Multimodal Prognosis of Biochemical Recurrence in Prostate Cancer from mpMRI and Clinical Data

Abstract

We develop and compare multimodal survival models to predict time to biochemical recurrence (BCR) in prostate cancer from multiparametric MRI (T2, ADC, HBV) and clinical data. We first establish classical baselines based on Cox models, gradient boosting, random survival forests and FastSurvivalSVM trained on MRI-only, clinical-only and concatenated MRI+clinical features. We then design a multimodal attention-based multiple instance learning (ABMIL) model that operates on slice-level MRI features, aggregates them into patient-level representations and fuses them with clinical embeddings. All models are trained using the provided 5-fold split, and evaluated with the concordance index (C-index). The best classical model (clinical-only Cox / DeepSurv) reaches a mean C-index of ≈0.80, while the multimodal ABMIL model achieves ≈0.72 and clearly outperforms MRI-only approaches. Ablation studies demonstrate that clinical variables are indispensable for robust prognosis and that learned pooling performs slightly better than simple mean pooling.

1. Data and preprocessing

We used the provided CHIMERA-like dataset consisting of:

Multiparametric MRI:T2-weighted, ADC, and high b-value diffusion (HBV) volumes

Clinical variables such as age at prostatectomy, Gleason grades, ISUP score, pre-operative PSA, pT-stage and binary indicators of capsular penetration, lymph-node and seminal vesicle invasion

Time-to-event labels: time to last follow-up/BCR (months) and BCR status (0/1)

Imaging preprocessing (01_convert_resample_to_nii.py)

Convert raw .mha mpMRI and prostate masks to NIfTI, resampling ADC and HBV to T2 geometry and resampling the prostate mask with nearest-neighbor interpolation.

Save per-patient NIfTI paths into outputs/paths_index.csv.

Slice and feature extraction (033_extract_mri_bags.py)

For each patient, we select a fixed number of prostate-containing axial slices from all three sequences.

Crop a centered 2D patch around the prostate and apply robust intensity normalization within the mask.

We use pretrained MobileNet-based backbones from https://github.com/StandWisdom/MRI-based-Predicted-Transformer-for-Prostate-cancer to extract 576-d features per slice per modality and concatenate them to a 1728-d triple-modality vector per slice.

Slice-level features for all slices of a patient are saved as a bag $X \in \mathbb{R}^{S\times1728}$ in outputs/mri_bags/*.npz with an index mri_bags_index.csv.

Clinical table and labels (02_build_labels_and_tabular.py)

Parse JSON clinical files into a tabular dataframe with:

Continuous features (age, PSA, etc.),

Ordinal/encoded variables (Gleason, ISUP, pT-stage),

Binary indicators for lymph nodes, margins, capsular and seminal vesicle invasion, lymphovascular invasion, and earlier therapy.

The survival outcome is defined as:

event = 1 if BCR occurs, 0 otherwise;

time = months from surgery to BCR or last follow-up.

2. Classical baseline models

We first extracted patient-level MRI embeddings by median pooling slice features across the prostate, concatenating T2/ADC/HBV embeddings into a 1728-dimensional vector per patient .

Using these MRI embeddings and the clinical vectors, we trained a panel of classical survival models:

Cox elastic-net (MRI-only, clinical-only, concatenated MRI+clinical)

Gradient boosting survival analysis (GBSA) with CoxPH loss

Random survival forest (RSF)

FastSurvivalSVM (linear survival SVM)

DeepSurv (PyTorch-based MLP survival model)

For each model family, we evaluated three settings:

MRI-only – imaging embeddings only

Clin-only – clinical variables only

Concat – concatenated MRI + clinical features

Training protocol

5-fold cross-validation strictly following data_split_5fold.csv.

Within each fold:

Train on 4 folds, test on the held-out fold.

Standardize features using StandardScaler fitted on the training data only.

Use small inner CV for hyperparameters (e.g., SVM α-grid, Cox α-grid)

Evaluation metric: Harrell's C-index on the test fold.

Model	Modality	Mean C-index ± SD
COX	Clin	0.806±0.100
COX	MRI+Clin	0.778±0.070
COX	MRI	0.395±0.160
DeepSurv	Clin	0.804±0.088
DeepSurv	MRI+Clin	0.747±0.123
DeepSurv	MRI	0.520±0.051
GBSA	Clin	0.720±0.109
GBSA	MRI+Clin	0.694±0.086

GBSA	MRI	0.457±0.030
RSF	Clin	0.187±0.099
RSF	MRI+Clin	0.253±0.083
RSF	MRI	0.550±0.134
SVM	Clin	0.757±0.069
SVM	MRI+Clin	0.704±0.111
SVM	MRI	0.410±0.055

Clinical-only Cox and DeepSurv achieve the best overall C-index (~0.80), indicating that the limited MRI embeddings alone are not sufficient to outperform well-structured clinical baselines on this dataset. MRI-only models perform substantially worse, especially for Cox and SVM, whereas multimodal concatenation recovers part of the performance.

3. Multimodal ABMIL architecture

We then designed a multimodal attention-based multiple instance learning (ABMIL) model based on https://github.com/DIAGNijmegen/CHIMERA/blob/main/task1_baseline/README.md that directly operates on slice-level triple-modality MRI features and fuses them with clinical data.

Instance encoder

Each slice feature vector x ∈ ℝ^{1728} is passed through a 2-layer MLP with LayerNorm and ReLU to produce a hidden representation h_i ∈ ℝ^{H_inst}.

Positional / modality encoding

When available, the slice index along the superior–inferior axis and the modality indicator (T2, ADC, HBV) are projected via a small MLP and added to the instance features to provide weak spatial and modality awareness.

Gated attention pooling

Slice embeddings for a patient {h_i} are aggregated into a patient-level representation using gated attention, where:

Attention weights w_i are computed from a gated MLP with tanh and sigmoid branches.

Weighted sum H = Σ_i w_i h_i forms the bag-level representation.

This enables the model to focus on the most prognostic slices instead of uniform pooling.

Clinical encoder and fusion

Clinical features are standardized and passed through a separate MLP to obtain a c ∈ $\mathbb{R}$^{H_cls} embedding.

Imaging and clinical embeddings are concatenated and passed through a fusion MLP.

A final linear layer predicts a scalar risk score, which is optimized using a Cox partial likelihood loss.

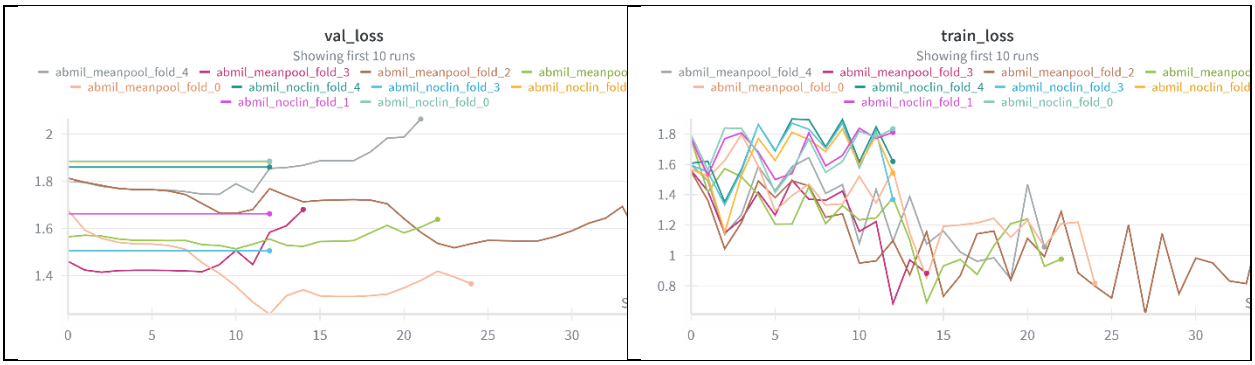
Training details

We use the same 5-fold CV split as for the classical models.

Within each fold, the training data are further split into train/validation (80/20) to monitor early stopping.

Optimization uses AdamW with cosine annealing schedule and early stopping based on validation loss.

We log train and validation losses, as well as test C-index per fold, to wandb.



4. Ablation study

We performed two sets of ablations:

Classical unimodal vs multimodal

Comparing MRI-only, Clin-only and MRI+Clin for Cox, DeepSurv, etc. (Table above).

ABMIL architectural ablations

Full model (“full”): attention pooling + clinical fusion.

No clinical (“noclin”): drop clinical branch, MRI-only MIL.

Mean pooling (“meanpool”): replace attention with simple mean over slices.

ABMIL results:

Full ABMIL: mean C-index = 0.707 ± 0.046

No clinical (MRI-only ABMIL): 0.415 ± 0.090

Mean pooling: 0.720 ± 0.050

Removing clinical features drastically degrades performance (~0.29 lower C-index), which confirms that clinical variables are crucial prognostic factors in this cohort. Interestingly, replacing attention with mean pooling slightly improves the C-index (0.720 vs 0.707), suggesting that with limited sample size the added flexibility of attention may not translate to better generalization, while mean pooling acts as a strong regularizer.

Comparing to classical baselines, ABMIL is competitive with some multimodal concatenation models, but still lags behind the best clinical-only Cox/DeepSurv baselines (~0.80), highlighting that the current MRI features, while informative, do not fully close the gap to clinical data.

Conclusion

In this assignment, we implemented and compared a range of classical and deep multimodal survival models to predict time to biochemical recurrence after prostatectomy. Clinical-only Cox and DeepSurv models achieved the strongest overall performance (C-index ≈ 0.80), while our multimodal ABMIL models reached C-indices around 0.71–0.72 and clearly outperformed MRI-only approaches. Across both classical and ABMIL settings, models trained solely on MRI embeddings performed poorly, with C-indices often close to 0.4–0.5, indicating that the available MRI-derived features were not sufficiently informative to rival clinical variables.

A likely explanation is that the slice and patient-level MRI features extracted from fixed pretrained CNNs, were not optimally adapted to this specific task and dataset. We did not finetune the backbone feature extractors on a larger, task-relevant cohort, and the limited sample size further constrained the ability of the attention-based MIL architecture to fully exploit subtle imaging cues.