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

Hieu Nguyen
nguyenhieu@uchicago.edu

CMSC 35440 Machine Learning in Biology and Medicine, Winter 2025, University of Chicago

Project overview

Goal:

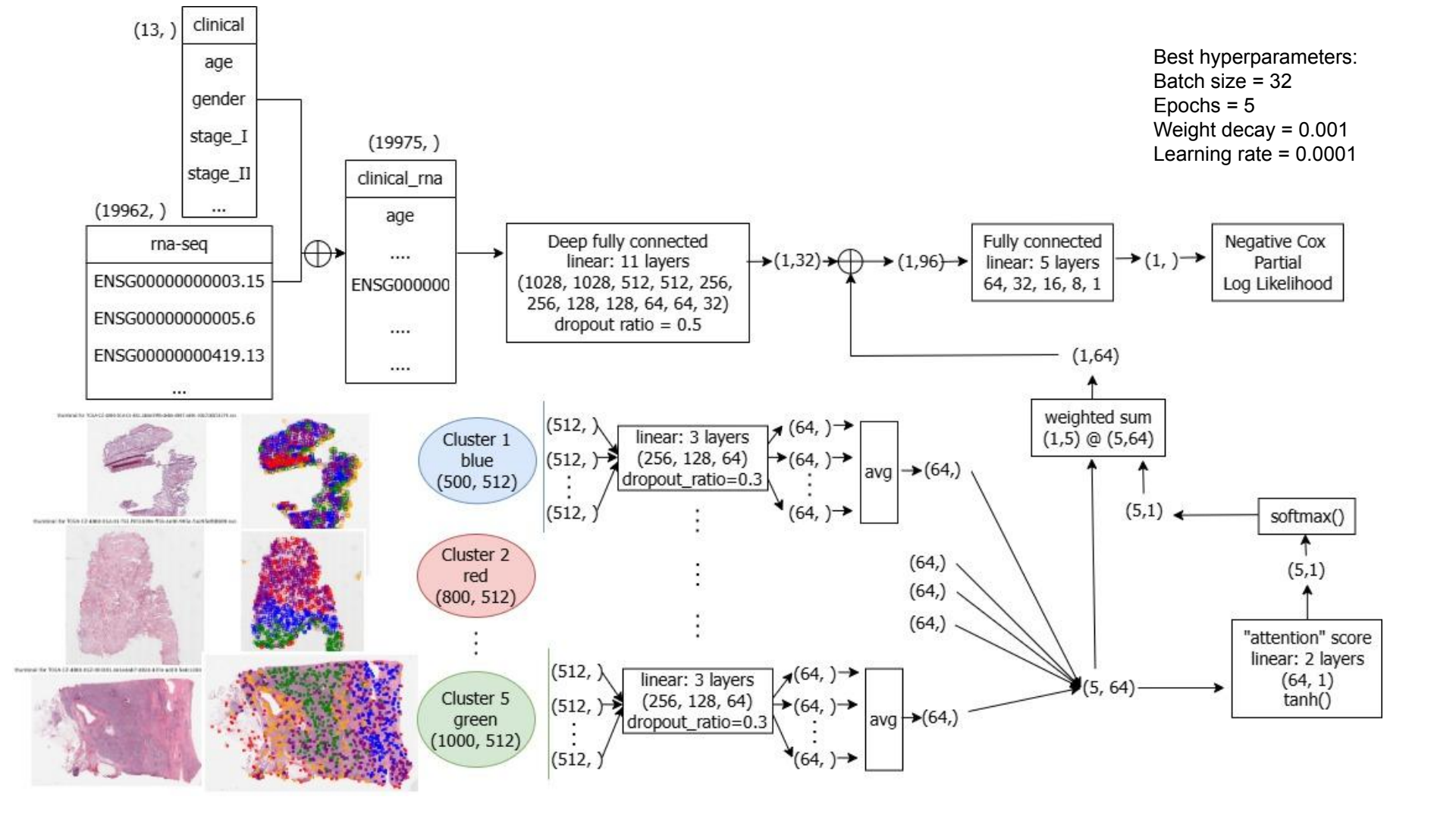
- Predict overall survival by stratifying two groups (high/low risk) for patients with clear cell renal cell carcinoma
- Extract features and fuse 3 modalities in an end-to-end model

Loss function: Negative Cox partial log-likelihood

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

Datasets and embeddings

- 533 patient-level samples with full data
- Train-validation-test: 70-15-15 ratio
- Labels: event (dead/alive), time (days from diagnosis to last checkup)
- Clinical data:
  - age (normalized), gender, tumor grade, staging (one-hot encoded)
  - 13 features
- RNA-seq gene expressions:
  - counts per million, log1p, and average counts for patients with multiple RNA samples
  - 19962 features
- Histopathology slides
  - Max 5 slides per patient, 1000 tiles per slide per patient
  - ResNet18 to convert these total tiles to i.e. (3000 tiles, 512) tensor
  - KMeans (k=5) to cluster these tiles

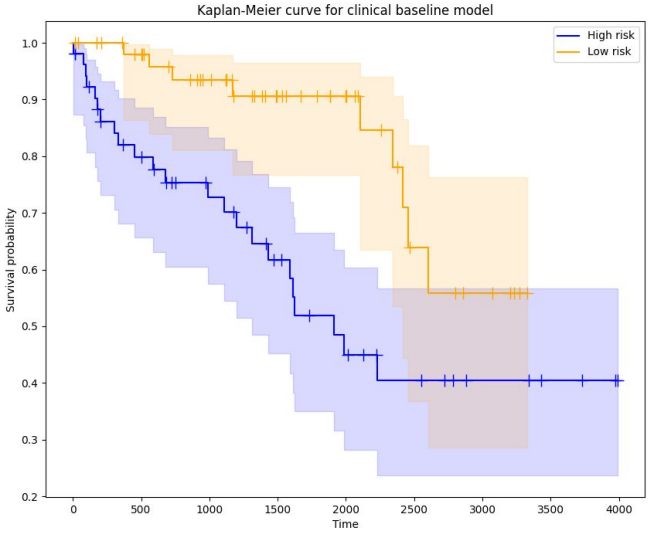

(13,)
clinical
age
gender
stage_I
stage_II
...
(19962,)
rna-seq
ENSG00000000003.15
ENSG00000000005.6
ENSG00000000419.13
...
(19975,)
clinical_rna
age
....
ENSG000000
....
....
Deep fully connected
linear: 11 layers
(1028, 1028, 512, 512, 256,
256, 128, 128, 64, 64, 32)
dropout ratio = 0.5
(1,32)
(1,96)
Fully connected
linear: 5 layers
64, 32, 16, 8, 1
(1,)
Negative Cox
Partial
Log Likelihood
(1,64)
Best hyperparameters:
Batch size = 32
Epochs = 5
Weight decay = 0.001
Learning rate = 0.0001
Cluster 1
blue
(500, 512)
Cluster 2
red
(800, 512)
Cluster 5
green
(1000, 512)
(512,)
(512,)
(512,)
linear: 3 layers
(256, 128, 64)
dropout_ratio=0.3
(64,)
(64,)
(64,)
avg
(64,)
weighted sum
(1,5) @ (5,64)
(5,1)
softmax()
(5,1)
(64,)
(64,)
(64,)
(5, 64)
"attention" score
linear: 2 layers
(64, 1)
tanh()

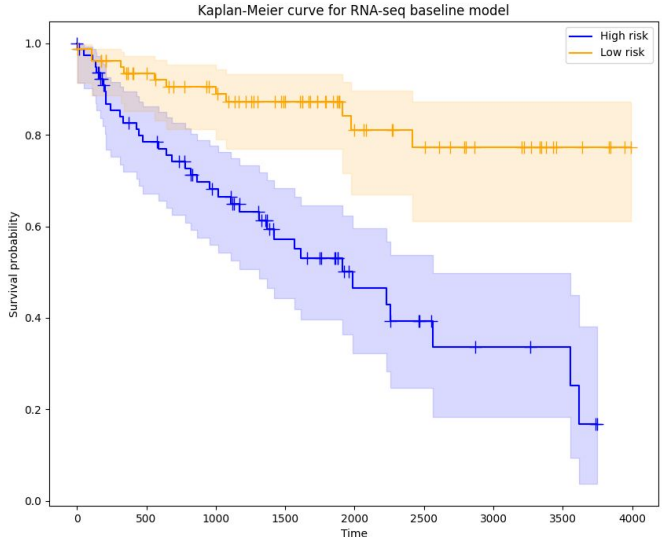
Results - Concordance Index

	Features	Model	**c-index (std)**
Baseline - Clinical	age, gender, tumor grade, staging	CoxPH penalizer l1_ratio=0.01	**0.7631 (0.0514)**
Baseline - RNA-seq gene expressions	PCA, n_components=16	CoxPH	**0.7326 (0.0410)**
Baseline - Histopathology slides	ResNet18 with ImageNet weights => (512,) vectors KMeans clustering (k=5)	Cox partial likelihood Feed-forward networks on each cluster Attention softmax layer at the end epochs=5, lr=0.001, batch size=32	**0.6868 (0.0560)**
Multimodality	Clinical+RNA: (32,) Histopathology: (64,) Concatenated: (96,)	Cox partial likelihood Feed-forward networks	**0.7753 (0.0427)**

mean c-index is calculated on test set (n=80) with 300 bootstraps with replacement

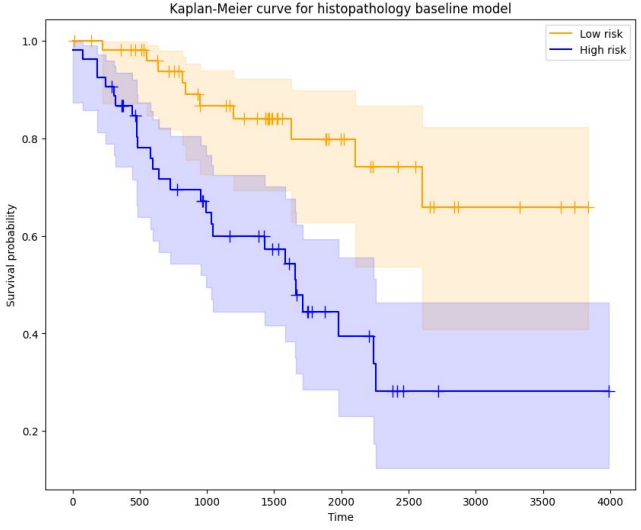
Wide CI
Not as clearly stratified

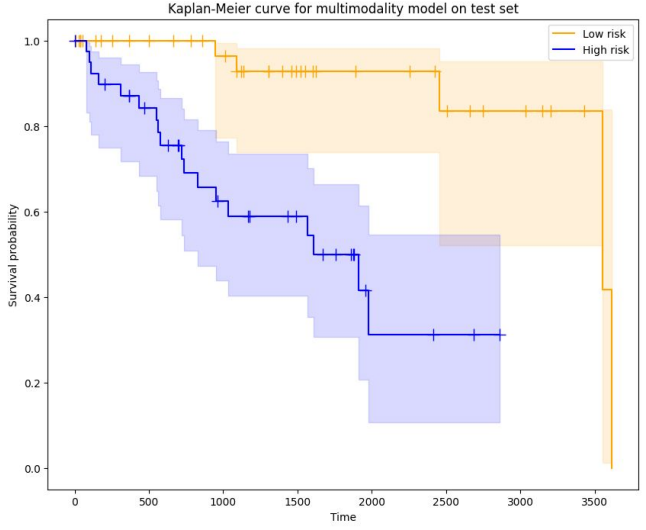




Moderate CI
Clearly stratified

Wide CI
Moderately stratified





Wide CI
Clearly stratified

Challenges and lessons learned

- Histopathology slides
  - Svs files are very large, needed to immediately delete the files after tiling
  - Only able to work with level=2 resolution which might not be as correct as the lower levels
  - Foundation models might identify ROIs better
- Model architecture
  - The choices for the layers, neurons, regularizations, and training hyperparameters were somewhat arbitrary
    => more slurm jobs help
  - Stratifying the data splitting or cross-validation might yield more accurate results
- Not much meaningful clinical inference yet so explainability is necessary
  - Grad-CAM on attention heatmap for the histopathology slides
  - Integrated gradient attribution for the gene expressions

Q&A