (SNN, GCN, CNN, bimodal fusion, trimodal fusion)
- $F(\cdot)$: Model output, which is hazard in survival outcome prediction

The IGs of all patients are computed. Then they are aggregated by taking the average among IGs of all patients.

Mathematically, if the partial derivative of the model F with respect to x_i at point $x' + \alpha(x - x')$ is positive, then a slight move $\varepsilon > 0$ from this point towards the next point $x' + (\alpha + \varepsilon)(x - x')$ leads to increasing model output (such as a hazard) F, and vice versa. The integral $\int_{\alpha=0}^{\alpha=1}$ takes an average of all increments or decrements caused by these small movements along the path from the baseline x' to the current point x. This is exactly the positive or negative contribution of feature x_i.

In TCGA-GBMLGG (Figure 10) , IDH mutation and PTEN are identified as the most influential among all 320 genes. Specifically, the IDH mutation exhibits higher Integrated Gradient (IG), correlating with increased mortality in IDH wild-type astrocytoma patients. Conversely, it correlates with decreased mortality in IDH-mutated astrocytoma and oligodendroglioma patients. PTEN has a similar trend as the IDH mutation gene in IDHwt-ATC and ODG patients, but different trends in IDHmut-ATC patients.

In TCGA-KIRC (Figure 11), the TIPARP, CYP3A7, DDX43 are most influential among all 362 genes. It can be observed that the CYP3A7, DDX43 has a negative effect on survival. Conversely, the APCDOIL seems to be negatively correlated with mortality.

Overall, the Integrated Gradient (IG) contributions analyzed in these two models show strong agreement with each other: the rank of each gene is similar. Additionally, they exhibit a high level of agreement with the findings reported in the original paper [1], with any slight differences possibly arising from variations in IG implementation.

14.2 Gradient-weighted Class Activation Mapping (Grad-CAM) to Interpret CNN

Gradient-weighted Class Activation Mapping (Grad-CAM) (14) is employed to interpret histology CNNs. Grad-CAM emphasizes the significance of features concerning a single sample (local importance), contrasting with Integrated Gradients (IG), which highlights feature importance among all samples (global importance).

$$\text{Grad-CAM}(x) = \mathrm{ReLU}\left(\sum_k W_k(x) \odot A_k(x)\right) \tag{14}$$

$$W_k(x) = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\partial L}{\partial A_{k,i,j}}(x) \tag{15}$$

- x: Input image
- Grad-CAM: The importance for making prediction with respect to each pixel in x

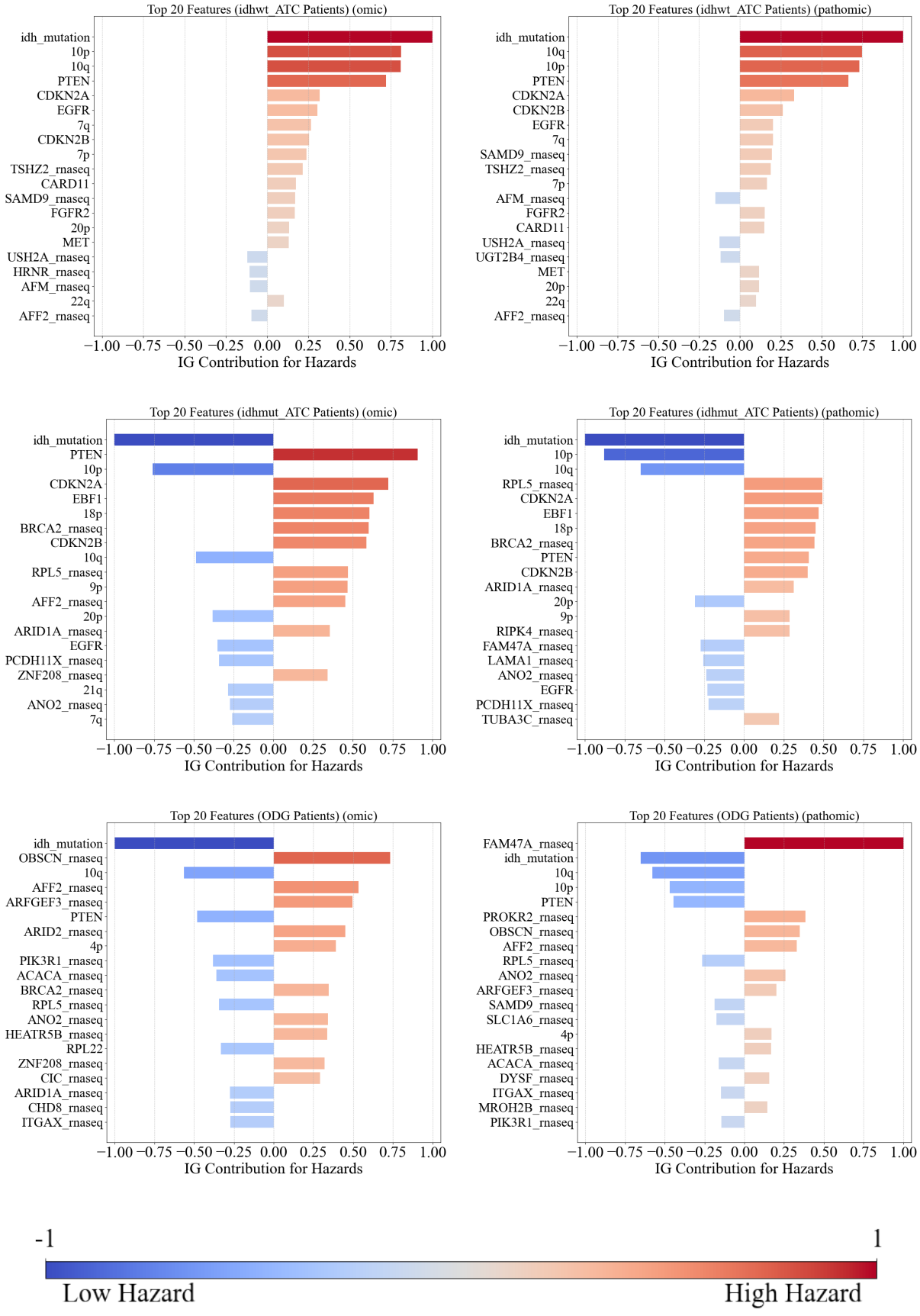

Figure 10: Comparison of Uni-modal (Omic, left) and the Multi-modal (Pathomic, right) IG Interpretatbility in TCGA-GBMLGG: The IDH mutation, PTEN are most influential genes among all 320 genes

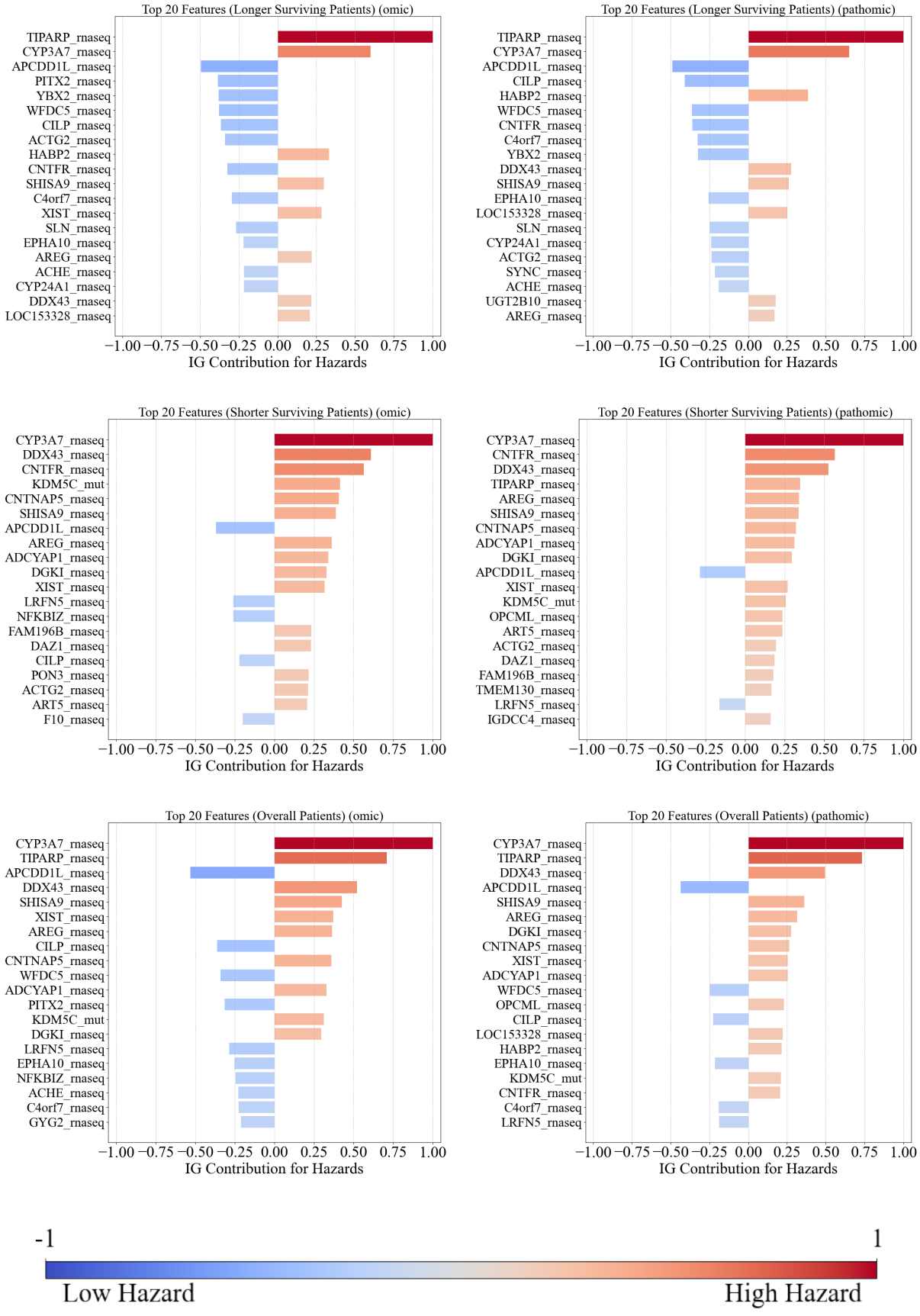

Figure 11: Comparison of Uni-modal (Omic, left) and the Multi-modal (Pathomic, right) IG Interpretatbility in TCGA-KIRC: The CYP3A7, DDX43 are most influential genes among all 362 genes

- W_k: Weights, representing the gradient-weighted average of the k-th channel of target layer .
- A: Activation map of the target layer, which is of shape (channel k, row i, column j).
- A_k: Activation map of the k-th channel for target layer.
- $A_{k,i,j}$: Activation value at position (i, j) in target layer k.
- L: Loss Function, the CoxLoss (11)

Combining the magnitudes of the gradients W_k and the magnitude of the latent space A_k yields the Grad-CAM.

The results for TCGA-GBMLGG and TCGA-KIRC are displayed in Figure 12 and 13. For both gliomas and renal cells, Grad-CAM focuses primarily on cell nuclei stained purple, followed by the cytoplasm and fibers. In TCGA-KIRC, Grad-CAM can precisely localize each cell nucleoli, whereas in TCGA-GBMLGG, Grad-CAM creates a patch of brightness around the nucleoli, roughly indicates their positions, and creates extra brightness if there are high-density regions of nucleoli. This demonstrates the importance of nucleoli in prediction.

15 Conclusion

In this project, the original paper [1] is reproduced and our results are similar to the original one.

In WHO Grade Classification of TCGA-GBMLGG task, the trimodal fusion network (CNN $\otimes$ GCN $\otimes$ SNN) ranks the highest in AUC, AP, F1-score, and F1 Grade II, Grade III, and Grade IV scores.

In Survival Outcome Prediction of TCGA-GBMLGG task, the trimodal fusion network (CNN $\otimes$ GCN $\otimes$ SNN) by original paper ranks the third highest with 0.8254 C-Index. The pre-aggregate design improves the C-Index to the second highest 0.8318 ((∪ CNN) $\otimes$ SNN) and highest 0.8361 ((∪ CNN) $\otimes$ (∪ GCN) $\otimes$ SNN). The stratification effect of (CNN $\otimes$ GCN $\otimes$ SNN) and ((∪ CNN) $\otimes$ (∪ GCN) $\otimes$ SNN) on low-to-high hazard patients are better than the other deep learning models, and on mid-to-high hazard patients are better than CNN model. According to IG, the IDH mutation and PTEN are the most influential genes. The Grad-CAM tells that the CNN network pays the highest attention to the nucleoli and then the surrounding fried-eggs-like patches.

In Survival Outcome Prediction of TCGA-KIRC task, the bimodal fusion network (CNN $\otimes$ SNN) ranks the highest with 0.7187 C-Index, followed by trimodal fusion network (CNN $\otimes$ GCN $\otimes$ SNN) with 0.7144 C-Index. The trimodal fusion (CNN $\otimes$ GCN $\otimes$ SNN) has the best low-to-high stratification performance among all models. According to IG, the CYP3A7 and DDX43 are the most influential genes. The Grad-CAM tells that the CNN network pays the highest attention to the nucleoli and the edges of cells.

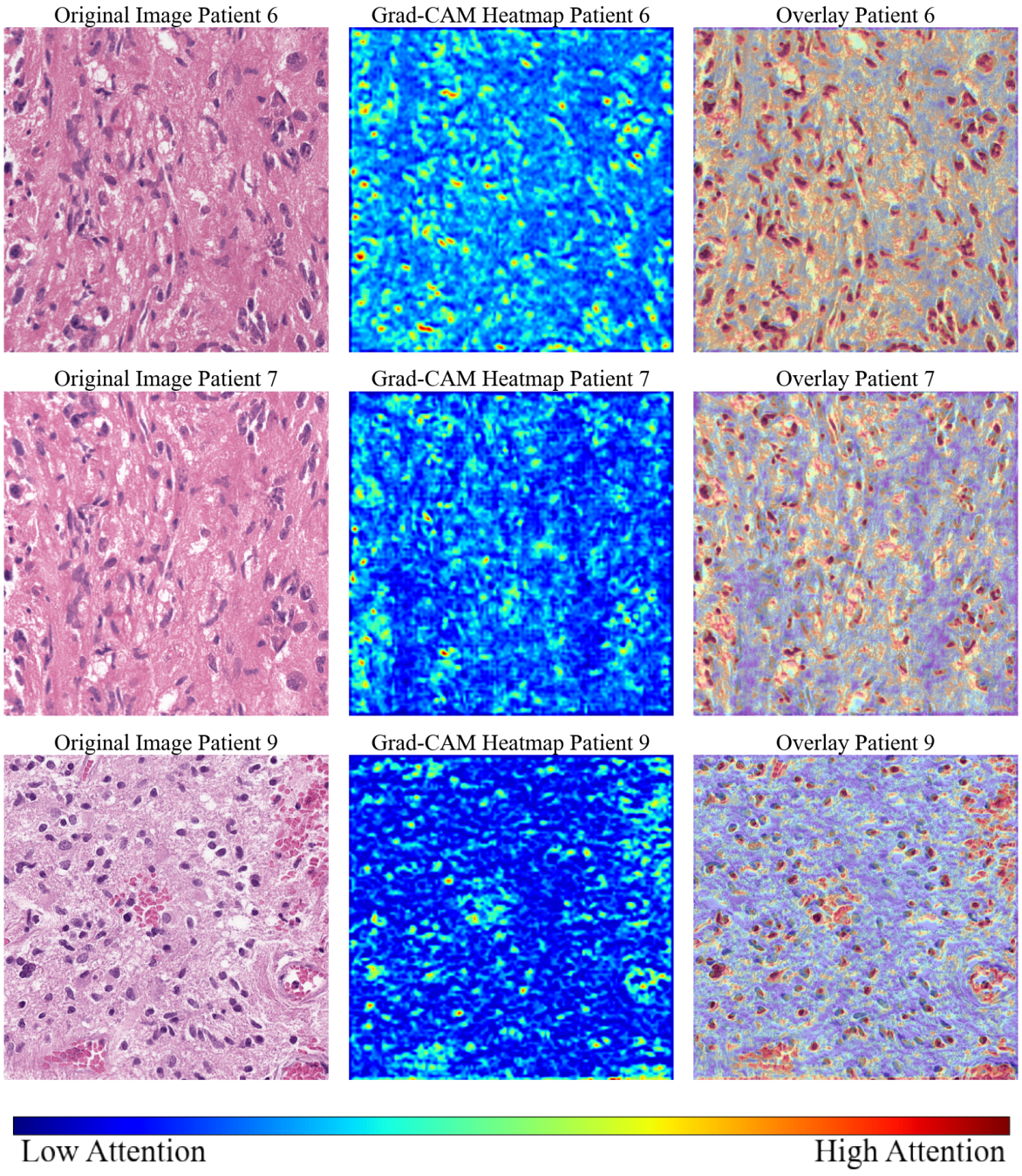

Figure 12: Grad-CAM of 4 Patients in TCGA-GBMLGG

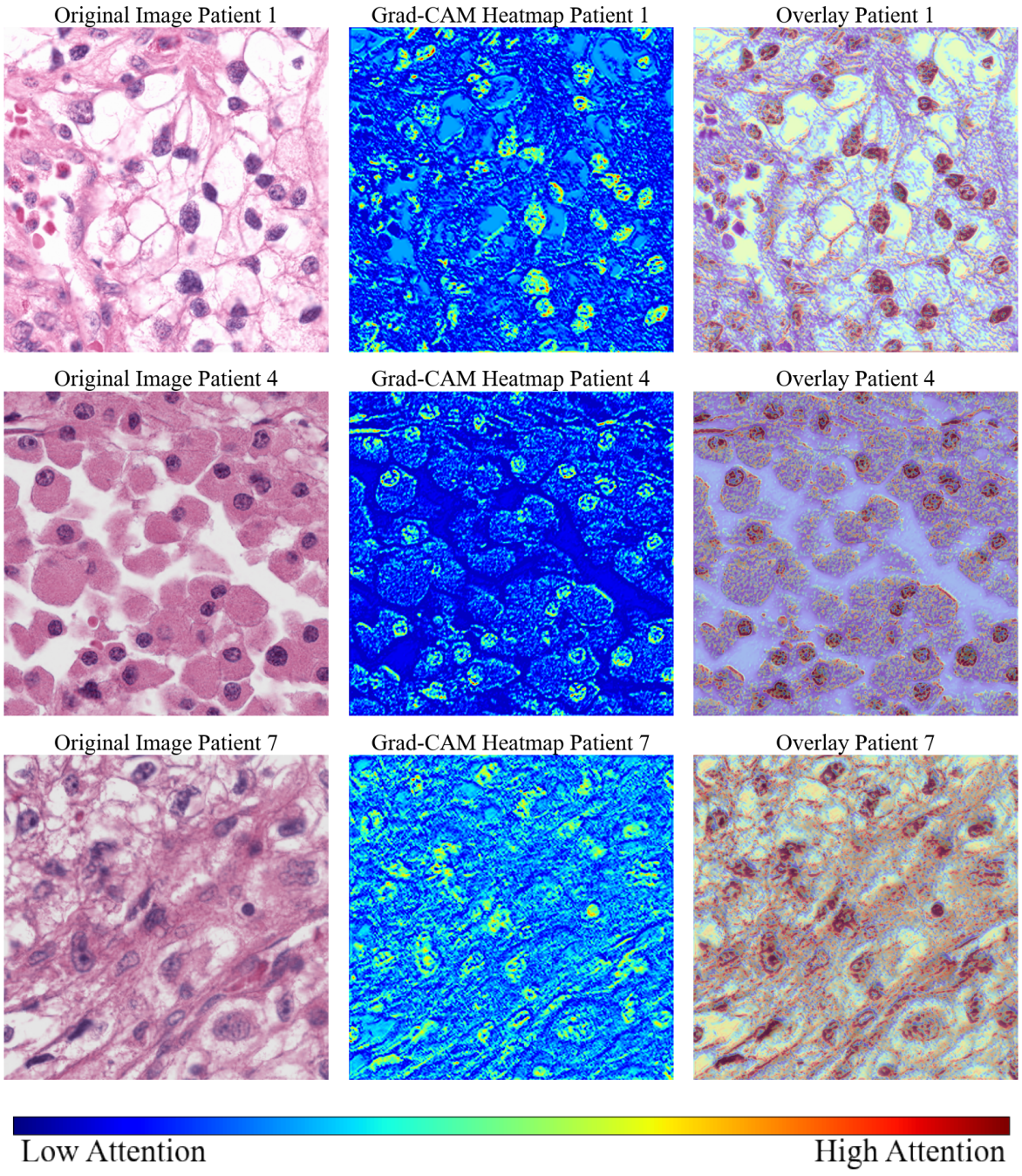

Figure 13: Grad-CAM of 4 Patients in TCGA-KIRC

References

[1] Richard J Chen, Ming Y Lu, Jingwen Wang, Drew FK Williamson, Scott J Rodig, Neal I Lindeman, and Faisal Mahmood. 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.

[2] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In International conference on machine learning, pages 3319–3328. PMLR, 2017.

[3] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision, pages 618–626, 2017.

[4] Hung-Chia Chen, Ralph L Kodell, Kuang Fu Cheng, and James J Chen. Assessment of performance of survival prediction models for cancer prognosis. BMC medical research methodology, 12:1–11, 2012.

[5] Shigehisa Kubota, Tetsuya Yoshida, Susumu Kageyama, Takahiro Isono, Takeshi Yuasa, Junji Yonese, Ryoji Kushima, Akihiro Kawauchi, and Tokuhiro Chano. A risk stratification model based on four novel biomarkers predicts prognosis for patients with renal cell carcinoma. World Journal of Surgical Oncology, 18:1–11, 2020.

[6] David N Louis. Who classification of tumours of the central nervous system. (No Title), 2016.

[7] Shuguang Zuo, Xinhong Zhang, and Liping Wang. A rna sequencing-based six-gene signature for survival prediction in patients with glioblastoma. Scientific reports, 9(1):2615, 2019.

[8] David R Cox. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2):187–202, 1972.

[9] Theodore Sakellaropoulos, Konstantinos Vougas, Sonali Narang, Filippos Koinis, Athanassios Kotsinas, Alexander Polyzos, Tyler J Moss, Sarina Piha-Paul, Hua Zhou, Eleni Kardala, et al. A deep learning framework for predicting response to therapy in cancer. Cell reports, 29(11):3367–3373, 2019.

[10] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. BMC medical research methodology, 18:1–12, 2018.

[11] Ole-Johan Skrede, Sepp De Raedt, Andreas Kleppe, Tarjei S Hveem, Knut Liestøl, John Maddison, Hanne A Askautrud, Manohar Pradhan, John Arne Nesheim, Fritz Albregtsen, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. The Lancet, 395(10221):350–360, 2020.

[12] Andrew Janowczyk and Anant Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. Journal of pathology informatics, 7(1):29, 2016.

[13] Ellery Wulczyn, David F Steiner, Zhaoyang Xu, Apaar Sadhwani, Hongwu Wang, Isabelle Flament-Auvigne, Craig H Mermel, Po-Hsuan Cameron Chen, Yun Liu, and Martin C Stumpe. Deep learning-based survival prediction for multiple cancer types using histopathology images. PloS one, 15(6):e0233678, 2020.

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

[15] Jason W Wei, Laura J Tafe, Yevgeniy A Linnik, Louis J Vaickus, Naofumi Tomita, and Saeed Hassanpour. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. Scientific reports, 9(1):3358, 2019.

[16] Ruoyu Li, Jiawen Yao, Xinliang Zhu, Yeqing Li, and Junzhou Huang. Graph cnn for survival analysis on whole slide pathological images. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 174–182. Springer, 2018.

[17] Mohammed Adnan, Shivam Kalra, and Hamid R Tizhoosh. Representation learning of histopathology images using graph neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 988–989, 2020.

[18] Yigit Ozen, Selim Aksoy, Kemal Kösemehmetoğlu, Sevgen Önder, and Ayşegül Üner. Self-supervised learning with graph neural networks for region of interest retrieval in histopathology. In 2020 25th International conference on pattern recognition (ICPR), pages 6329–6334. IEEE, 2021.

[19] Yiqing Shen, Bingxin Zhou, Xinye Xiong, Ruitian Gao, and Yu Guang Wang. How graph neural networks enhance convolutional neural networks towards mining the topological structures from histology. In ICML Workshop on Computational Biology, volume 8, 2022.

[20] Pushpak Pati, Guillaume Jaume, Antonio Foncubierta-Rodriguez, Florinda Feroce, Anna Maria Anniciello, Giosue Scognamiglio, Nadia Brancati, Maryse Fiche, Estelle Dubruc, Daniel Riccio, et al. Hierarchical graph representations in digital pathology. Medical image analysis, 75:102264, 2022.

[21] Iain Carmichael, Benjamin C Calhoun, Katherine A Hoadley, Melissa A Troester, Joseph Geradts, Heather D Couture, Linnea Olsson, Charles M Perou, Marc Niethammer, Jan Hannig, et al. Joint and individual analysis of breast cancer histologic images and genomic covariates. The annals of applied statistics, 15(4):1697, 2021.

[22] Nikhilanand Arya and Sriparna Saha. Multi-modal advanced deep learning architectures for breast cancer survival prediction. Knowledge-Based Systems, 221:106965, 2021.

[23] Yicheng Wang, Ye Luo, Bo Li, and Xiaoang Shen. Multi-modality fusion based lung cancer survival analysis with self-supervised whole slide image representation learning. In Chinese Conference on Pattern Recognition and Computer Vision (PRCV), pages 333–345. Springer, 2023.

[24] Wei Shao, Tongxin Wang, Liang Sun, Tianhan Dong, Zhi Han, Zhi Huang, Jie Zhang, Daoqiang Zhang, and Kun Huang. Multi-task multi-modal learning for joint diagnosis and prognosis of human cancers. Medical image analysis, 65:101795, 2020.

[25] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. Advances in neural information processing systems, 27, 2014.

[26] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784, 2014.

[27] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1):21–27, 1967.

[28] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018.

[29] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In International conference on machine learning, pages 4182–4192. PMLR, 2020.

[30] Wei-Chien Wang, Euijoon Ahn, Dagan Feng, and Jinman Kim. A review of predictive and contrastive self-supervised learning for medical images. Machine Intelligence Research, 20(4):483–513, 2023.

[31] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In International conference on machine learning, pages 2127–2136. PMLR, 2018.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

[33] J Martin Bland and Douglas G Altman. The logrank test. Bmj, 328(7447):1073, 2004.

A TCGA-GBMLGG

	TCGA-ID (TOTAL:769)	Genomics (Size: 320) (TOTAL:769)	ROI Images (Large Patches) (Size: $3 \times 1024 \times 1024$) (TOTAL:1505)	ROI Images (Small Patches) (Size: $3 \times 512 \times 512$) (TOTAL: 13545)	ROI Graphs (TOTAL:1505)
(1)	02-0054	DNA&RNA 02-0054	Image 02-0054-1	Image 02-0054-1-0-0	Graph 02-0054-1
(2)				Image 02-0054-1-0-256	
(3)				Image 02-0054-1-0-512	
(4)				Image 02-0054-1-256-0	
(5)				Image 02-0054-1-256-256	
(6)				Image 02-0054-1-256-512	
(7)				Image 02-0054-1-512-0	
(8)				Image 02-0054-1-512-256	
(9)				Image 02-0054-1-512-512	
(10)			Image 02-0054-2	Image 02-0054-2-0-0	Graph 02-0054-2
(11)				Image 02-0054-2-0-256	
(12)				Image 02-0054-2-0-512	
(13)				Image 02-0054-2-256-0	
(14)				Image 02-0054-2-256-256	
(15)				Image 02-0054-2-256-512	
(16)				Image 02-0054-2-512-0	
(17)				Image 02-0054-2-512-256	
(18)				Image 02-0054-2-512-512	
...	...	...	...	...	...
(13537)	02-0115	DNA&RNA 02-0115	Image 02-0115-1	Image 02-0115-1-0-0	Graph 02-0115-1
(13538)				Image 02-0115-1-0-256	
(13539)				Image 02-0115-1-0-512	
(13540)				Image 02-0115-1-256-0	
(13541)				Image 02-0115-1-256-256	
(13542)				Image 02-0115-1-256-512	
(13543)				Image 02-0115-1-512-0	
(13544)				Image 02-0115-1-512-256	
(13545)				Image 02-0115-1-512-512	

Table 6: TCGA-GBMLGG dataset: Each patient (TCGA-ID) has a unique genomics sequence, but can have a varying number of ROI image (Large Patches). Each ROI image can be split into 9 ROI images (Small Patches). Each ROI Graph is created from an ROI image (Large Patch).

B TCGA-KIRC

	TCGA-ID (TOTAL:417)	Genomics (SIZE:362) (TOTAL:417)	ROI Images (SIZE: $3 \times 512 \times 512$) (TOTAL:1260)	ROI Graphs (TOTAL:1260)
(1)	B0-4688	DNA&RNA B0-4688	Image B0-4688-1-ROI-1	Graph B0-4688-1-ROI-1
(2)			Image B0-4688-1-ROI-2	Graph B0-4688-1-ROI-2
(3)			Image B0-4688-1-ROI-3	Graph B0-4688-1-ROI-3
(4)	B0-4690	DNA&RNA B0-4690	Image B0-4690-1-ROI-1	Graph B0-4690-1-ROI-1
(5)			Image B0-4690-1-ROI-2	Graph B0-4690-1-ROI-2
(6)			Image B0-4690-1-ROI-3	Graph B0-4690-1-ROI-3
...	...	...	...	...
(1243)	T7-A92I	DNA&RNA T7-A92I	Image T7-A92I-1-ROI-1	Graph T7-A92I-1-ROI-1
(1244)			Image T7-A92I-1-ROI-2	Graph T7-A92I-1-ROI-2
(1245)			Image T7-A92I-1-ROI-3	Graph T7-A92I-1-ROI-3
(1246)			Image T7-A92I-2-ROI-1	Graph T7-A92I-2-ROI-1
(1247)			Image T7-A92I-2-ROI-2	Graph T7-A92I-2-ROI-2
(1248)			Image T7-A92I-2-ROI-3	Graph T7-A92I-2-ROI-3
(1249)	A3-3365	DNA&RNA A3-3365	Image A3-3365-1-ROI-1	Graph A3-3365-1-ROI-1
(1250)			Image A3-3365-1-ROI-2	Graph A3-3365-1-ROI-2
(1251)			Image A3-3365-1-ROI-3	Graph A3-3365-1-ROI-3
(1252)			Image A3-3365-2-ROI-1	Graph A3-3365-2-ROI-1
(1253)			Image A3-3365-2-ROI-2	Graph A3-3365-2-ROI-2
(1254)			Image A3-3365-2-ROI-3	Graph A3-3365-2-ROI-3
(1255)	MM-A564	DNA&RNA MM-A564	Image MM-A564-1-ROI-1	Graph MM-A564-1-ROI-1
(1256)			Image MM-A564-1-ROI-2	Graph MM-A564-1-ROI-2
(1257)			Image MM-A564-1-ROI-3	Graph MM-A564-1-ROI-3
(1258)			Image MM-A564-2-ROI-1	Graph MM-A564-2-ROI-1
(1259)			Image MM-A564-2-ROI-2	Graph MM-A564-2-ROI-2
(1260)			Image MM-A564-2-ROI-3	Graph MM-A564-2-ROI-3

Table 7: TCGA-KIRC dataset: Each patient (TCGA-ID) has a unique genomics sequence. 413 patients have 3 ROIs each, and another 3 patients (TCGA-T7-A92I, TCGA-A3-3365, TCGA-MM-A564) have 6 ROIs each. Each ROI Graph is created from an ROI image.

C Kaplan-Meier Plots and Histograms

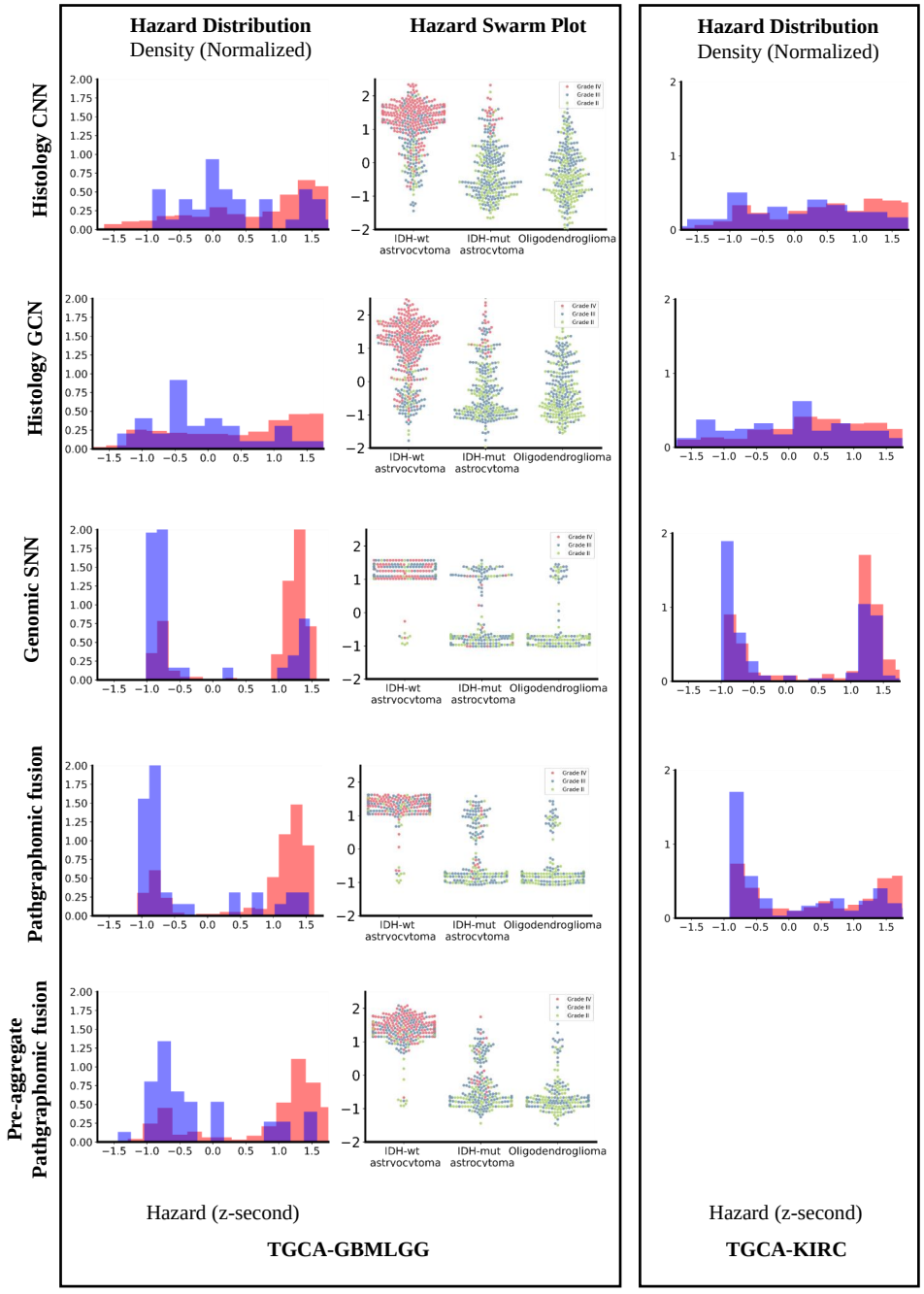

Figure 14: Hazard Histogram and Swarm Plots for TCGA-GBMLGG and TCGA-KIRC

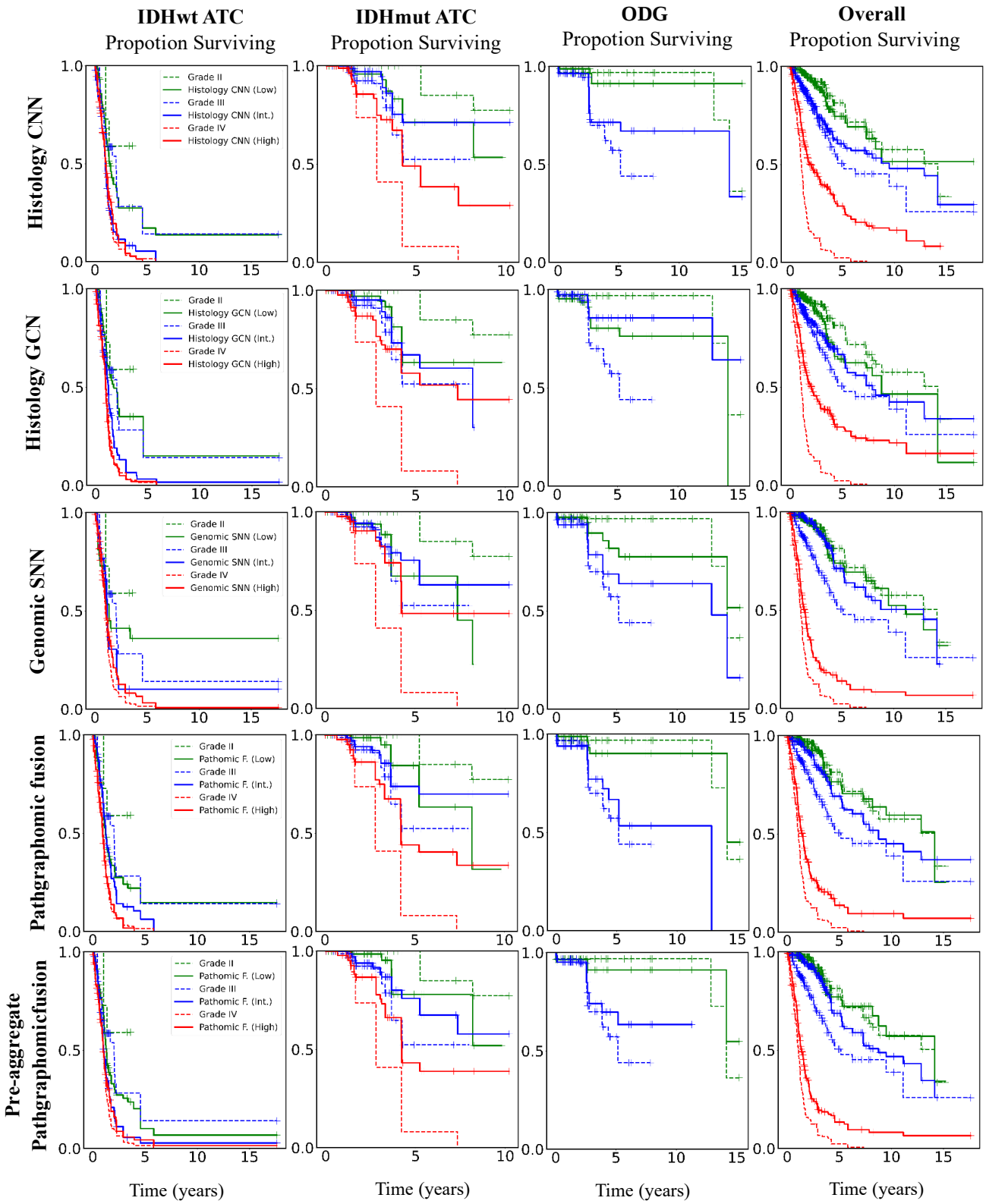

Figure 15: Kaplan-Meier Plot in TCGA-GBMLGG

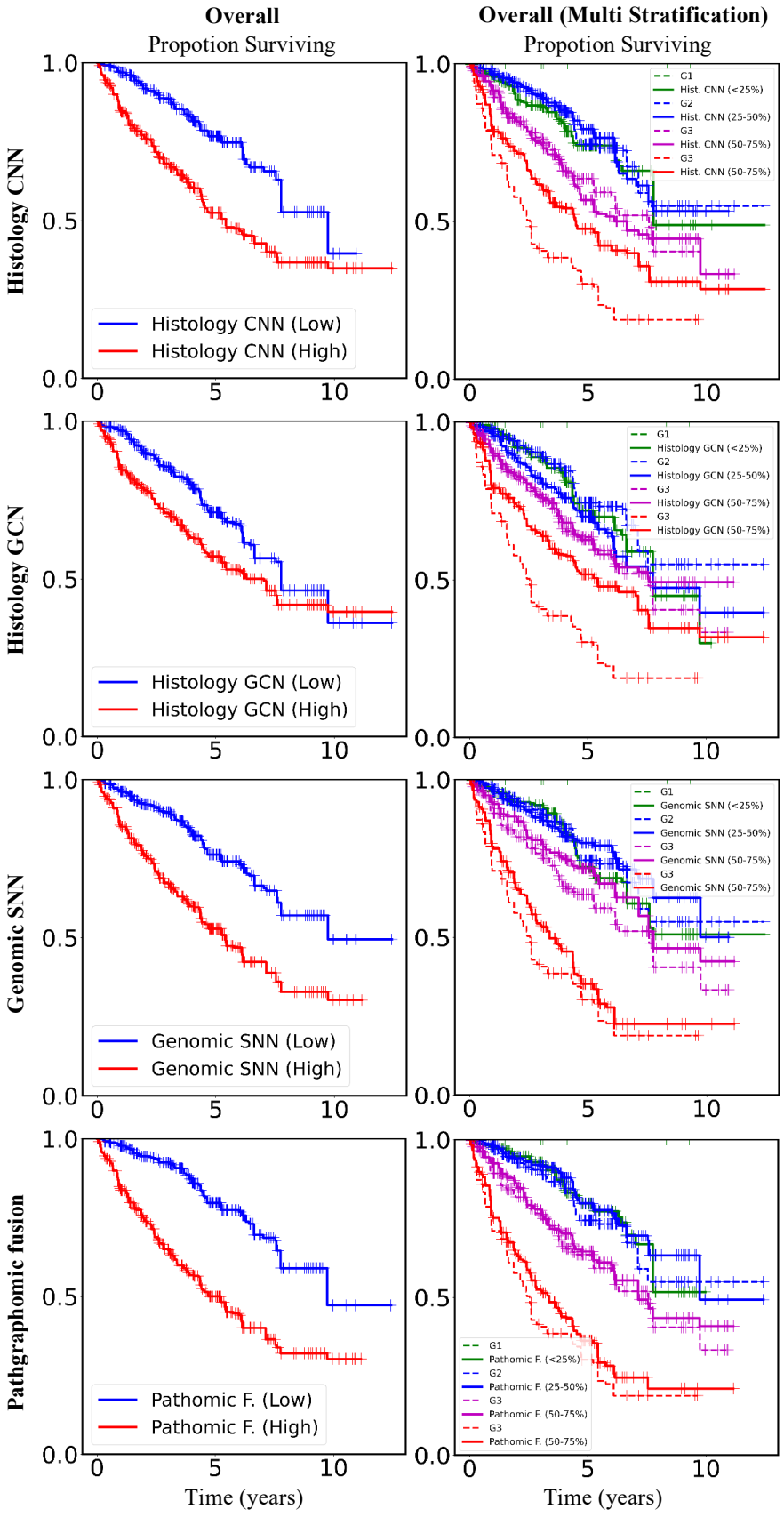

Figure 16: Kaplan-Meier Plot in TCGA-KIRC

D Generative Tools

The following are generated or revised by Chatgpt: 1. GradCAM classes in Gradcam_path_GBMGLL.py 2. GradCAM classes in Gradcam_path_KIRC.py 3. The pre-aggregated Bimodal and Trimodal Fusion in networks.py incorporates the Self Attention structure from ChatGPT and the original network structure. We integrated the concept and architecture from ChatGPT to establish a foundational structure, which we subsequently adapted by adjusting parameters and hyperparameters to suit our specific needs. 4. The Chatgpt is applied to polish the sentences, detect grammatical errors.