Appendix A. Deep Learning-based Survival Analysis

Survival analysis is a task that models the time to an event, where the outcome of the event is not always observed. Such events are called censored, in which the date of the last known encounter is used as a lower bound of the survival time. For the task of cancer survival outcome prediction, an uncensored event would be patient death, and a censored event would include either patient survival or last known follow-up.

Let T be a continuous random variable that represents patient survival time, and the survival function $S(t) = P(T \geq t_0)$ be the probability of a patient surviving longer than time t_0. We can denote the probability that an event occurs instantaneously at a time t (after t_0) as the hazard function $\lambda(t)$. Integrating the hazard function over the time between t and t_0 gives us the survival function [80].

$$\lambda(t) = \lim_{\partial t \to 0} \frac{P(t \leq T \leq t + \partial t | T \geq t)}{\partial t}, S(t) = \exp\left(-\int_0^t \lambda(x)\partial x\right).$$

The most common semi-parametric approach for estimating the hazard function is the Cox proportion hazards model, which assumes that the hazard function can be parameterized as an exponential linear function $\lambda(t|x) = \lambda_0(t)e^{\beta x}$, where $\lambda_0(t)$ is the baseline hazard that describes how the risk of an event changes over time, β are model parameters that describe how the hazard varies with covariates / features X of a patient. In the original model, the baseline hazard $\lambda_0(t)$ is left unspecified, making it difficult to estimate β, however, the Cox partial log-likelihood can be derived that expresses the likelihood of an event to be observed at time t for β, X [81].

$$l(\beta, X) = -\sum_{i \in U}\left(X_i\beta - \log\sum_{j \subset R_i} e^{X_j\beta}\right), \frac{\partial l(\beta, X)}{\partial X_i} = \delta(i)\beta - \sum_{i,j \in C_j, U} \frac{\beta e^{X_i\beta}}{\sum_{k \in C_j} e^{X_k\beta}}$$

where U is the set of uncensored patients, R_i is the set of patients whose time of death or last follow-up is later than i. From the partial log-likelihood, β can be estimated using iterative optimization algorithms such as Newton-Raphson or Stochastic Gradient Descent. To train deep networks for survival analysis, features from the hidden layer are used as covariates in the Cox model, with the derivative of the partial log-likelihood used as error during back-propagation. To evaluate the performance of networks for survival analysis, we use the Concordance Index (c-Index), which measures the concordance of ranking of predicted hazard scores with the ground truth survival times of patients. To demonstrate how well Pathomic Fusion performs over other models, we used the c-Index as a comparative performance metric to measure how well each model is able to predict hazard scores amongst patients (higher is better). Our baseline for clinical practice was using the ground truth molecular subtypes as covariates in a Cox proportional hazard model - the canonical regression technique for modeling survival distributions. P-Values were calculated using the Log Rank Test, which we used to assess low vs. high risk stratification on all datasets, low vs. intermediate and intermediate vs. high (33-66-100 percentile) risk stratification in glioma, and 25-50-75-100 percentile risk stratification in CCRCC [69].

Appendix B. Implementation Details

A. *Inclusion Criteria for Genomic and Transcriptomic Features*

In our analysis on the merged TCGA-GBMLGG and TCGA-KIRC projects, we use 320 and 357 genomic features respectively. Genomic features include mutation (e.g. - binary indication of mutation status for IDH1 gene, 0/1) and copy number variation (CNV) (e.g. - amplified / deleted copies for genes and chromosomal regions). Copy number variation measurement in TCGA uses the Affymetrix SNP 6.0 array to identify repeated copies of genomic regions, with the final output as segment mean values (amplified regions have positive values, deleted regions have negative values). For TCGA-GBMLGG, mutation and CNV data used in our analysis were curated from the same set of genomic features used in Mobadersany *et al.* [29]. Genes curated include EGFR, MDM4, MGMT, MYC and BRAF, which are implicated in oncogenic processes such as angiogenesis, apoptosis, cell growth, and differentiation. For TCGA-KIRC, we used the most amplified /deleted genes (all CNVs with greater than 7% amplification or deletion), which yielded 117 CNV features. For both projects, we included RNA-Seq expression, which is measured as the quantified bulk abundance of mRNA transcripts. Using cBioPortal, we selected the top 240 differentially expressed genes for both projects [65]. Because our genomic features do not share any explicit spatial or temporal dependencies with each other, we feed our features

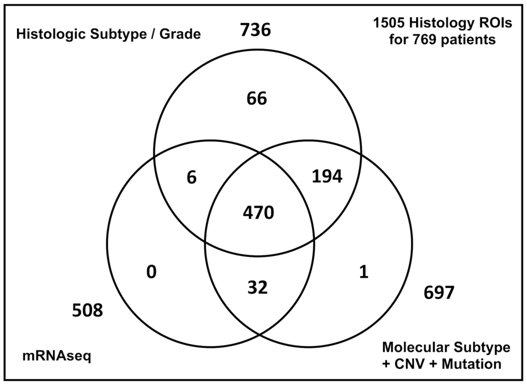

Fig. 7: Missing data in the TCGA-GBMLGG project. Because not all genomic features are available for all patients, Pathomic Fusion was trained with different subsets of the data than the unimodal networks for both survival outcome prediction and grade classification.

through a Self-Normalizing Feedforward Network to learn a low-dimensional representation before fusing it with histology and cell graph modalities.

B. Data Missingness and Alignment in TCGA-GBMLGG

To draw comparisons with the previous state-of-the-art, we used the existing curated TCGA-GBMLGG data found in the supplement of [29], which required in careful handling of missing values in multimodal data. For each patient, 1-3 20x 1024×1024 histology ROIs (0.5 μ / pixel) from diagnostic slides, and 320 genomic features were used. Of the 769 patients, 72 patients have missing molecular subtype (IDH mutation and 1p19q codeletion), 33 patients have missing histological subtype and grade labels, and 256 patients have missing mRNA-Seq data (Fig. 6). Because multiple ROIs from diagnostic slides were obtained for some patients, each image was treated as a single data point in cross-validation, with the genomic and ground-truth label information copied over. A 15-fold Monte-carlo cross-validation was conducted using the same train-test splits as the supplement of [29], which were generated randomly with 80% training and 20% testing (split by TCGA ID). Depending on the task (survival prediction vs. grade classification) and combination of modalities used (histology vs. genomic vs. histology+genomic), different subsets of the train split were used to train the unimodal and multimodal networks due to missingness. In validating our models on the test splits of the cross-validation with missing data, test splits were standardized to exclude all missing data across all models (center overlap in Figure 5.). Data missingness was not an issue in working with CCRCC, with all unimodal and multimodal networks trained with the same train-test splits in a 15-fold cross-validation.

C. Network Architectures

Three different network architectures were used to process the three modalities in our problem: 1) a VGG19 CNN with batch normalization for histology images, 2) a GCN for cell spatial graphs, and 3) a Feedforward Self-Normalizing Network for molecular profiles. The VGG19 network consists of 16 convolutional, 3 fully connected and 5 max pooling layers, with 512 × 512 sized images used as input. Dropout probabilities of 0.25 were applied after the first two fully connected layers (of size 1024), with a mild dropout (p=0.05) applied after the last hidden layer (of size 32). Our GCN consisted of 3 GraphSAGE and Self-Attention Pooling layers with hidden dimension 128, followed by two linear layers of size 128 and 32. Lastly, our Genomic SNN consists of 4 consecutive blocks of fully-connected layers with dimensions [64, 48, 32, 32], ELU activation, and Alpha Dropout. For survival outcome prediction, all networks were activated using the Sigmoid function, with the output scaled to be between -3 and 3. For grade classification, all networks were activated using the Log Softmax to compute scores for each of the 3 WHO Grades.

Our multimodal network architectures consist of two components: 1) Gating-based Modality Attention, and 2) fusion by Kronecker Product. Each modality was gated using three linear layers, with the second linear layer used to compute the attention scores. For survival outcome prediction, the genomic modality was used to gate over the image and graph modalities, while for grade classification, the histology image modality was used to gate over the genomic and graph modality. Additional dimension reduction of the gated unimodal feature representations was performed to reduce the output size of the Kronecker product feature space in the trimodal network. In our trimodal network, for survival outcome prediction, the first and third linear layers for the genomic modality have 32 hidden units to maintain the feature map dimension, with the linear layers in the image and graph modalities having 16 hidden units in order to transform the feature representations into a lower dimension. For grade classification, we maintained the feature dimension of our histology image modality instead, and reduced the dimension of the graph and genomic modalities. No feature dimension reduction was done in bimodal networks in any tasks. For feature fusion, the Kronecker product of the respective unimodal feature representations for each modality was computed, creating feature maps of size: $[33 \times 33], [33 \times 33], [33 \times 17 \times 17]$ for our CNN$\otimes$SNN, GCN$\otimes$SNN, and CNN$\otimes$GCN$\otimes$SNN. To use the unperturbed unimodal features, we appended 1 to each feature vector before computing the Kronecker Product. Dropout layers with probability $(p = 0.25)$ were inserted after gating and computing the multimodal tensor.

D. Experimental Details

Pathomic Fusion was built with PyTorch 1.5.0, PyTorch Geometric 1.5.0, Captum 0.2.0, and Lifelines 0.24.6. Node features for cell graph construction were calculated by: 1) segmenting each nuclei, 2) using the Contours Features Toolbox in in OpenCV 4.2.0, 3) the Texture Feature Toolbox in , 4) self-supervised deep features using Contrastive Predictive Coding, and 5) PyFlann 1.6.14 for graph construction. Resources used in our experimentation include 12 Nvidia GeForce RTX 2080 Tis on local workstations, and 2 Nvidia Tesla V100s on Google Cloud. The Histology CNN was initialized using pretrained weights from ImageNet, followed by finetuning the network using a low learning rate of 0.0005 and a batch size of 8. Random crops of 512 × 512, color jittering, and random vertical and horizontal flips were performed of data augmentation. The Histology GCN and Genomic SNN were initialized using the self-normalizing weights from Klambeur *et al.* [58], and trained with a learning rate of 0.002 with a batch size of 32 and 64 respectively. For the Genomic SNN, a mild $\mathcal{L}_1$ regularization was also used with hyperparameter value 3e-4 to enforce feature sparsity. All networks were trained with the same epochs using the Adam optimizer, dropout probability $p = 0.25$, and a linearly decaying learning rate scheduler.

After training the Histology CNN, for each 1024×1024 histology ROI, we extract $[32 \times 1]$ embeddings from 9 overlapping 512×512 patches, which we pair with their respective cell graph and genomic feature input as input into Pathomic Fusion. For the Histology GCN and Genomic SNN, we first trained their respective unimodal networks with the aforementioned training details, and then trained the last linear layers of the multimodal network with the unimodal network modules frozen with a learning rate of 0.0001 and Adam solver. At epoch 5, we unfroze the genomic and graph networks, and then trained the network for 25 more epochs using a learning rate of 0.0001, Adam solver, and a linearly decaying learning rate scheduler.

E. Evaluation Details

The predicted hazard and grade scores from each unimodal and multimodal network were evaluated on the test splits of the 15-fold cross-validation. To use the entire 1024×1024 histology image for CNN-based survival outcome prediction on TCGA-GBMLGG, similar to previous work, we computed the mean of hazard predictions from 9 overlapping 512×512 image crops across all histology ROIs belonging to each patient. For plotting the Kaplan-Meier curves, we pooled predicted hazards from all of the test splits in the 15-fold cross-validation and plotted them against their survival time. For creating the Swarm plots, we z-scored predicted hazards in each split before pooling so that scores for low vs. intermediate risk would have similar ranges in visualization. For grade classification on TCGA-GBMLGG, we used the max softmax activation score from overlapping 512×512 patches to determine class. For CNN-based survival outcome prediction on CCRCC, we similar computed the mean of hazard predictions from 512×512 histology ROIs for each patient.

Appendix C. Ablation Studies and Comparative Analysis

TABLE III: Concordance Index & statistical significance of Pathomic Fusion and ablation experiments in glioma survival prediction.

Model	c-Index ↑	[0,50] vs. (50,100] ↓	[0,33] vs. (33,66] ↓	[33,66] vs. (66,100] ↓
Cox (Age+Gender)	0.732 ± 0.012*	1.90e-92	1.48e-38	5.93e-27
Cox (Grade)	0.738 ± 0.013*	6.00e-255	9.57e-23	2.94e-66
Cox (Molecular Subtype)	0.760 ± 0.011*	2.07e-228	4.65e-26	1.45e-51
Cox (Grade+Molecular Subtype)	0.777 ± 0.013*	5.29e-215	1.14e-40	5.02e-52
Histology CNN	0.792 ± 0.014*	5.09e-40	1.77e-07	6.61e-25
Histology GCN	0.746 ± 0.022*	1.62e-21	2.29e-03	4.20e-15
Genomic SNN	0.808 ± 0.014	1.52e-52	0.153	2.16e-81
SCNN (Histology Only) [29])	0.754*	2.08e-61	-	-
GSCNN (Histology+Genomic) [29])	0.781*	3.08e-64	-	-
Pathomic F. (GCN⊗SNN)	0.812 ± 0.010*	1.06e-55	0.209	3.09e-80
Pathomic F. (CNN⊗SNN)	0.820 ± 0.009*	5.18e-57	0.103	4.56e-79
Pathomic F. (CNN⊗GCN⊗SNN)	**0.826 ± 0.009***	7.09e-57	2.68e-03	5.82e-74

$^{**}p < 0.05$

TABLE IV: Concordance Index & statistical significance of Pathomic Fusion and ablation experiments in CCRCC survival prediction.

Model	c-Index ↑	[0,50] vs. (50,100] ↓	[0,25] vs. (25,50] ↓	[25,50] vs. (50,75] ↓	[50,75] vs. (75,100] ↓
Cox (Age+Gender)	0.630 ± 0.024*	1.27e-16	0.108	1.2e-05	0.360
Cox (Grade)	0.675 ± 0.036*	4.42e-17	1.25e-07	4.52e-04	0.513
Histology CNN	0.671 ± 0.023*	3.87e-16	0.481	1.74e-04	4.19e-04
Histology GCN	0.648 ± 0.031*	4.23e-02	0.012	0.651	0.144
Genomic SNN	0.685 ± 0.024*	8.84e-19	0.480	0.013	7.07e-16
Pathomic F. (GCN⊗SNN)	0.688 ± 0.029*	1.65e-17	0.301	0.069	1.79e-12
Pathomic F. (CNN⊗SNN)	0.719 ± 0.031*	1.11e-27	0.772	7.00e-6	5.77e-12
Pathomic F. (CNN⊗GCN⊗SNN)	**0.720 ± 0.028***	2.48e-24	0.087	9.37e-3	1.08e-14

$^{**}p < 0.05$

F. Ensembling Effects

In order to further validate the performance improvement in multimodal networks presented in Table I we conduct ensambling experiments. In other words, for a fair comparison we compare the performance of our multimodal networks against CNN⊗CNN, GCN⊗GCN and SNN⊗SNN in order to rule out ensembling effects causing the observed improvement. Table II summarizes these results and we demonstrate that the improvement from Pathomic Fusion is greater than fusing the same modality using the same architecture.

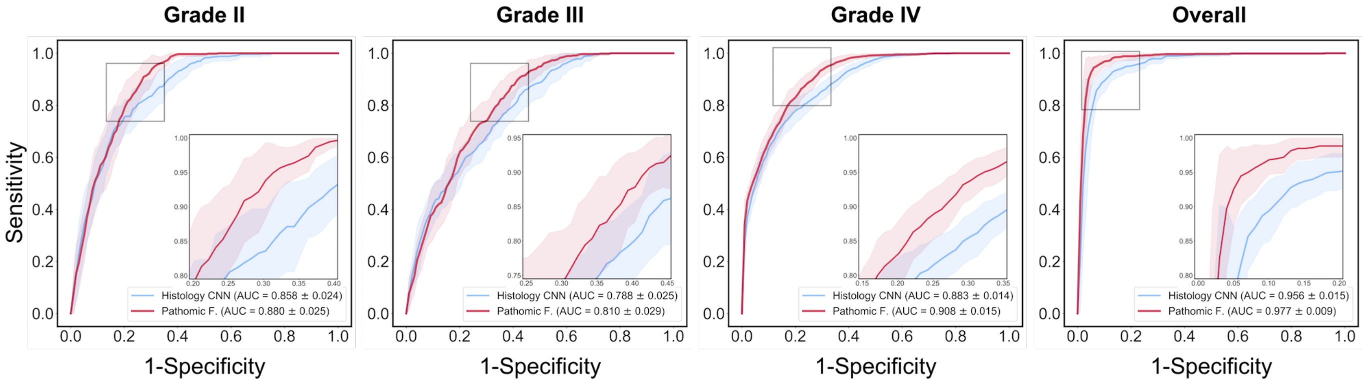

Fig. 8: Comparative analysis of AUC curves for Histology CNN and Pathomic Fusion in grade classification. The confidence interval is representative of the 15-fold cross validation. Since grade is usually determined via histology, our Genomic SNN that uses only CNV, gene mutation and chromosome deletion is not predictive of grade. Pathomic Fusion has greater AUCs in all cases, and performs particularly well on grade IV potentially because IDH mutation and 1p19q co-deletion from the genomic profile aid in discriminating Grade IV astrocytomas and Grade II/III oligodendrogliomas.

TABLE V: Comparative analysis of the ensembling effects of unimodal networks. Overall, ensemble models of Histology CNN, Histology GCN, and Genomic SNN caused networks to overfit and decrease in performance. Improvements made by ensembling were marginal compared to improvements made by Pathomic Fusion.

Model	c-Index ↑	AUC ↑	AP ↑	F1-Score (Micro) ↑	F1-Score (Grade IV)	c-Index ↑
Histology CNN	0.750 ± 0.010	0.883 ± 0.008	0.793 ± 0.017	0.717 ± 0.017	0.873 ± 0.013	0.671 ± 0.023
Histology (CNN + CNN)	0.749 ± 0.010	0.888 ± 0.007	0.807 ± 0.013	0.715 ± 0.021	0.879 ± 0.015	0.671 ± 0.023
Histology GCN	0.722 ± 0.014	0.849 ± 0.011	0.764 ± 0.012	0.665 ± 0.019	0.849 ± 0.011	0.648 ± 0.031
Histology (GCN + GCN)	0.720 ± 0.014	0.851 ± 0.015	0.763 ± 0.021	0.650 ± 0.023	0.812 ± 0.022	0.643 ± 0.027
Genomic SNN	0.808 ± 0.014	0.853 ± 0.012	0.729 ± 0.018	0.652 ± 0.015	0.857 ± 0.017	0.684 ± 0.025
Genomic (SNN + SNN)	0.794 ± 0.014	0.850 ± 0.012	0.725 ± 0.019	0.651 ± 0.018	0.856 ± 0.017	0.684 ± 0.025
Pathomic F. (GCN⊗SNN)	0.812 ± 0.010	0.897 ± 0.010	0.812 ± 0.016	0.714 ± 0.018	0.902 ± 0.014	0.686 ± 0.024
Pathomic F. (CNN⊗SNN)	0.820 ± 0.009	0.905 ± 0.010	**0.833 ± 0.016**	0.730 ± 0.019	0.913 ± 0.011	0.719 ± 0.031
Pathomic F. (CNN⊗GCN⊗SNN)	**0.826 ± 0.009**	**0.908 ± 0.008**	0.828 ± 0.016	**0.749 ± 0.020**	**0.920 ± 0.014**	**0.720 ± 0.028**

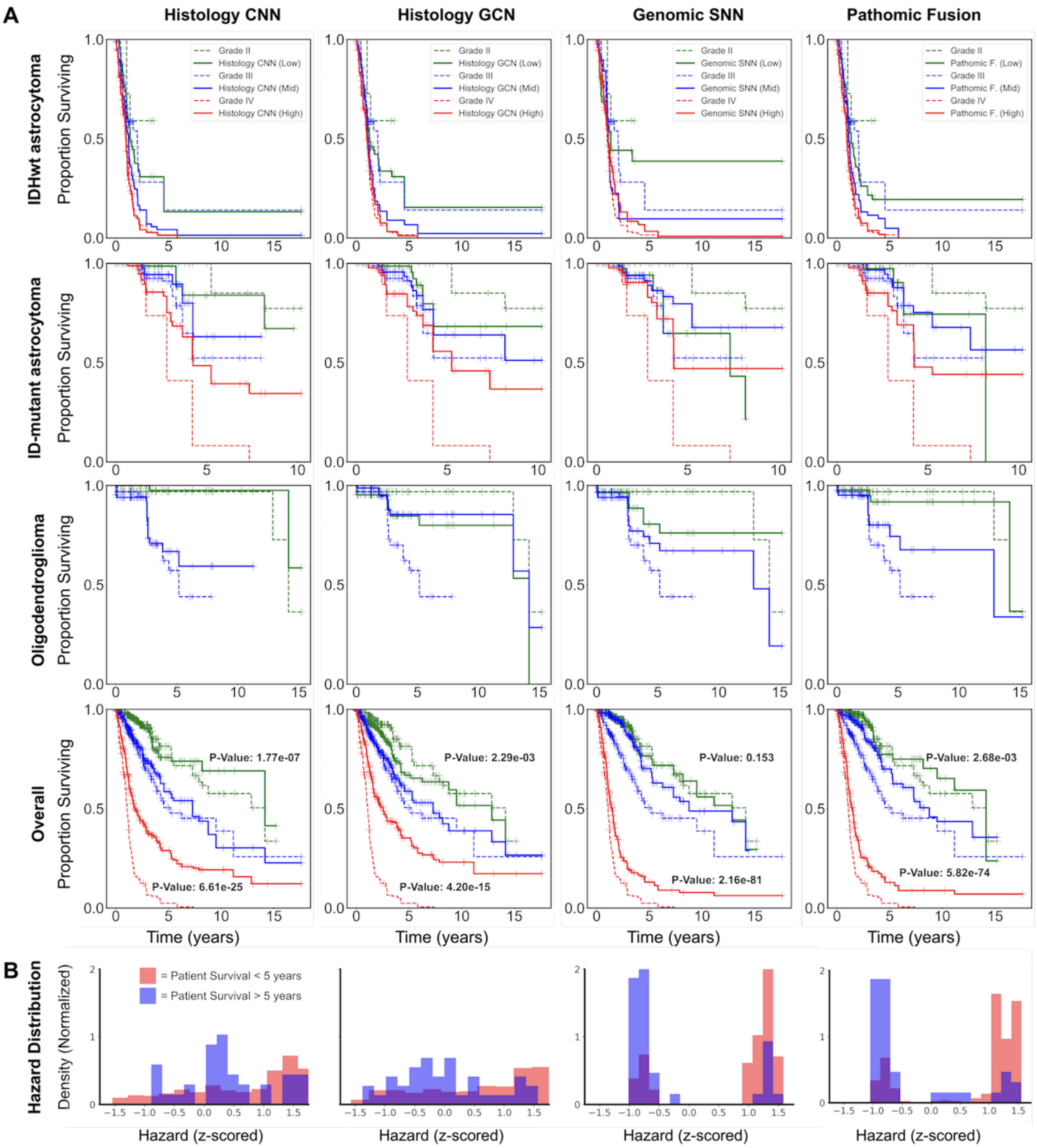

Fig. 9: **A.** TCGA-GBMLGG Kaplan-Meier comparative analysis of Histology CNN, Histology GCN, Genomic SNN, and Pathomic Fusion with respect to IDHwt ATCs, IDHmut ATCs, ODGs, and all molecular subtypes in stratifying WHO Grades II, III, and IV using the 33-66-100 percentile of hazard predictions. Overall, we observe that this heuristic has similar stratification of patients as the WHO grading system, with Pathomic Fusion having the closest resemblance. **B.** Distribution of hazard predictions for Histology CNN, Histology GCN, Genomic SNN, and Pathomic Fusion. Histology CNN and Histology had similar skewed distributions of hazard. Qualitatively, the distribution of hazard predictions by Genomic SNN is divided into clusters, while Pathomic Fusion produced is able to delineate three clusters.

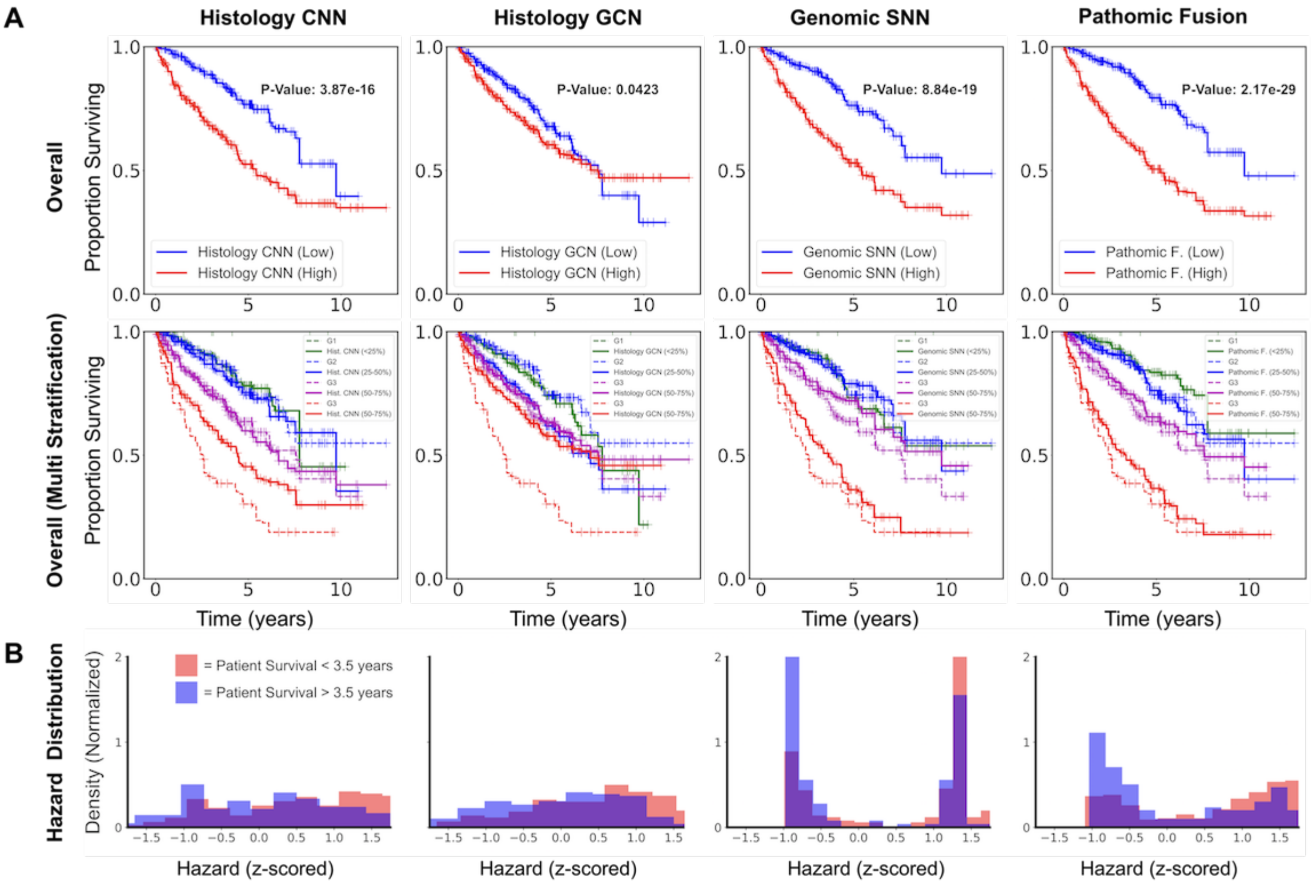

Fig. 10: **A.** CCRCC Kaplan-Meier comparative analysis of Histology CNN, Histology GCN, Genomic SNN, and Pathomic Fusion in stratifying low vs. high survival using the 50-100 percentile of hazard predictions, and Fuhrman Grades I, II, III, and IV using the 25-50-75-100 percentile of hazard predictions. In performing fine-grained stratification with four risk categories, Pathomic Fusion performed the best in disentangling each category. **B.** Distribution of hazard predictions for Histology CNN, Histology GCN, Genomic SNN, and Pathomic Fusion.

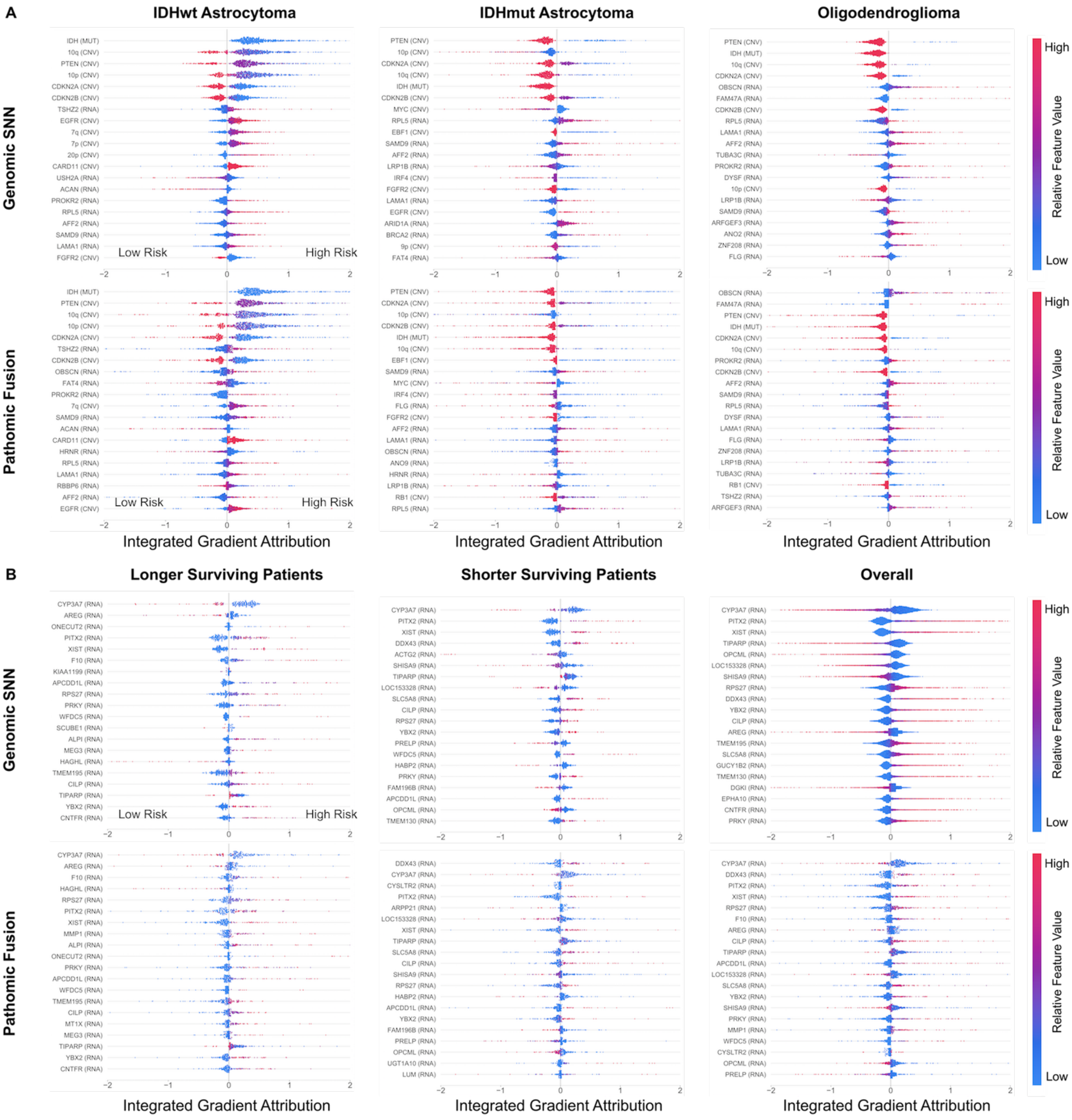

Fig. 11: Genomic SNN and Pathomic Fusion global explanation across patient cohorts in TCGA-GBMLGG and CCRCC. Attribution color corresponds to low (blue) vs. high (red) feature value, and attribution direction corresponds to how the gene feature value contributes to low risk (left) vs. high risk (right). Data points in the summary plots correspond to local explanations made by Integrated Gradients when attributing features for a given sample. Top 20 features were ranked by mean absolute attribution. **A.** Global explanation for each molecular subtype in TCGA-GBMLGG. OBSCN, FAT4, HRNR, RPL5, RBBP6, RB1, and ANO9, FAM47A feature importance increased when conditioned on morphological features, while EGFR feature importance decreased. **B.** Global explanation for longer surviving, shorter surviving, and all patients in TCGA-KIRC. Longer and shorter surviving patient cohorts were defined by the top 25 longest and shortest surviving patients respectively. MMP1, HAGHL, and ARRP21 feature importance increased when conditioned on morphological features.