itleMultimodal Prostate Cancer Recurrence Prediction on CHIMERA Task 1 with Streaming S3 Data

Group #FYS-STK3155/4155
University of Oslo, Department of Physics
(Dated: December 16, 2025)

We present a simple, reproducible baseline for the CHIMERA challenge (Task 1: prostate cancer biochemical recurrence prediction) that integrates clinical variables and radiology (mpMRI) features while streaming data directly from the public AWS S3 bucket, avoiding full local storage. We implement a lightweight pipeline that: (i) reads patient-level clinical JSONs and MRI volumes from S3; (ii) produces compact MRI features using intensity histograms; (iii) encodes clinical covariates; and (iv) trains a logistic regression classifier with stratified k-fold cross-validation. On the full public training set we obtain mean AUROC $= 0.679 \pm 0.103$, AUPRC $= 0.579 \pm 0.122$, balanced accuracy $= 0.608 \pm 0.101$, and F1 $= 0.341 \pm 0.249$ (5 folds). The code and configuration are structured for easy extension to histopathology and alternative models (e.g., XGBoost or attention-based MIL). This report documents data handling, methods, results, limitations, and next steps, satisfying the Project 3 requirements for structure, reproducibility, and critical assessment.

I. INTRODUCTION

Predicting biochemical recurrence (BCR) after prostatectomy is clinically important for patient counseling and treatment planning. The CHIMERA challenge [1] offers a benchmark dataset that combines multimodal sources (histopathology, radiology, and clinical data). While many high-performing approaches rely on heavy local preprocessing and storage, clinical teams often need lean pipelines that operate within infrastructure constraints. Our goal is therefore to deliver a compact, streaming-first baseline that can be run end-to-end with minimal local storage, yet remains faithful to sound machine learning practice (clear train–validation splits, fixed seeds, and consistent preprocessing).

Concretely, we address CHIMERA Task 1 (prostate BCR prediction). We fuse clinical variables with compact mpMRI features extracted from MHA/NIfTI volumes. We emphasize: (i) robustness of ingestion (direct S3 reads), (ii) simple, interpretable models (logistic regression), and (iii) clear evaluation (AUROC, AUPRC, balanced accuracy, F1). Though modest, this baseline is a strong foundation for subsequent extensions to histopathology features and multimodal fusion strategies.

II. MATERIALS AND METHODS

A. Dataset and task

We focus on CHIMERA Task 1 (public training split), where the objective is binary classification of biochemical recurrence (BCR). Labels are provided (or can be inferred) from each patient's clinical JSON (key "BCR", values in {0,1}). Radiology data consist of per-patient mpMRI volumes (e.g., T2w, ADC, HBV) provided as MHA and/or NIfTI files. Histopathology slides and precomputed features are also available but were not used in the core baseline to prioritize speed and streaming simplicity.

B. Streaming S3 ingestion

All data are accessed directly from the public bucket `s3://chimera-challenge/v2/task1/` via `s3fs`/`fsspec`. For formats that require a local file path (e.g., MHA reading via SimpleITK), we use a small, bounded temporary cache (few files, auto-evicted), ensuring we never persist the dataset locally. Clinical JSONs are parsed in-memory. This design keeps the storage footprint minimal and makes the pipeline portable.

C. Feature engineering

Clinical. We build a column-wise preprocessing pipeline: numeric features are imputed with the median and standardized; categorical features are imputed with the most frequent value and one-hot encoded. If clinical variables are not in the manifest, we parse the per-patient clinical JSON from S3 and assemble a flat tabular representation.

MRI. For MRI we extract compact intensity features from the primary sequence (e.g., T2w) using histograms with 32 bins (robust range via 1–99% quantiles). This yields a small, fixed-length descriptor per patient that is efficient to stream and train on. SimpleITK handles MHA; nibabel can be used for NIfTI.

D. Models

We use logistic regression as a transparent baseline (`sklearn`), trained on the concatenation of clinical and MRI features. All seeds are fixed. We scale features with a standardizer in a pipeline. While we planned to compare XGBoost and small MLP heads, we report here the logistic regression baseline and leave more complex models to future extensions.

E. Evaluation

We perform stratified 5-fold cross-validation with patient-level splits. We report:

- AUROC (area under the ROC curve),
- AUPRC (area under the precision–recall curve) due to class imbalance,
- balanced accuracy (mean of sensitivity and specificity), and
- F1-score at threshold 0.5.

We summarize per-fold metrics and provide the mean and standard deviation across folds.

III. RESULTS AND DISCUSSION

Table I summarizes 5-fold cross-validation performance (clinical+MRI baseline). Metrics are computed by streaming data directly from S3, demonstrating that a compact pipeline can run without local copies of the dataset. As expected for a simple linear model on compact features, the predictive performance is modest but meaningful.

Interpretation. AUROC of approximately 0.68 suggests the combination of basic clinical covariates and simple MRI histograms contains useful prognostic signal. The AUPRC around 0.58 is reasonable under class imbalance. The linear decision boundary likely underfits the multimodal structure; non-linear models (e.g., gradient boosting) and richer MRI features (texture, radiomics) should improve both discrimination and calibration.

Ablations and scope. We prioritized MRI+clinical due to cost and latency: histopathology feature files are large (gigabytes each). Our code supports WSI features via streaming `.pt`/`.npy` files and can aggregate them (mean/max) to a patient vector; enabling this is a straightforward switch in the configuration but may substantially increase runtime and memory. Future work will add XGBoost, calibration (Platt/isotonic), and early-fusion with WSI features.

Limitations. This baseline: (i) ignores spatial patterns in MRI, (ii) does not yet include histopathology features in the reported results, (iii) uses a single operating threshold (0.5) for F1, and (iv) does not address domain shift. Nonetheless, its simplicity makes it a reliable reference point and a solid foundation for more advanced multimodal methods.

TABLE I. Stratified 5-fold CV performance (clinical+MRI). Mean ± std across folds.

oprule Metric	AUROC	AUPRC	Bal. Acc.	F1
Mean ± std	0.679 ± 0.103	0.579 ± 0.122	0.608 ± 0.101	0.341 ± 0.249

IV. CONCLUSION

We implemented and evaluated a streaming-first baseline for CHIMERA Task 1 using clinical variables and compact MRI features, trained with logistic regression and validated via 5-fold cross-validation. The approach achieves AUROC $\approx$ 0.68 and AUPRC $\approx$ 0.58, with minimal local storage requirements and clear reproducibility. The pipeline is intentionally simple and designed for extension: adding WSI features, non-linear models (XGBoost), and calibration are natural next steps that we expect to improve performance while maintaining reproducibility.

REPRODUCIBILITY AND AVAILABILITY

Code (data loaders, feature extractors, CV runner) is provided under the `Code/` folder. Configuration files document modalities and S3 paths. Experiments can be reproduced by (i) generating a manifest from S3, and (ii) running the cross-validation script with fixed seeds. The environment used was a conda environment ("ml") with `SimpleITK`, `s3fs`, `fsspec`, `scikit-learn`, `pandas`, and `numpy`.

ACKNOWLEDGMENTS

We thank the CHIMERA organizers for providing public access to multimodal data via Grand Challenge and AWS S3.

[1] CHIMERA Challenge: Combining Histology, Medical Imaging, and Molecular Data for Prognosis and Diagnosis. https://chimera.grand-challenge.org/chimera/