A Deep Learning Multimodal Approach to Survival Risk Prediction in Kidney Cancer Patients

Hieu Nguyen
University of Chicago
nguyenhieu@uchicago.edu

Abstract

Cancer prognosis relies on heterogeneous data sources. To leverage the Cancer Genome Atlas (TCGA) data, deep learning models extract complex features and predict clinical outcomes. This work proposes a multimodal fusion strategy integrating clinical, RNA-seq, and histopathology data for survival prediction. Efficient embeddings and feature extraction models are constructed for each modality, followed by concatenation to optimize the negative Cox partial log-likelihood. Validated on TCGA's clear cell renal cell carcinoma dataset, the model achieves a high concordance index, primarily driven by clinical and RNA-seq features. This study lays the groundwork for future foundation model experiments and interpretability methods, such as attention-based heatmaps and integrated gradients, to enhance clinical relevance.

Code and trained model are made available at: https://github.com/quanghieu31/multimodal-ccRCC.

1 Introduction

Cancer prognosis via survival prediction aids biomarker discovery, patient stratification, and treatment response [1]. Understanding tumor microenvironments has improved prognosis [2], and advances in medical technology have enabled machine learning to analyze complex cancer data [3, 4].

Oncologists rely on both histology and genomic data, but histology often lacks genomic integration [1]. Whole-slide images provide morphological details but vary by pathologist, while genomic data offer molecular insights but cannot isolate tumor-specific transcripts. Fusing these modalities enhances early survival prognosis and personalized treatment [1], with deep learning well-suited for this task.

This study explores a deep learning approach with structured embeddings and feature extraction for survival risk prediction (Figure 3) in clear cell renal cell carcinoma (ccRCC), the most common kidney cancer [5]. The model integrates clinical, RNA-seq, and histopathology data, optimizing Cox partial likelihood and evaluating performance via the concordance index (c-index) and Kaplan-Meier curves. Since prognosis depends on multiple factors, early risk identification is key, as smaller tumors generally indicate better survival and treatment feasibility.

Contributions: (1) rationale for embeddings in each modality and (2) clustering and attention-based feature extraction for histopathology slides.

2 Related Work

Cancer prognosis via survival prediction has gained traction due to the availability of cancer data. RNA-seq expression data have been used in regression models to assess gene contribution to sur-

CMSC 35440: Machine Learning in Biology and Medicine, Winter 2025, University of Chicago, IL. (LaTeX template borrowed from NeurIPS 2023)

vival, while other modern transcriptomic sequences accelerate the understanding of mutations and abnormalities [6, 7]. Multiple Instance Learning networks are recent developments that improve the feature extraction for whole slide images with efficient computation [8]. Another study develops a deep learning system to predict disease-specific survival across ten cancer types using histopathology images, employing a weakly supervised approach without pixel-level annotations and testing three survival loss functions [9]. It significantly improves risk stratification over traditional clinical variables, achieving high concordance index and retaining predictive value in multivariable analyses. Another one classifies non-small cell lung cancer types and predicts mutations on whole-slide images [10]. The traditional method often uses clinical features like tumor grades, staging, gender, or age, and fits into a Cox proportional hazards model [11].

Multimodal fusion via deep learning introduces more methods to experiment. It aims to understand the heterogeneity across data modalities, make discoveries and interpretations, and enhance human-computer interaction, especially in medicine. There are many ways to combine the modalities in this field. Natural language processing and convolutional neural networks are used to extract clinical features across ICU sources to predict clinical intervention [12]. Most works have focused on establishing correspondences between histology tissue and genomics: Pathomic Fusion model integrating histopathology and genomic features for cancer prognosis in glioma and ccRCC, correlation between gene expressions and cellular features or joint analysis of breast cancer histologic images and genomic covariates [1, 13, 14]. $TransSurv$ is one of many prominent models that combine multi-scale pathological and multi-omics representations in a transformer-based mechanism to generate comprehensive, interpretable survival predictions [15].

3 Dataset and Embeddings

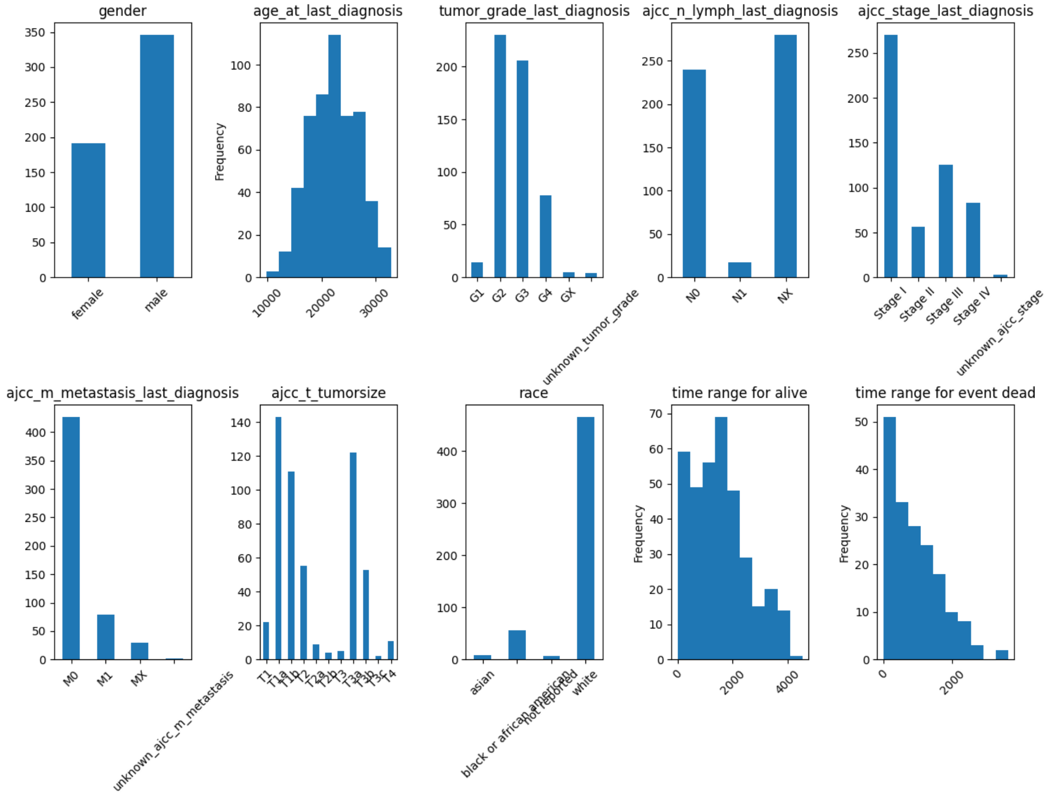

Figure 1: Clinical feature distributions

Three data modalities come from the TCGA-KIRC, hosted on the Genomic Data Commons Data Portal, National Cancer Institute [3]. A few patients with missing RNA-seq gene expression data are removed. In total, there are 533 patient-level samples with full clinical data, RNA-seq expressions, and histopathology slides. Events are balanced in which two-third of the samples are alive and the rest are dead.

Most categorical clinical variables are imbalanced, which may lead to biased predictions. Only gender, age, staging, and tumor grade are selected as the clinical features used in this study. The first three features align with the choices made in the literature [1] but I added tumor grade since tumor grade and staging are not highly correlated and could contribute to the survival probability.

Variable distributions are in Figure 1. The age variable is standardized because it is normally distributed and has relatively large values compared to the other variables. Gender is binary encoded. Tumor grade and staging are one-hot encoded because linear models (Cox proportional hazards model) and neural networks (fully connected layers) will be used on these data. That is, linear models and neural networks often access the influence of each category independently. If label encoding is used and ordinal relationship assumption might not be accurate, using one-hot encoding is more reliable and convenient since we only have a few categories per tumor grade or staging (low cardinality). After pre-processing, there are 13 clinical features.

RNA-seq gene expression counts are high dimensional (19962 genes per sample). Only protein-coding genes are retained before normalization to reduce noise from non-coding RNA, ensuring biological relevant expressions [16].

The data from TCGA-KIRC also have several types of normalizations: raw counts, FPKM, or TPM. Initially, I had some confusion about the within-sample and between-sample normalizations. FPKM is not suitable for between-sample comparisons because the total number of FPKM normalized counts for each sample will be different [17]. For example, sampleA has a higher proportion of counts for XCR1 (5.5/1,000,000) than SampleB (5.5/1,500,000), despite identical FPKM values. Thus, XCR1 counts (or any gene) cannot be directly compared between samples due to differing total normalized counts. TPM is counts per length of transcript per million reads mapped that normalizes sequencing depth and gene length. TPM could be useful for between-sample comparisons. But since I want to start from a simpler normalization method first in this study, CPM (counts scaled by total number of reads) is decided to apply to the protein-coding RNA genes to ensure that the gene matrices used in all analyses were the same [16, 17]. In particular, CPM of a gene count j for a patient i is,

$$CPM(i,j) = \frac{\text{rawcount}_{i,j}}{\text{totalreads}_i} \cdot 10^6$$

RNA-seq experiments generate different total read counts per sample, so dividing by total reads ensures gene expression values are comparable across samples. Then, often raw counts for each gene are small and so the ratio by total reads is also small, thus a multiplication of 10^6 avoids tiny decimal values. This scaling factor is also consistent with other popular normalization methods.

A log1P transformation is also applied on each gene to account large counts. Finally, since some patients have more one RNA samples, these will be average so that every patient has one single 19962-dimension vector of normalized gene counts.

As the final modality, each patient has different number of the histopathology slides. The maximum number of slides being considered in this study is capped at 5. Since the slides are large in size, the standard practice [10] is to get 20x magnification tiles (500x500) and 1000 random tiles are extracted from each slide. Also, the resolution level is 2 since my machine cannot read higher resolution (i.e. lower levels) due to memory issue. Then, a ResNet18 with ImageNet weights (the last classification layer is dropped) is used to convert these tiles into 512-dimensional representations. The choice for the number of 1000 is based on a few experiments to see which number gives not too dense and not too sparse tiles across slides on average. The rationale for ResNet18 is that it is a common, state-of-the-art feature extraction model for image data, and its parameters are reasonable to run on my machine. ResNet18 helps detect the key features on each tile like the edges or color intensity.

Before inputting these features into the main model, a KMeans ($k = 5$) clustering method is used to group the tiles into 5 clusters. This approach was also done in [8]. The purpose is to localize the representations with an expectation that some clusters might be clinically different than the others; that is, some clusters might observe tumor cells or complications. The choice for number 5 is based on intuition and would require more experiments to find the most optimal number.

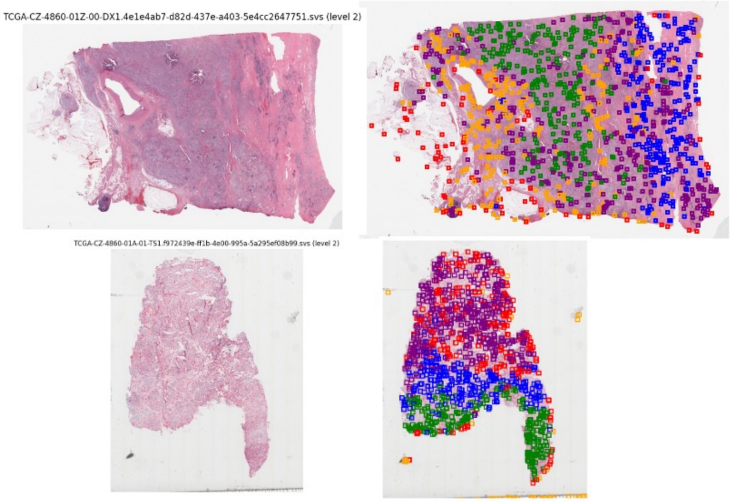
TCGA-CZ-4860-01Z-00-DX1.4e1e4ab7-d82d-437e-a403-5e4cc2647751.svs (level 2)
TCGA-CZ-4860-01A-01-TS1.f972439e-ff1b-4e00-995a-5a295ef08b99.svs (level 2)

Figure 2: An example of clustering on 2 histopathology slides

Therefore, each patient's histopathology slides can be represented as a list of 5 tensors, each tensor can have different shapes. For example, if a patient has 3 slides, which means 3000 tiles, then after clustering, they can have 5 tensors of shape (500, 512), (1000, 512), (300, 512), (200, 512), and (1000, 512) respectively. The clustering is applied across the total tiles of all the slides (Figure 2). That is, cluster 1 (blue) can have tiles from all the slides.

4 Methodology

4.1 Loss Function and Evaluation

The Cox proportional hazards model is a common semi-parametric approach for estimating the hazard function in survival analysis [11]. It assumes that the proportional hazards assumption which is satisfied in this dataset and that the hazard function can be parameterized as an exponential linear function $h(t|x) = h_0(t)e^{\beta x}$ at time t and with features x of a patient and learnable model parameters β, in which $h_0(t)$ is the baseline hazard that reflects how the risk of an event changes over time. It is hard to specify $h_0(t)$ for each patient, challenging the β training. Fortunately, the Cox partial log-likelihood (PL), without needing to specify $h_0(t)$ measures the relative probability of a patient's risk against the other patients who are still censored [18]. Assuming these probabilities are independent between the patients, the PL of n patients in a batch with observed/uncensored events (dead, $D = 1$),

$$PL(\beta, t) = \prod_{i \in D=1}^{n} \frac{e^{\vec{x}_i^T \beta}}{\sum_{j \in R(t)}^{|R(t)|} e^{\vec{x_j}^T \beta}}$$

in which $R(t) = \{j | t_j > t\}$ is the risk set (i.e. patients that are still censored after time t). Treating this as a loss function in a maximum likelihood estimation, it is more convenient to work with derivatives by turning it into a summation with logarithm and negating.

$$loss(\beta, t) = -\sum_{i \in D=1}^{n} \left(\vec{x_i}^T \beta - \ln \left(\sum_{j \in R(t)}^{|R(t)|} e^{\vec{x}_j^T \beta} \right) \right)$$

Intuitively, the goal is to maximize the difference between the predicted risk score from the model, $\vec{x_i}^T \beta$, of an uncensored/dead patient and the total predicted exponential risked scores of the ones in the risk set. Profiles of ones who pass away earlier should be considered at higher risk.

Regarding evaluation, concordance index measures ranking accuracy - for any two patients, the one predicted to be at higher risk actually experiences the event before the other [19]. It is the percentage of all comparable pairs (at least one must be uncensored). Secondly, the Kaplan-Meier is used to plot the survival probability over time for two groups divided at the predicted median risk.

4.2 Model Architecture

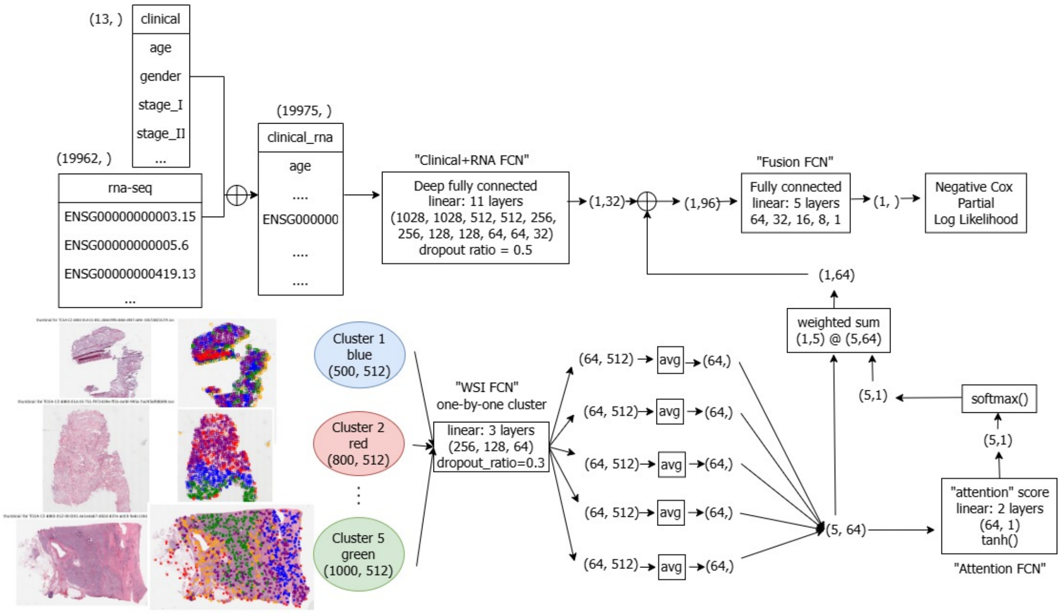

Figure 3: Model overview. “FCN" means fully connected network.

In Figure 3, the first part of the model is a simple concatenation of the processed clinical and RNA-seq expression features. The concatenated features are 19975 in dimension while there are 533 samples which is a relatively smaller number. Thus, overfitting is the primary issue to consider.

A deep fully connected neural network with heavy regularization (Clinical+RNA FCN) is a solution to this high dimensionality and feature extraction problem. Deep networks can learn hierarchical feature representations, while regularization techniques mitigate overfitting by enforcing sparsity and reducing reliance on specific feature [20]. In this case, dropout ratio is significantly useful because it prevents co-adaptation of neurons due to the deep fully connected layers [21]. By randomly dropping units, dropout reduces the risk of overfitting to the small training set for better generalization. Furthermore, gradually lowering the dimensions in each layer helps in progressive feature extraction, ensuring that the network captures essential information while filtering out noise. This structured reduction also aims to improve generalization by preventing the network from memorizing small training samples and learn lower-dimensional and more meaningful feature representations. For the clinical and RNA-seq data, the output dimension is 32.

In extracting the lower-dimensional features from the histopathology images, a notable thing is that the clustering is different between the samples. That is, cluster 1 of a sample and cluster 1 of another sample can refer to different characteristics of its corresponding set of tiles. The hypothesis is that the fully connected linear model for whole slide images (WSI FCN) can learn the difference between the clusters through training and attention so that it would finalize the weights to effectively reduce the 512 to 64 dimensions for each cluster. In general, WSI FCN takes one at a time, condenses it to 64 dimensions and applies an average pooling, before stacking the 5 64-dimensional clusters into a small Attention FCN with Tanh activation and then softmax to calculate the weights for each cluster. Intuitively, the goal of these two models is learn the difference between the clusters and how to place appropriate weights on each of them rather than deterministically indicating a cluster, say, is a tumor or not.

The Attention FCN model outputs 5 scores that are transformed with softmax to determine the weights being applied to each of the 5 clusters. The final output of the histopathology images is a tensor of shape 64.

The final concatenated features have the dimension of 96 before inputting to the Fusion FCN model. This linear model is relatively small and fuses the three modalities into one network before calculating the predicted risk score for the loss function. This is because all of the heavy and complex computations are expected to happen in the other feature extraction models. The Fusion FCN aims to be simpler and captures the interactions of the modalities.

4.3 Experiment Set-up

The training started with sets of conventional hyperparameters and regularization. This must also be constrained by the number of samples in training. That is, the training, validation, and test sets are of the 70/15/15 ratio. The experiments combinations of batch size of (32, 64), epochs of (5, 10, 15), learning rates of (0.001, 0.0001), weight decays of (0, 0.0001) and several attempts to change the dimensions of layers and number of layers in the Clinical+RNA FCN and WSI FCN models.

5 Results

The hyperparameters that give the highest c-index include batch size of 32, epochs of 5, weight decay of 0.001, and learning rate of 0.0001. The dropout ratios are 0.5 for the Clinical+RNA FCN and 0.3 for the WSI FCN model. In neural Clinical+RNA FCN, it seems that more hidden layers and dimensions per layer are correlated with lower validation loss and high c-index (holding others constant).

The multimodality model is compared to 3 baseline unimodal models. The clinical baseline is based on the age, gender, tumor grade, and staging covariates, as discussed in Section 2. They are fitted into a CoxPH model with a L1 penalizer of 0.01 due to the sparsity and high multicollinearity. The RNA-seq gene expression baseline involves a Principal Component Analysis (PCA) with 16 principal components that are reduced from 19962 genes. They are fitted into a CoxPH model without any regularization. Then, a baseline model on the histopathology slides utilizes the same embeddings as in the multimodality model and the same negative Cox partial log-likelihood function. That is, this baseline fits ResNet18 outputs with KMeans ($k = 5$) clustering along with the WSI FCN and Attention FCN feature extraction models. It has epochs of 5, learning rate of 0.001, and batch size of 32. Another simpler model could be used to evaluate this baseline and will be done in the near future. Table 1 summarizes the models and the c-index results.

Model	c-index (std)
Baseline - Clinical	0.7631 (0.0514)
Baseline - RNA-seq	0.7326 (0.0410)
Baseline - Histopathology	0.6868 (0.0560)
Multimodality	0.7753 (0.0427)

Table 1: C-index results. Mean c-index is calculated on test sets with 300 bootstraps with replacement.

All models have good results. The baseline model for histopathology slides has the c-index that is relatively lower than the other three. Meanwhile, the multimodal model has the highest c-index Clinical and RNA-seq features likely contribute most significantly to the multimodal model. That is, the two sets of features have more explaining power for the high c-index. It implies that the histopathology might take a small part in the multimodal effects.

Based on these c-index results only, combining the clinical, genomic, and histopathology slides, the approach outperforms unimodal models although the histopathology features do not add significant improvement over the clinical or RNA-seq gene expressions alone.

Nonetheless, the integration may provide some information in distinguishing survival Kaplan-Meier curves (Figure 4) of less aggressive tumors, especially in the heterogeneous nature of tumor microenvironment. We observe that the multimodality version allows for more fine-grained stratification of survival probabilities between the low and high risk groups. This information can be used to benefit risk cohorts and personalized treatments. Other than the RNA-seq baseline model, the other three Kaplan-Meier curves have wide confidence intervals. This might indicate a high variance among the two risk groups that needs to be investigated further in the future.

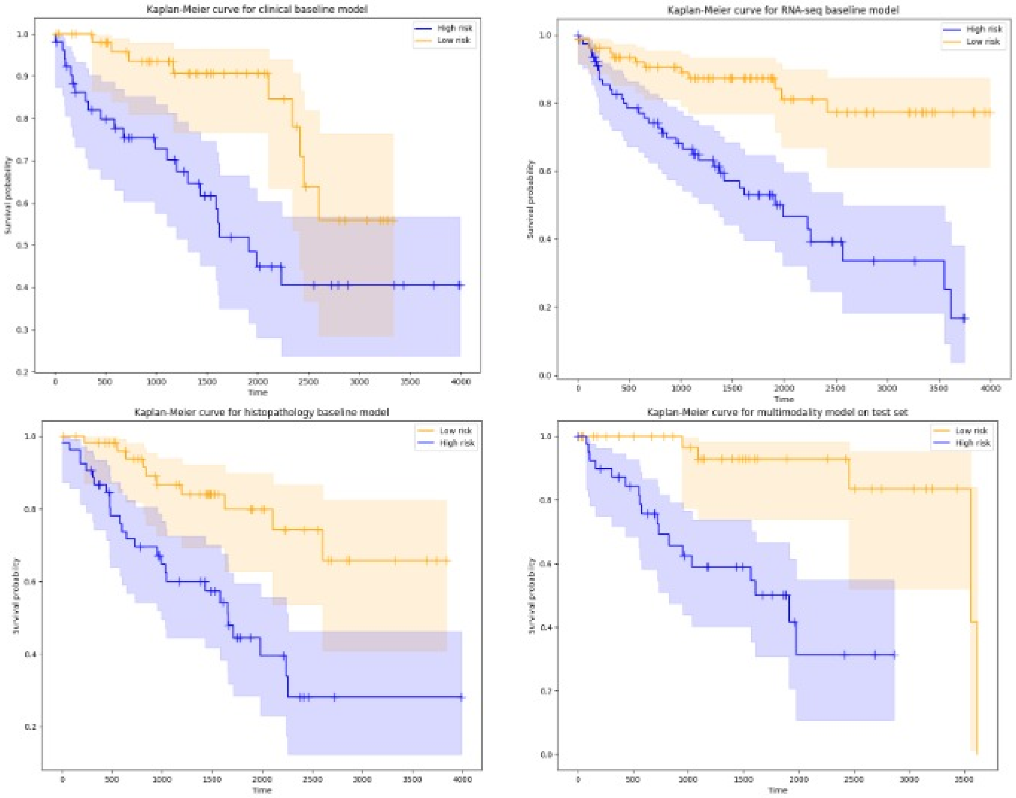

Figure 4: Kaplan-Meier curves

6 Conclusion and Lessons Learned

This study demonstrates that integrating clinical data, RNA-seq gene expressions, and histopathology images improves survival risk prediction compared to unimodal approaches. The multimodal fusion model achieved the highest concordance index, with clinical and RNA-seq features contributing most significantly. While histopathology images added some predictive power, their effect was relatively minor, suggesting the need for more experiments on feature extraction techniques to better capture their prognostic value.

One key takeaway is that foundation models for histopathology images could significantly enhance feature extraction by identifying the regions of interest better on the images. In this study, histopathology slides were processed using ResNet18 embeddings followed by clustering as in a separate step of the whole model. But more sophisticated models, such as self-supervised vision transformers or domain-specific histopathology networks, may better capture spatial and morphological patterns in tumor microenvironments. Future work should explore pre-trained histopathology-specific models or fine-tune large-scale vision models on relevant datasets. For example, UNI is a popular general-purpose self-supervised model for pathology that is pretrained on more than 100 million images from 100,000 diagnostic H&E-stained WSIs including ones from TCGA [22]. Not only does it reduce the preprocessing workload for histopathology images, but it also provides rich, domain-specific feature representations that could improve downstream survival prediction. This allows the network to learn hierarchical representations directly from raw whole-slide images in a fully end-to-end manner.

There is also a strong need for multimodal interpretability of the model. Survival risk prediction is mostly useful in cohort-based treatment and planning. One important questions that can be answered using the multimodality approach is the ability to evaluate the impacts of different data sources [1]. For example, given a prediction on survival risk, we want to attribute how local pixel regions in histopathology images, SHAP values for clinical features, and integrated gradients for RNA-seq data are used to produce that prediction. At the same time, a simpler approach would be to visualize the softmax attention scores from my own model. Overall, multimodal interpretability can help identify novel integrative biomarkers with diagnostic, prognostic, and therapeutic relevance [1].

Lastly, several choices in this model architecture can be arbitrarily empirical. More experiments can help with learning these embeddings better. For example, more combinations for the regularizations (dropout rate, weight decay) and training (learning rate, epochs, batch size, number of clusters) can be tested. During training and evaluation, stratifying the data splitting based on event distribution or cross-validation might also help.

The two biggest lessons I learned from this project are (1) the importance of thoroughly understanding the data and designing appropriate embedding strategies with clear rationale, and (2) critically reading research papers and public code to selectively extract only what is necessary for my work.

References

[1] Richard J. Chen, Ming Y. Lu, Jingwen Wang, Drew F. K. 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, April 2022.

[2] Douglas Hanahan and Robert A. Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674, March 2011.

[3] Allison P. Heath, Vincent Ferretti, Stuti Agrawal, Maksim An, James C. Angelakos, Renuka Arya, Rosita Bajari, Bilal Baqar, Justin H. B. Barnowski, Jeffrey Burt, Ann Catton, Brandon F. Chan, Fay Chu, Kim Cullion, Tanja Davidsen, Phuong-My Do, Christian Dompierre, Martin L. Ferguson, Michael S. Fitzsimons, Michael Ford, Miyuki Fukuma, Sharon Gaheen, Gajanan L. Ganji, Tzintzuni I. Garcia, Sameera S. George, Daniela S. Gerhard, Francois Gerthoffert, Fauzi Gomez, Kang Han, Kyle M. Hernandez, Biju Issac, Richard Jackson, Mark A. Jensen, Sid Joshi, Ajinkya Kadam, Aishmit Khurana, Kyle M. J. Kim, Victoria E. Kraft, Shenglai Li, Tara M. Lichtenberg, Janice Lodato, Laxmi Lolla, Plamen Martinov, Jeffrey A. Mazzone, Daniel P. Miller, Ian Miller, Joshua S. Miller, Koji Miyauchi, Mark W. Murphy, Thomas Nullet, Rowland O. Ogwara, Francisco M. Ortuño, Jesús Pedrosa, Phuong L. Pham, Maxim Y. Popov, James J. Porter, Raymond Powell, Karl Rademacher, Colin P. Reid, Samantha Rich, Bessie Rogel, Himanso Sahni, Jeremiah H. Savage, Kyle A. Schmitt, Trevar J. Simmons, Joseph Sislow, Jonathan Spring, Lincoln Stein, Sean Sullivan, Yajing Tang, Mathangi Thiagarajan, Heather D. Troyer, Chang Wang, Zhining Wang, Bedford L. West, Alex Wilmer, Shane Wilson, Kaman Wu, William P. Wysocki, Linda Xiang, Joseph T. Yamada, Liming Yang, Christine Yu, Christina K. Yung, Jean Claude Zenklusen, Junjun Zhang, Zhenyu Zhang, Yuanheng Zhao, Ariz Zubair, Louis M. Staudt, and Robert L. Grossman. The nci genomic data commons. *Nature Genetics*, 53(3):257–262, February 2021.

[4] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17, 2015.

[5] National Cancer Institute. Clear cell renal cell carcinoma, 2024. Accessed: 2024-03-08.

[6] 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), February 2019.

[7] Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, June 2011.

[8] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789, October 2020.

[9] 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, June 2020.

[10] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L. Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 24(10):1559–1567, September 2018.

[11] D. R. Cox. Regression models and life-tables. *Journal of the royal statistical society series b-methodological*, 34:187–220, 1972.

[12] Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding with deep neural networks. In Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 322–337. PMLR, 18–19 Aug 2017.

[13] 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, and J. S. Marron. Joint and individual analysis of breast cancer histologic images and genomic covariates. *The Annals of Applied Statistics*, 15(4), December 2021.

[14] Vaishnavi Subramanian, Benjamin Chidester, Jian Ma, and Minh N. Do. Correlating cellular features with gene expression using cca. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, page 805–808. IEEE, April 2018.

[15] Zhilong Lv, Yuexiao Lin, Rui Yan, Ying Wang, and Fa Zhang. Transsurv: Transformer-based survival analysis model integrating histopathological images and genomic data for colorectal cancer. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(6):3411–3420, November 2023.

[16] Hatice Büşra Lüleci, Dilara Uzuner, Müberra Fatma Cesur, Atılay İlgün, Elif Düz, Ecehan Abdik, Regan Odongo, and Tunahan undefinedakır. A benchmark of rna-seq data normalization methods for transcriptome mapping on human genome-scale metabolic networks. *npj Systems Biology and Applications*, 10(1), October 2024.

[17] HBC Training. Dge count normalization, 2024. Accessed: 2024-03-08.

[18] D. R. COX. Partial likelihood. *Biometrika*, 62(2):269–276, August 1975.

[19] FRANK E. HARRELL, KERRY L. LEE, and DANIEL B. MARK. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387, February 1996.

[20] Stefano Recanatesi, Matthew Farrell, Madhu Advani, Timothy Moore, Guillaume Lajoie, and Eric Shea-Brown. Dimensionality compression and expansion in deep neural networks, 2019.

[21] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, January 2014.

[22] Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, March 2024.