Prostate - MRI

The first dataset is relatively small, but well-curated. It includes 26 patients with biopsy-confirmed prostate cancer who also underwent radical prostatectomy. Altogether it contains about 22,000 MRI images, all acquired on a 3-Tesla Philips Achieva scanner using both endorectal and phased-array coils, so the image quality is fairly high. In addition, there’s one pathology image per patient, stored in JPG format, which provides Gleason grading and tumor size information.

DICOM database inside 3D Slicer, where multiparametric MRI sequences like T2, diffusion, and DCE are stored and managed before analysis

the T2-weighted MRI views of the prostate — axial, sagittal, and coronal. These provide the in vivo anatomical detail that we want to align with the ex vivo pathology slices.

One of the strengths here is that every patient has paired MRI and pathology, with standardized Gleason grading and clinical details. The MRI acquisition was high-field and high-resolution, and the pathology was sectioned in the same plane as MRI, which helps for correspondence.

However, the main limitation is that the pathology is only available as low-resolution JPGs, not whole slide images. That means we can’t directly apply modern pathology embedding methods such as Virchow. And although the MRI–pathology alignment exists, there’s no deformable registration provided.

Reference dataset

To test MRI preprocessing and feature extraction pipelines (e.g., loading DICOMs, slice selection, feature extraction with MONAI/ResNet).

To validate MRI–pathology correspondence concepts (e.g., whether MRI signal abnormalities match pathology-reported tumor size and Gleason score).

To practice building a small-scale manifest (linking MRI slice indices with pathology outcome labels).

Prostate-Fused-MRI-Pathology (2023)

The second dataset is more recent and addresses some of those issues. It contains 28 patients with multiparametric MRI, including T1, T2, diffusion, and dynamic contrast-enhanced imaging — in total about 32,500 DICOM images. Importantly, 16 of those patients also have digitized histopathology in the form of whole slide images, with 114 pseudo-whole mounts in TIFF format. Each slide also has an accompanying XML annotation file. For 15 patients, MRI and histology have been deformably co-registered, and there’s a correspondence Excel file that maps histology slices directly to MRI slices.

1. don’t know thickness of slides but the standard MRI slice thickness = ~3–4 mm. Histology section thickness = ~4 μm

2. For WSI analysis (Virchow, embeddings): no thickness needed.

For MRI ↔ histology alignment: thickness is conceptually important, but in this dataset they solved it by giving you correspondence tables. it bypasses exact micrometer thickness and instead maps histology slices (A, B, C…) to MRI slices (#12, #13, #14…)

T2: Best sequence for anatomy (you can see the zonal anatomy: peripheral zone, transition zone, capsule, seminal vesicles). Tumors often appear as dark (hypointense) areas in the normally bright peripheral zone.

Dynamic Contrast-Enhanced MRI (DCE): changes frame-by-frame (every few seconds after injection). Tumors often enhance early and intensely, then wash out. Functional / vascular information (blood supply) Helps distinguish tumor vs benign lesions when diffusion is inconclusive.

A = closest to apex (bottom of prostate).

Next letters = moving upward.

Last letter = near base (top, near bladder)

The limitations here are mainly scale and complexity. There are only 28 patients, and only 16 with WSIs. The pathology images are extremely large — about 76 gigabytes in total — so they require specialized handling with tools like OpenSlide. And while deformable registration is provided, it’s still imperfect, so there may be mismatches between histology and MRI.

Prostate-MRI-US-Biopsy

The third dataset is the largest and most clinically oriented. It includes about 837 participants. All underwent multiparametric MRI with T2, DWI, ADC, and in some cases DCE. The unique feature of this dataset is the linkage to MRI-targeted, ultrasound-guided biopsy cores. For each core, there are pathology results: Gleason scores, the length of cancer within the core, and whether it was malignant or benign. In addition, there are 3D fiducial models that show biopsy needle tracks and prostate surface reconstructions. The dataset is distributed as MRI DICOM images, spreadsheets containing biopsy metadata, and 3D overlays that can be visualized in software such as 3D Slicer.

So how can we use these datasets together? The biopsy dataset is particularly valuable for MRI–pathology correlation, because it allows us to train MRI embeddings against pathology labels such as Gleason score or cancer length. Even though it doesn’t include whole slide images, it still provides weak cross-modal alignment, since pathology outcomes are linked back to MRI coordinates.

The key idea is to combine datasets: the fused MRI–pathology dataset is small, but it gives us direct histology–MRI slice correspondence. The biopsy dataset is much larger, and provides large-scale MRI–pathology supervision. Together, they give us both depth and scale.

My next steps are threefold. First, I will build a single manifest across datasets

Basic Identifiers

patient\_id → anonymized ID (e.g., aaa0051, biopsy123).

dataset\_name → which dataset: Prostate-MRI, Fused, Biopsy.

collection\_site → if available (e.g., NCI Bethesda).

MRI Information

modality → T2, DWI, ADC, DCE, etc.

series\_uid → DICOM series UID (unique key for MRI series).

series\_description → scanner label (e.g., “T2 TSE cor”, “DWI b800”).

slice\_idx → relevant slice index (from correspondence table if mapped).

file\_path\_mri → path to the DICOM folder or converted NIfTI.

mri\_resolution → voxel size (mm).

notes\_mri → quality issues (motion, artifacts, missing slices).

Histopathology Information

wsi\_file → path to WSI (TIFF) or low-res JPG.

wsi\_format → TIFF, SVS, JPG.

annotation\_file → XML polygon file, or “None”.

stain → H&E (default here).

slice\_label → histology label (A, B, C…) if applicable.

file\_size → optional, helps manage large WSIs.

notes\_wsi → e.g., “correction.tif version”, “missing annotation”.

Pathology Labels

gleason\_primary → e.g., 3.

gleason\_secondary → e.g., 4.

gleason\_score → e.g., 3+4=7.

tumor\_length\_mm → cancer length in biopsy core or histology slice.

benign\_malignant → binary label.

Correspondence / Alignment

t2\_slice\_mapped → MRI slice index aligned to histology.

landmarks → notes from correspondence table (e.g., “cystic space”, “calcification”).

registration\_method → deformable, rigid, or “table only”.

alignment\_notes → free text (confidence, maybe slice mismatch).

Second, on the pathology side, I’ll process the whole slide images into embeddings using OpenSlide for patch extraction, the XML annotations for region selection, and Virchow for feature representation.

Third, on the MRI side, I’ll extract slice features using SimpleITK to identify the exact T2 slices indicated in the correspondence tables, and then apply deep models such as MONAI networks, ResNets, or vision transformers to extract embeddings. The ultimate goal is to bring both modalities into a shared feature space for cross-modal learning.