University of Padova

Department of Mathematics "Tullio Levi-Civita"

Master Thesis in Data Science

Developing a Deep Learning Model for Predicting Treatable Molecular Alterations from H&E-Stained Histological Section Slides

Supervisor
Prof. Emanuela Leonardi
University of Padova

Co-supervisor
Prof. Nadieh Khalili
Radboud University Medical Center

Master Candidate
Amirhosein Kalhor

Student ID
2072056

Academic Year
2024-2025

"I suppose therefore that all things I see are illusions; I believe that nothing has ever existed of everything my lying memory tells me. I think I have no senses. I believe that body, shape, extension, motion, location are functions. What is there then that can be taken as true? Perhaps only this one thing, that nothing at all is certain."
— Rene Descartes

Abstract

Determining biomarkers—such as microsatellite instability (MSI), TP53 mutation, and other molecular alterations—is crucial for guiding effective cancer treatment strategies. However, the most reliable way to identify these genomic features currently, requires whole-genome sequencing, a procedure that is costly for widespread clinical use. As a result, treatment decisions often depend mainly on individual pathologists who analyze whole slide images (WSIs) based on cell architecture, staining patterns, and other histopathological cues; and molecular analysis are only used when treatments have not been effective. Deep neural networks have proven highly effective in learning complex data representations for image classification, offering a potential solution to this problem. In this work, we present a deep learning framework for prostate cancer that classifies WSIs according to their likelihood of having specific genomic alterations—such as MSI or TP53 mutation—using histopathological features alone. To accomplish this, we leverage deep foundational models, such as UNI, to extract feature embeddings from pathology pictures and incorporate data from multiple cancer types, thereby enhancing the network's generalization capabilities. By employing this model, clinicians can identify patients at higher risk for alterations without relying on costly whole-genome sequencing for every patient. Our research will stratify patients into low risk and high-risk groups for these biomarkers, allowing for a cost-effective introduction of molecular diagnostics tools with more precision, to become part of the routine clinical procedures.

Contents

Listing of figures

Listing of tables

Listing of acronyms

WSI ..... whole slide image

TMB ..... Tumor Mutation burden

MSI ..... micro satellite instability

H and E ..... hematoxylin and eosin

ROI ..... Regions of Interest

AUC-ROC ..... Area Under the Receiver Operating Characteristic curve

MIL ..... Multi Instance Learning

TP53 ..... tumor protein p53 gene

AUPRC ..... Area Under the Precision-Recall Curve

IHC ..... immunohistochemistry

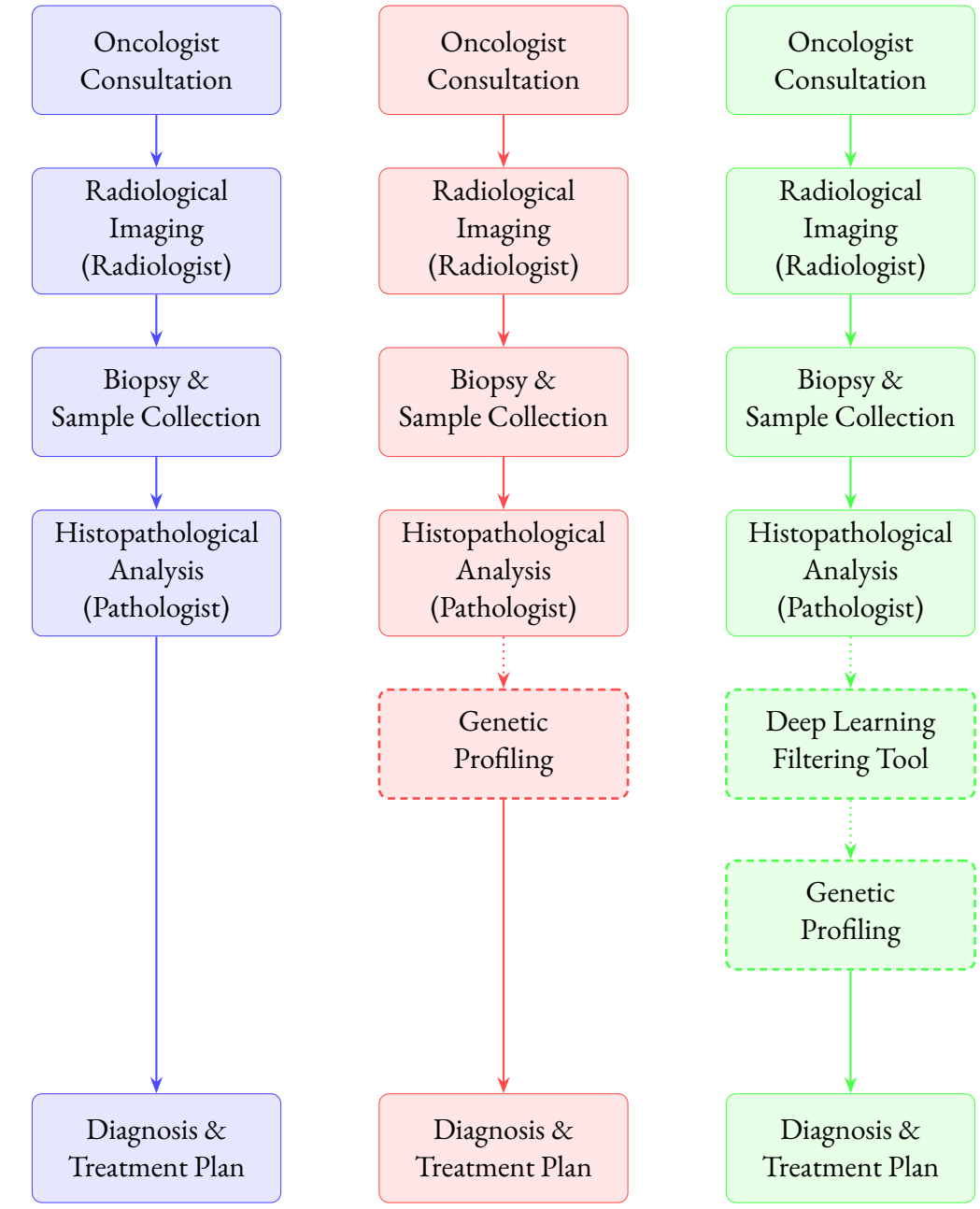
1 Introduction

This chapter examines the clinical and histopathological context of the role and significance that TP53 mutation plays in cancer diagnosis, reviews relevant deep-learning literature, and states the hypotheses that guided our work.

1.A1 General Workflow in Cancer Diagnosis and clinical motivation

Cancer diagnosis and treatment are multi-step procedures that involve collaboration among radiologists, pathologists, and oncologists to ensure accurate diagnosis and optimal patient treatment. Sometimes, a tissue specimen is obtained via biopsy and processed into histopathology slides through fixation, embedding, sectioning, and staining. After scanning or mounting, a pathologist examines the slide for morphological and immunohistochemical markers—e.g. assigning a score using scoring systems such as Gleason in case of prostate cancer[9] to classify tumor type and grade. Finally, in collaboration with the oncology team, a treatment plan is devised and its efficacy is periodically reassessed with follow-up imaging or repeat biopsy.

This process although effective, contains serious drawbacks. It is known that certain genetic alterations such as TP53 mutation, as will be explained later, can be a cause of and therefore a biomarker for more aggressive and more lethal forms of the disease. They can also be helpful in determining the course of treatment since they will respond better to certain drugs and other drugs will not be as effective in treating those cancers. Currently the only reliable path to determining these markers is whole genome sequencing. This approach is too costly to be used for every patient, so it is not a regular part of the diagnosis. In this work we aim to utilize the power of AI and implement a deep learning pipeline to stratify patients ,based on the histopathology slides of their tissue, into high risk and low risk buckets for TP53, focusing the whole genome sequencing on a subsection of patients, and potentially making it feasible to be incorporated into the diagnosis procedure.



(a) Current Basic Approach **(b)** Hypothetical Costly Approach **(c)** Cost-Effective Approach

Figure 1.1: Flow Chart for Cancer Diagnosis. it is worth noting that this is a very generalized chart. as mentioned before some cancers such as colorectal cancer have had some molecular testing incorporated in the procedure. This mostly reflects the procedure for prostate Approaches. "Dotted boxes are hypothetical steps that are currently not implemented in the procedure."

1.A2 Knowledge Gap, Hypotheses and Specific Objective

As of this writing we don't have a tool that can predict TP53 mutation based on histopathology slides with adequate precision, an area where the only trusted approach remains the high cost whole genome sequencing. multiple measures have been taken to train a deep learning model for this aim ,explored further in the review section of this chapter, however, their success remains limited. Here we are attempting to change that

We aim to rectify this By utilizing a few new techniques and improve the state of the art (see 3 section).

Our attempt is guided by the hypothesis that Certain latent features in histopathological slides encode molecular alterations that may elude routine microscopy and that also at least some of these features are conserved across cancer types—more likely for a ubiquitous mutations such as *TP53*, therefore An AI model trained on multiple cancer types simultaneously, will better stratify patients into **low risk** (low likelihood of actionable alteration) and **high risk** groups, reserving DNA/RNA sequencing for the latter and making precision oncology more affordable. This allows for an affordable approach to incorporate whole genome sequencing in the diagnosis procedure by reserving it only for a subset of cancer patients

1.A3 Background

1.A3.1 Histopathology Slides and Whole Slide Images (WSIs)

Histopathology slides are glass preparations of biopsy tissue processed by fixation, paraffin embedding, microtomy, and staining, enabling visualization of cellular and architectural features essential for diagnosis.

Whole-slide images (WSIs) are high-resolution scans of histopathology slides, capturing the entire tissue section at multiple magnifications and enabling interactive pan-and-zoom navigation on-screen. Typical WSI resolutions range from 0.25 to 0.5 µm per pixel, producing file sizes that can exceed several gigabytes (GB). Such large images pose challenges for GPU-based deep-learning pipelines, which we address in Section 3. WSIs thus form the primary data modality in computational pathology and in the experiments of this thesis.

1.A3.2 Prostate Cancer

Prostate cancer (PCa) is the second most-diagnosed cancer in men worldwide, accounting for approximately 15 % of new male cancer cases[3]. PCa is clinically heterogeneous: many tumors remain indolent, whereas others progress to aggressive, metastatic disease with high mortality. Key molecular drivers of progression include alterations in tumor-suppressor genes such as TP53. To explain the histological correlates of aggressive behavior, we first review the normal prostate microarchitecture.

The prostate is an exocrine gland located below the bladder, which encircles the urethra. Histologically, it comprises glandular structures lined by tall columnar or pseudostratified epithelium, underpinned by a basal cell layer and embedded within fibromuscular stroma . Prostate adenocarcinoma, the most prevalent malignancy of

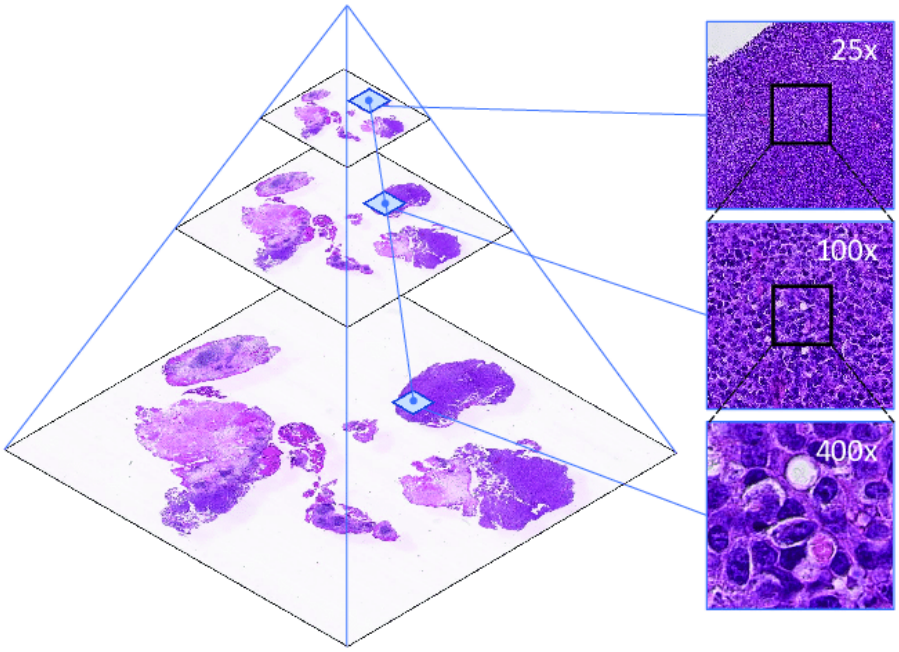


Figure 1.2: The structure of a WSI. slides are scanned at several magnification levels and saved in files with special formats[20]

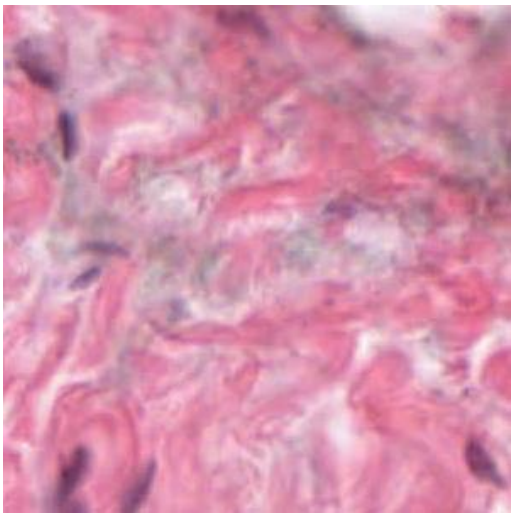
this gland, originates from the secretory epithelium and is characterized by the absence of basal cells, an infiltrative growth pattern, and cytological features such as prominent nucleoli. Diagnosis is often established through histopathological evaluation of needle biopsy specimens, with the Gleason grading system employed to assess tumor differentiation and predict prognosis.

1.A3.3 TP53 and Genome Sequencing

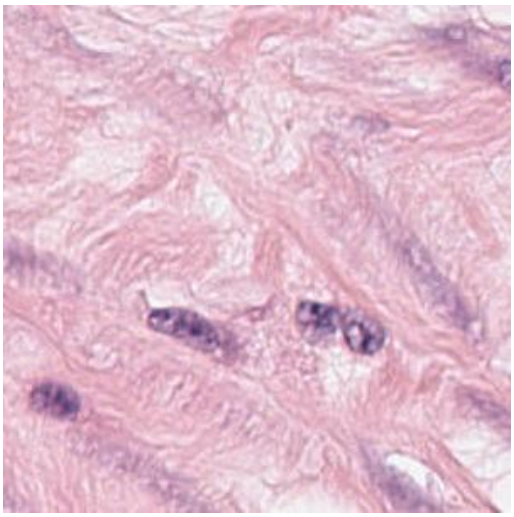
TP53 is the most frequently mutated gene across human malignancies, with somatic alterations detected in approximately 50–60 % of all tumors. In prostate adenocarcinoma, TP53 mutations occur in about 6–7 % of primary cases and rise to nearly 40 % in advanced disease[8]. These alterations span ovarian, colorectal, lung, head-and-neck, and many other cancers, making TP53 inactivation the single most common tumor-suppressor event in oncology[2].

The TP53 gene encodes the p53 protein, a master regulator of the cellular stress response: it enforces cell-cycle arrest, triggers apoptosis, and facilitates DNA repair[16]. In prostate cancer (PCa), TP53 mutations—especially in advanced or metastatic cases—ablate these tumor-suppressive activities, leading to increased genomic instability and resistance to standard therapies.

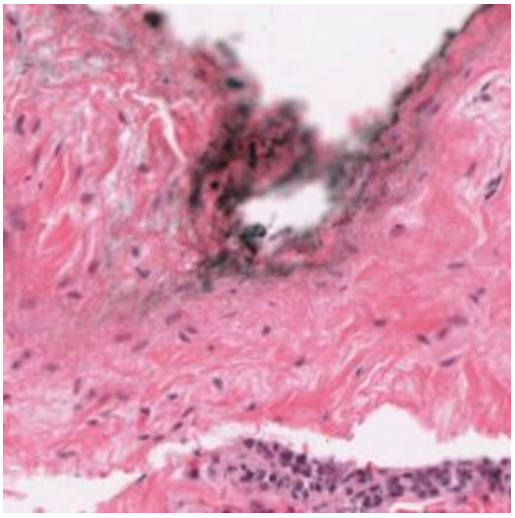
Clinically, PCa tumors harboring TP53 mutations tend to have higher Gleason scores, more frequent lymphovascular invasion, and worse overall survival compared to TP53-wildtype tumors[8]. Moreover, co-occurrence of TP53 alterations with other tumor suppressors (e.g. RB1, PTEN) defines a molecularly aggressive subtype with rapid progression and poor prognosis.



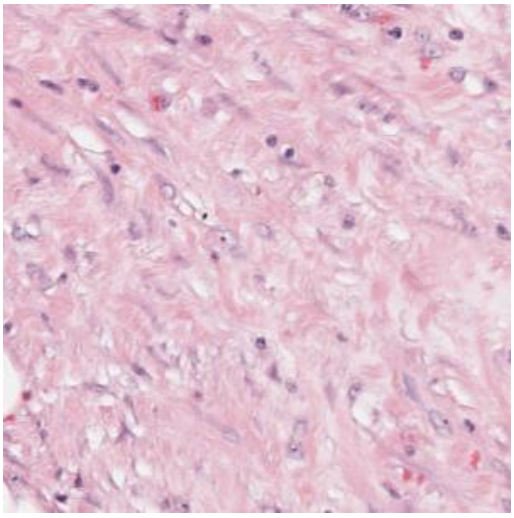
(a) Pathology Image 1.



(b) Pathology Image 2.



(c) Image 1 at a different magnification.



(d) Image 2 at a different magnification.

Figure 1.3: Examples of different magnification of WSIs, images in each column is from the same WSI.

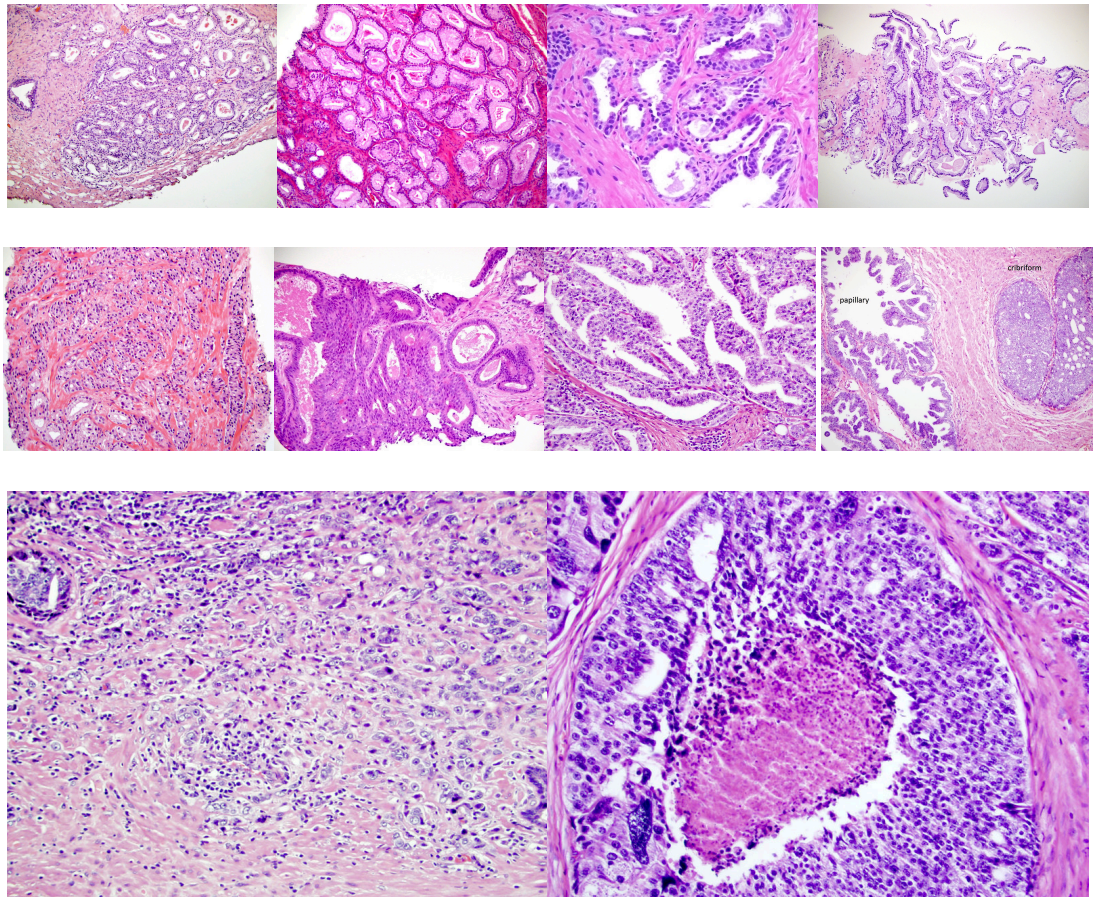


Figure 1.4: Representative pathology images demonstrating Gleason scores[12]. Row 1 (4 images): Gleason score 3; Row 2 (4 images): Gleason score 4; Row 3 (2 images): Gleason score 5. The Gleason score is a grading system for prostate cancer based on histological patterns. It is determined by summing the two most prevalent patterns (each scored 1 to 5). This score predicts cancer aggressiveness and guides treatment decisions.

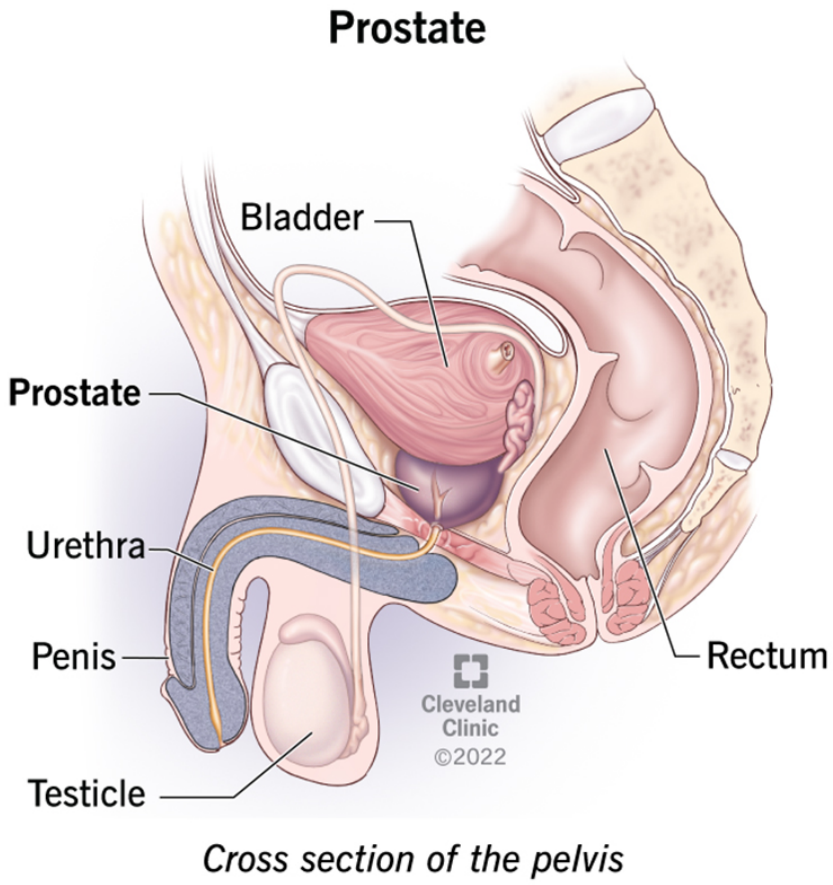


Figure 1.5: schematic image of prostate positioning within the body[1]

Approximately 80–90 % of TP53 alterations are missense substitutions in its DNA-binding core domain. These mutants not only lose transactivation of genes controlling cell-cycle arrest, DNA repair, and apoptosis, but also frequently inhibit residual wild-type p53 in a dominant-negative manner. In addition, certain p53 mutants gain novel oncogenic functions—driving proliferation, invasion, metabolic shifts, and immune evasion—that further fuel tumor progression[6].

Consequently, the presence of TP53 mutations not only serves as a valuable prognostic biomarker but also helps guide treatment decisions, as tumors harboring these mutations often exhibit resistance to conventional therapies.

Mutant p53 proteins often exhibit extended half-lives, resulting in nuclear accumulation that can be visualized as intense, diffuse staining on p53 immunohistochemistry (IHC), whereas truncating or null variants show absent ("null") staining[17]. This differential IHC pattern indicates that even standard histology images may contain latent visual cues of TP53 status that a sufficiently powerful deep-learning model could exploit.

These molecular insights have significant therapeutic implications. For instance, TP53 mutations have been linked to resistance to androgen deprivation therapy, a cornerstone of PCa treatment. Emerging strategies aim to

restore p53 function or exploit vulnerabilities in TP53-deficient tumors, offering potential avenues for targeted therapies. As our understanding of TP53's role in PCa deepens, it holds promise for refining prognostic models and guiding personalized treatment approaches.

Due to the costly nature of determining this crucial marker, treatment is often prescribed without this information. However due to the nature of the mutation's effects, we hypothesize that there may be visible changes in the cell or tissue structure particularly in the nucleus that can be observed in the pathology image of the tissue, paving the way for powerful imaging models to capture this pattern, potentially offering an economically viable path of incorporating this information into the diagnostics procedure.

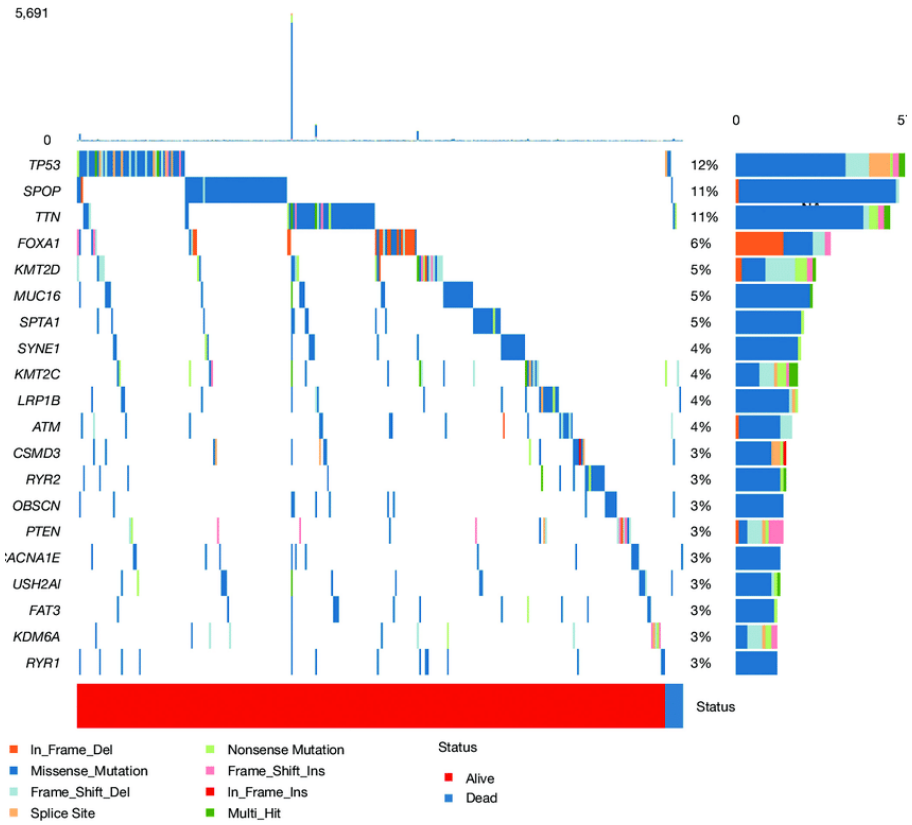


Figure 1.6: mutation map of genes present in prostate cancer[10] this combines the primary and advance cases

1.A4 Literature Review

Deep-learning work over the last five years shows that the histomorphology of routine H&E whole-slide images (WSI) carries signals that allow in-silico detection of these alterations—often before they are obvious to a pathologist.

Evidence in prostate cancer. The current benchmark for WSI-based *TP53* prediction in prostate adenocarcinoma is the tile-level ResNet-18 model of Pizurica *et al.* ("TiDo"). Trained on TCGA-PRAD patients, TiDo achieved a mean **AUC 0.71** on held-out TCGA data, with similar performance during cross-validation *in-house*. External testing on a multifocal Ghent cohort confirmed generalisability at both patient and lesion level (primary-tumour biopsies and prostatectomies AUC 0.65, lymph-node metastases AUC 0.68).

Interpretability analyses revealed that the network weighs stromal patterns—collagen remodelling, cancer-associated fibroblasts—at least as much as epithelial morphology, explaining many apparent false-positive calls[19]

Evidence in other cancers and cross-cancer transfer. A pan-cancer Nature Communications study[18] trained fully supervised CNNs for *TP53* status in nineteen TCGA cohorts, reporting slide-level **AUCs 0.65–0.80** within the tissue of origin and when networks were transferred across tissues, as in breast to lung classification, performance held at 0.62–0.72 .These positive cross-classification results suggest that a subset of *TP53*-associated patterns is shared across organs. More recently, Arslan *et al.* performed 12093 molecular prediction experiments spanning 32 cancer types. Single-nucleotide variants in driver genes were moderately predictable (median **AUC 0.636±0.117**); p53 protein over-expression was one of the most robust signals, detectable in six cancers with **AUCs 0.672–0.835**. Importantly, the authors observed little dependence on tumour purity, reinforcing the notion that both tumour cells and micro-environment contribute to the phenotype. Additionally, it was suggested that some cancers, including breast and bladder, have an easier to detect patterns which can contribute to better cross model prediction. This observation informed our training set later and was put into test.

1.A5 Thesis Outline

First and foremost we will explore the dataset, the reason for its selection and its prepration. Then we will establish several parallel pipelines for preprocessing our slides and their classification as outlined in the table of content and the 3 and 4 chapters. We will follow this by several new additional modifications and changes that we have implemented in our target and loss functions and examine their effectiveness in improving the results followed by a discussion of their shortcomings. we will finish by summerising the overall work in 5 chapter and what possible routes exist for potential improvement and a more Robust modeling in future work

2 Dataset

2.A1 Data Sources and Rationale

For model development, validation, and testing, we utilized publicly available data from The Cancer Genome Atlas (TCGA) [21]. TCGA provides comprehensive multi-modal datasets—including genomic, clinical, and imaging data—across numerous cancer types via the Genomic Data Commons (GDC) and The Cancer Imaging Archive (TCIA). TCGA is the standard dataset used in many similar works allowing for better comparison. Leveraging such an extensive and diverse dataset allows the assessment of model generalizability and effectiveness for predicting TP53 mutation status from whole-slide histopathological images (WSIs).

Specifically, we retrieved hematoxylin and eosin (H&E)-stained WSIs from the following TCGA cancer cohorts:

- Prostate Adenocarcinoma (PRAD-TCGA)
- Colorectal Adenocarcinoma (COAD-TCGA)
- Breast Invasive Carcinoma (BRCA-TCGA)
- Lung Adenocarcinoma (LUAD-TCGA)
- Bladder Urothelial Carcinoma (BLCA-TCGA)

The inclusion of diverse cancer cohorts was partly motivated by findings from Noorbakhsh *et al.* [18], suggesting that broader morphological diversity enhances predictive performance in histopathology models. however as previously mentioned they used a cross cancer training and fell short of pan cancer approach.

2.A2 Inclusion Criteria and Filtering

We applied rigorous quality-control measures to ensure data integrity. WSIs were excluded if they suffered from Missing or inconsistent pixel-spacing metadata, critical for accurate WSI interpretation. In addition the absence of annotated TP53 mutation status was another cuase for removal, as the primary predictive task necessitates labeled data. These measures ensured high-quality, reliable input data for model training and evaluation.

2.A3 Class Distribution and Imbalance

Our dataset exhibited substantial variability in TP53 mutation prevalence across cancer types (Table 2.1). Notably, prostate adenocarcinoma showed marked class imbalance, with only 45 positive cases among 397 total cases. This imbalance poses significant modeling challenges, particularly for prostate-specific models.

Cancer type	Total	TP53 +	TP53 −
Breast Invasive Carcinoma	971	631	340
Bladder Urothelial Carcinoma	383	188	195
Prostate Adenocarcinoma	397	45	352
Lung Adenocarcinoma	1,064	526	538
Colorectal Adenocarcinoma	461	216	245
Total	3,276	1,606	1,670

Table 2.1: Number of patients by cancer type and TP53 mutation status. breast invasive carcinoma and lung adenocarcinoma compose the largest cohorts spanning over sixty percent of the total

Figure 2.1 highlights this imbalance, underscoring the challenge posed by the prostate cohort compared to other cancers. Despite this imbalance, recent studies [15, 4] have shown that standard imbalance-correction methods often degrade model performance in histopathology tasks; thus, we refrained from applying resampling or class-weighting in our high performing models.

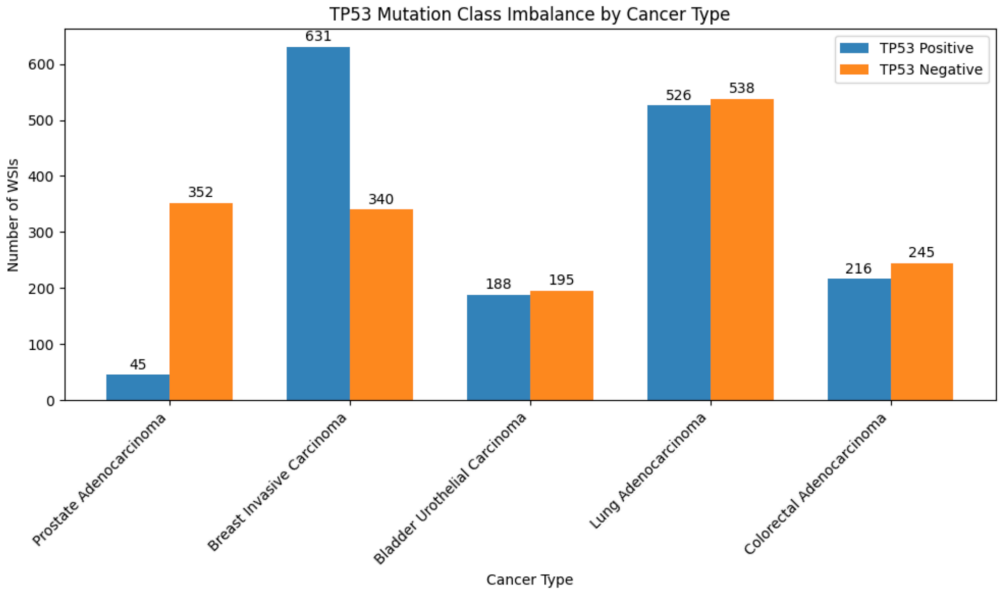


Figure 2.1: Visualization of the class imbalance between TP53-positive and TP53-negative samples across cancer types, particularly pronounced in prostate adenocarcinoma (PRAD-TCGA). Such imbalance complicates effective model training and evaluation and is traditionally assumed that to require mitigating measures.

2.A4 Data Splits for Training and Validation

For rigorous evaluation, the prostate adenocarcinoma cohort (PRAD-TCGA) was partitioned into three stratified cross-validation folds based on TP53 status. Within each fold, 20% of the training data was reserved as a validation set for hyperparameter tuning, maintaining stratification.

We defined three training scenarios to examine how dataset diversity influences predictive performance:

1. **Prostate-only:** PRAD-TCGA exclusively, serving as a baseline.
2. **Three-cancer:** PRAD-TCGA combined with BLCA-TCGA and BRCA-TCGA.
3. **Five-cancer:** Three-cancer set expanded with LUAD-TCGA and COAD-TCGA.

Table 2.2 details the TP53 status distribution across these dataset splits.

These splits allowed us to quantify the benefits of adding diverse cancers and identify whether improvements plateau beyond a certain morphological diversity threshold, as proposed by Noorbakhsh *et al.* [18].

Dataset split	Total	TP53 +	TP53 −
Prostate-only Training	222	24	198
Three-cancer Training	1,576	843	732
Five-cancer Training	3,101	1,585	1,515
Prostate-only Validation	53	6	47
Prostate-only Test	122	15	108

Table 2.2: Distribution of TP53-positive and negative samples across training, validation, and test splits. lack of adequate TP53 positive data in prostate cancer is visible in the training and test sets since they remain pure prostate data through out the training.

2.A5 Ethical Compliance and Data Usage

This study used publicly available, fully anonymized WSIs and genomic data from TCGA. Under TCGA Human-Subjects Data Policies, this constitutes "non-human-subjects research," exempting additional Institutional Review Board (IRB) oversight and informed consent. Analyses adhered to the Declaration of Helsinki, and reporting followed the TRIPOD-AI checklist for transparency in AI-based predictive modeling.

3 Models

3.A1 Data Preprocessing

3.A1.1 Overview

Whole-slide images (WSIs) range from tens of megabytes to several gigabytes, far exceeding GPU memory limits. We therefore tile each WSI into non-overlapping patches, then discard background via simple thresholding. We devise two pipelines designed to extract patches and create embeddings from the aformentioned patches producing a final tensor that is used for training. Those piplines are:

- **CLAM–UNI:** CLAM creates patches, discarding the non tissue parts of the slide, then UNI encodes each into a 1024-D vector. the final output is a matrix with a width of 1024, representing the vector sizes and a length that is determined by number of patches extracted from our WSI.
- **Virchow2–PRISM:** Virchow2 encodes up to 100 k mixed-magnification tiles into 2560-D vectors; PRISM aggregates them into a single 1280-D slide token.

3.A1.2 Details of the Embedding Pipelines

CLAM–UNI Pipeline

CLAM region proposal. We use the CLAM toolbox with default settings[14]. WSIs are segmented and tiled into 256×256 patches. CLAM records the coordinates, instead of saving a myraid of small images, increasing efficiancy in memory usage and processing.

UNI encoding. Selected patches (256 × 256) feed into the UNI model (ViT-L/16, 307 M parameters) pre-trained with DINO-v2 on 100 k H&E WSIs[5]. UNI's embedding dimension is 1024, and yields stable features across magnifications. CLAM's preprocessing code aligns coordinates, requiring no further registration.

Summary. With defaults, CLAM–UNI outputs a $[N_{\text{NumPatches}} \times 1024]$ feature tensor per slide.

Virchow2–PRISM Pipeline

Virchow2 embeddings. We employ Virchow2, a pathology-tailored Vision Transformer (ViT-H/14) with 632 M parameters. It processes 224 × 224 px tiles extracted at 5×, 10×, 20×, or 40× magnification, feeding each tile through 32 transformer layers with 16 attention heads and a 1280-dimensional token embedding space. To form a single descriptor for each tile, Virchow2 concatenates the CLS token (1280 D) with the mean of all patch tokens (1280 D), yielding a rich 2560-dimensional embedding that captures both global context and fine-grained tissue morphology[23]

PRISM aggregation. PRISM is a multi-modal foundation model that learns to summarize an entire specimen (one or more WSIs) into a single slide-level representation. PRISM's Perceiver-IO slide encoder uses 513 learned latent vectors (each 1280 D) and eight cross-attention blocks. The first block attends directly to all tile embeddings and the second block begins to share keys/values via a cache. Blocks 3–8 reuse the same cross-attention weights, streaming in cached context to save memory. Each block is followed by a six-layer latent self-attention transformer (with 8 heads), and throughout, all latent tokens are kept at 1 280 dimensions. After eight such stages, PRISM outputs One "slide embedding" ($[1 \times 1280]$) – a global summary used for classification [7]

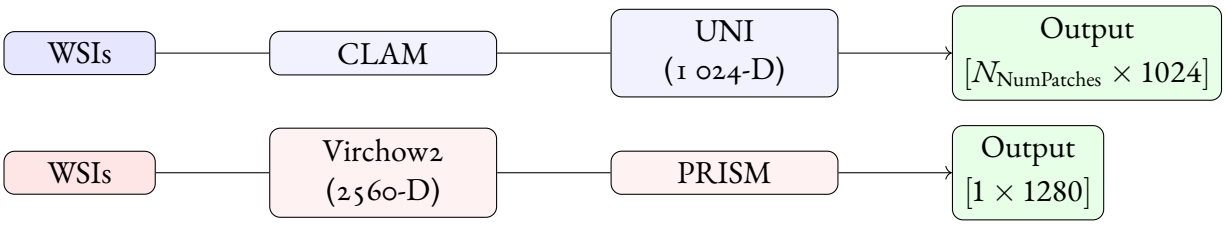


Figure 3.1: Embedding pipelines with default settings. Top: CLAM–UNI produces $[N_{\text{NumPatches}} \times 1024]$ ROI embeddings. Bottom: Virchow2–PRISM yields $[1 \times 1280]$ slide representations.

3.A2 Compound Training Objective

This is a new compound loss that we used in some of our models. The idea was to use contrastive loss as part of the total loss in order to force the model to close the gap between representations of TP53 positive cases among the cancer types. We combined three components into our slide-level loss: (i) a primary binary cross-entropy (BCE) for TP53, (ii) an auxiliary cross-entropy over cancer type, and (iii) supervised contrastive loss.

Supervised contrastive loss. Following Khosla *et al.*[13], let $z_i \in \mathbb{R}^d$ be the L_2-normalized embedding of slide i. Define the positive set $\mathcal{P}(i) = \{j : y_j = y_i\}$. Then for each anchor i,

$$\mathcal{L}_{\text{con}}(i) = -\frac{1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp\big(z_i \cdot z_p/\tau\big)}{\sum_{a \neq i} \exp(z_i \cdot z_a/\tau)},$$

and $\mathcal{L}_{\text{con}} = \frac{1}{B} \sum_i \mathcal{L}_{\text{con}}(i)$.

Auxiliary head. Let ℓ_{prim} be the logit for TP53 and $\boldsymbol{\ell}_{\text{aux}} \in \mathbb{R}^C$ the logits over C cancer types. We define

$$\mathcal{L}_{\text{prim}} = -\big[y_{\text{TP53}} \log \sigma(\ell_{\text{prim}}) + (1 - y_{\text{TP53}}) \log\big(1 - \sigma(\ell_{\text{prim}})\big)\big],$$

$$\mathcal{L}_{\text{aux}} = -\sum_{c=1}^{C} 1[y_c = c] \ \log\big(\text{softmax}(\boldsymbol{\ell}_{\text{aux}})_c\big).$$

Total loss. The final objective is

$$\mathcal{L} = \mathcal{L}_{\text{prim}} + \lambda_{\text{aux}} \mathcal{L}_{\text{aux}} + \lambda_{\text{con}} \mathcal{L}_{\text{con}}, \quad \lambda_{\text{aux}}, \lambda_{\text{con}} \in [0.1, 1].$$

Gradients from all three terms backpropagate into the shared encoder. During inference, only the primary head is retained.

3.A3 Model Architectures

3.A3.1 Modified Attention-based Deep MIL (mAbMIL)

Our primary slide-level model is a variant of Attention-based Deep Multiple Instance Learning (AbMIL)[11]. We briefly recap the original architecture and then describe our modifications:

1. **Instance encoder:** a CNN (or linear projection) maps each patch embedding $x_i \in \mathbb{R}^{d_{\text{in}}}$ to a hidden vector $h_i \in \mathbb{R}^H$.
2. **Attention pooling:** either vanilla or gated attention computes weights a_i over instances:

$$A_{\text{vanilla}}: \quad Z = \tanh(W_1 H), \ A = \text{softmax}(W_2 Z),$$

$$A_{\text{gated}}: \quad U = \sigma(W_u H), \ V = \tanh(W_v H), \ A = \text{softmax}\big(W_w (U \odot V)\big).$$

3. **Bag classifier:** the slide-level embedding $z = \sum_i a_i h_i$ passes to one or more heads (primary TP53, optional auxiliary).

We project h_i through an additional linear layer before attention to control model capacity. In addition, We included an auxiliary cancer-type head in some of our models as a tuned hyperparameter so model can get a better

representation of cancer types, potentially leading to better generalization and as explained before used a compound loss with some of our models instead of the simple loss originally proposed. On top of that We adjustd attention hidden size L and number of heads to better reflect the complexity of our data and optimize AUROC on validation folds,

3.A3.2 MLP with Attention Pooling and Two-Stage Training(MAPT)

We also implement a lightweight MIL variant combining multi-head attention pooling with an MLP bag classifier, trained in two stages:

Architecture.

- **Instance encoder:** a single linear layer (1024→1024) + ReLU.
- **Attention pooling:** H learned queries attend over all N instances, producing a bag embedding of size H.
- **MLP classifier:** three hidden layers (1024→512→128→16→1) with batch-norm, dropout, and ReLU; final sigmoid logit predicts TP53.

Two-stage training.

1. **Stage 1:** all layers trained on the multi-cancer dataset for 10 epochs.
2. **Stage 2:** freeze the first two MLP layers, fine-tune the final two layers on the prostate-only training split for 5 epochs.

This strategy leverages general histology features learned from diverse tumors, then refines prostate-specific decision boundaries.

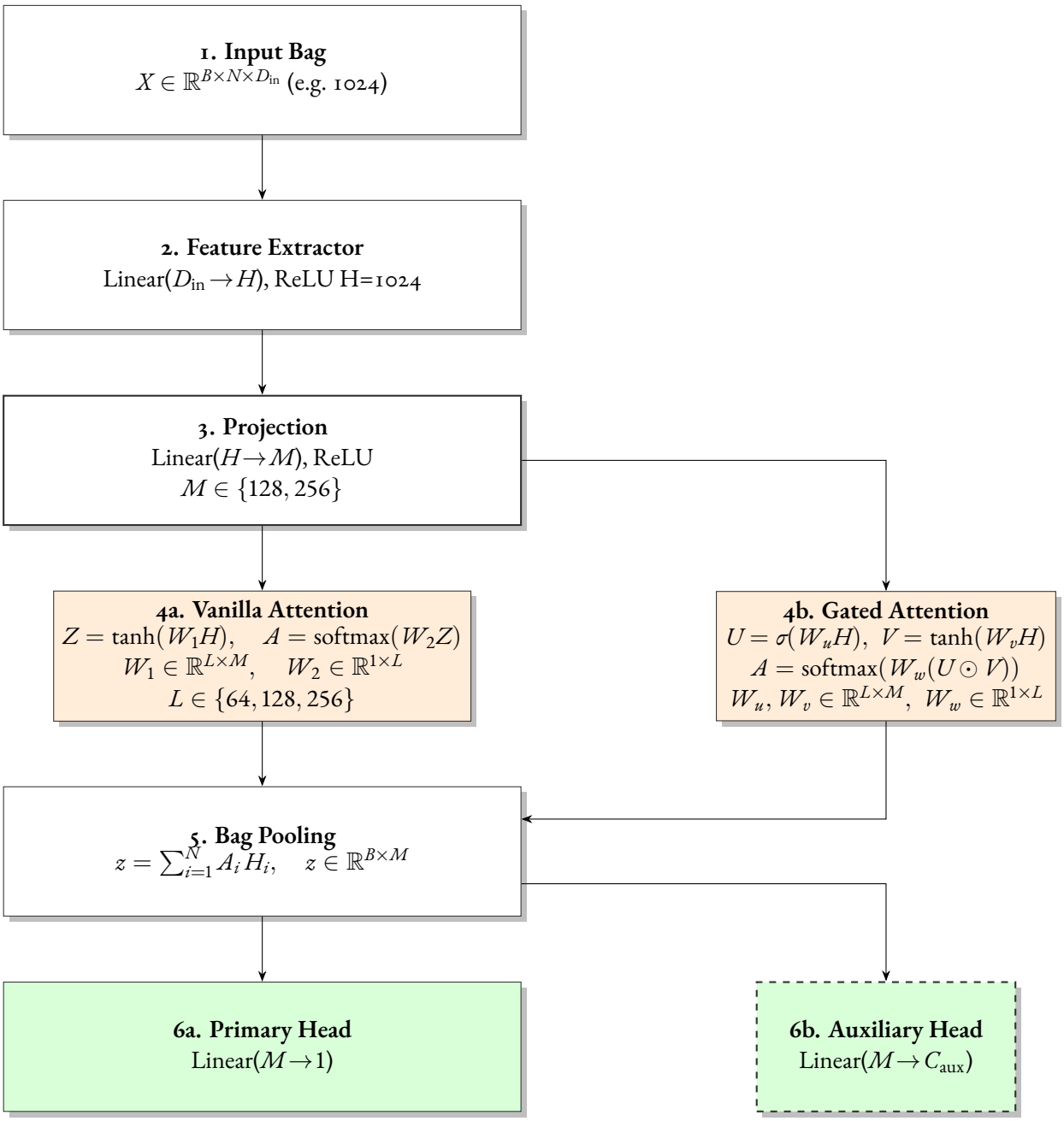


Figure 3.2: Architecture of the modified mAbMIL model. Dashed components are optional (controlled by hyperparameters)

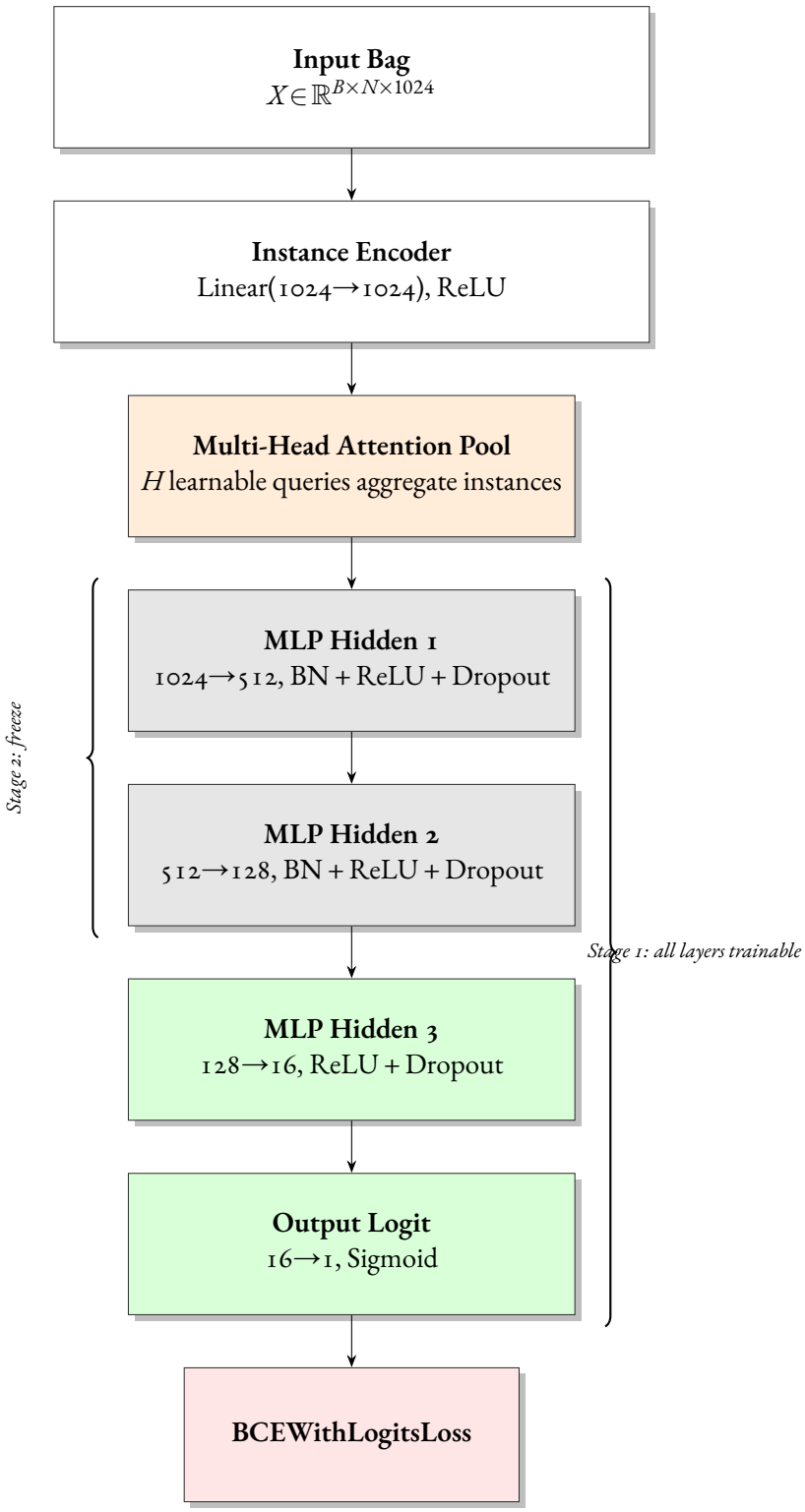


Figure 3.3: Two-stage training for the attention-pooling MLP model. Gray layers are frozen in Stage 2; green layers are fine-tuned on prostate data only.

4 Results

4.A1 Overview of Experiments

We evaluated each model on the held-out test set (122 whole-slide images). Training was conducted under three data regimes:

- **Prostate-only:** TCGA–PRAD slides only.
- **Three-Cancer:** PRAD + BLCA + BRCA cohorts.
- **Five-Cancer:** Three-Cancer + LUAD + COAD cohorts.

Each regime was run with and without our compound loss (auxiliary head + supervised contrastive). Hyperparameters were tuned using the Baysian Hyperband algorithm[22]. Our approach in training and tuning was comprehensive and thorough. We assessed model discrimination using the area under the receiver-operating-characteristic curve (AUC-ROC). The ROC curve plots true positive rate (TPR) against false positive rate (FPR) over all thresholds. An AUC of 1.0 indicates perfect discrimination, whereas 0.5 corresponds to random performance. During training, we monitored the validation loss (primary BCE) for early stopping, and tracked the validation AUC-ROC as a complementary metric to ensure robust generalization. To gauge reliability we computed 95% confidence intervals (2000 resamples) on the test set.

4.A1.1 Main pipeline : mAbMIL + CLAM-UNI, Baseline Performance (No Auxiliary, No Contrastive)

Table 4.2 summarises the discrimination achieved by our CLAM–UNI embeddings coupled to the modified *modified multiple-attention bag MIL* (mAbMIL) aggregator using only the standard BCE loss. we removed the small

Parameter	Description
Optimizer	Adam with a learning rate of 0.0005, betas (0.9, 0.999), and weight decay 10^{-4}.
Epochs	The model is trained for a predetermined number of epochs (default: 10) with early stopping based on validation loss.
$\lambda_{\text{aux}}, \lambda_{\text{con}}$	Both set to 0.1.

Table 4.1: The table listing the best hyperparameters used for training

CNN network in AbMIL, meant to extract information from raw images, and replaced it's output by UNI embeddings instead. we used CLAM for patching and coordinate extraction in creating the attention heatmaps. In Contrastive-AUX cases we replaced the BCE loss of AbMIL with a compound loss mentioned in chapter 3

Dataset	**AUC**	Std. Dev.	**95% CI**
Prostate-only	0.713	0.026	[0.481, 0.903]
Three-Cancer	0.780	0.026	[0.701, 0.851]
Five-Cancer	0.791	0.009	[0.721, 0.859]

Table 4.2: Baseline AUC-ROC for mAbMIL + CLAM–UNI (BCE only). Adding BLCA and BRCA data increases AUC and narrows the CI width, indicating both higher accuracy and greater stability. Extending to five cancers yields only a marginal additional gain, with an almost identical CI width; the improvement is not statistically significant.

A few key observations can be made here. First is that adding BLCA and BRCA raises the AUC from 0.713 to 0.780 and causes the 95 % CI width to shrink (0.422 → 0.150), confirming greater consistency. Extending to Five-Cancer offers only a marginal point-estimate increase (0.780 → 0.791) and a comparable CI width (0.138), rendering the effort statistically insignificant. This indicates the performance increase gained by a pan cancer approach, presenting strong indication that not only there are possible TP53 morphological features visible in WSIs but it is also strong evidence that there are TP 53 specific morphologies preserved across different tissue and cancer types

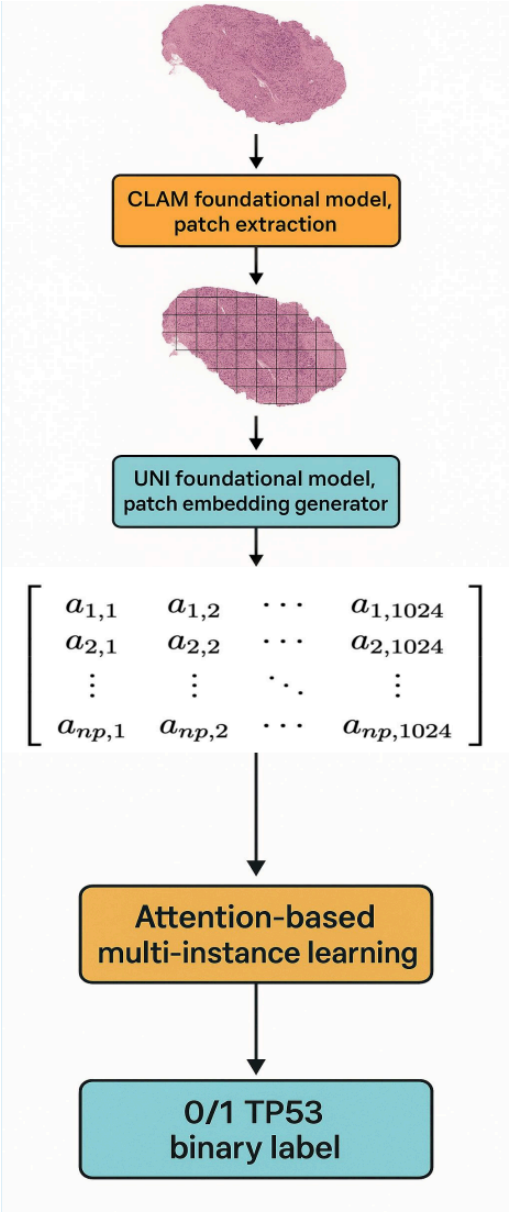
4.A1.2 mAbMIL + CLAM-UNI + Auxiliary Head and Contrastive Loss

We next added the auxiliary cancer-type head and supervised contrastive loss ($0.1 < \lambda_{\text{aux}} = \lambda_{\text{con}} < 1$) to add more regularization and improve generalization of the model. contrastive loss forces the positive TP53 classes closer together, encouraging the model to treat the cohort as more homogeneous and reducing cancer type effects. auxiliary head allows the model to learn cancer specific features and allows the model to better discriminate between cancers, reducing the chance that cancer specific features are effecting the output.
With Three-Cancer data the auxiliary + contrastive objective pushes the AUC upward (0.780 → 0.800), but the 95% CIs overlap heavily ([0.701, 0.851] vs [0.720, 0.875]); the improvement is therefore not statistically significant. In the Five-cancer setting the compound loss slightly depresses the point estimate, again within overlapping

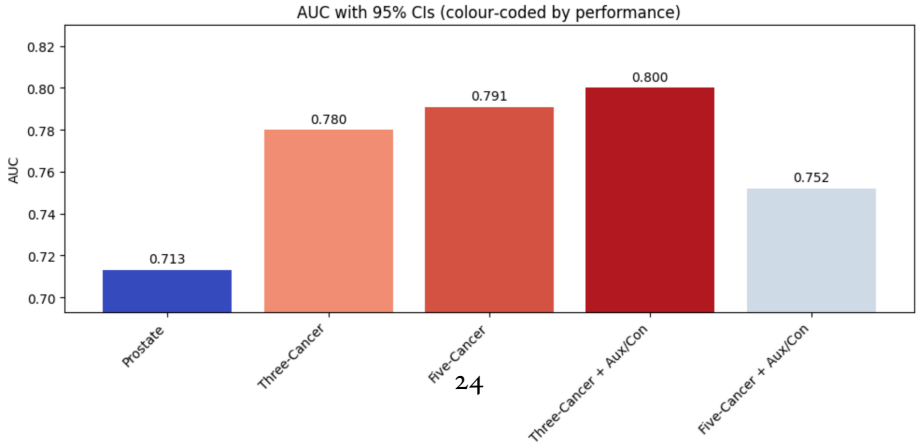
Table 4.3: mAbMIL + CLAM–UNI with auxiliary and contrastive loss. changes in confidence interval and AUC are not statistically significant compared to baseline

Dataset	**AUC**	Std. Dev.	**95% CI**
Three-Cancer + Aux/Con	0.800	0.013	[0.720, 0.875]
Five-Cancer + Aux/Con	0.752	0.018	[0.680, 0.823]

intervals. We conclude that the auxiliary/contrastive head mainly tightens variance (fold-level SD drops from 0.026 to 0.013) rather than shifting the mean, creating a more robust model with no observable increase in performance



(a) Schematic of the CLAM–UNI + mAbMIL pipeline.



(b) AUC with 95 % confidence intervals – gives an at-a-glance sense of accuracy and reliability. A clear increase in the pan cancer setting compared to prostate alone

4.A2 Visualization of Attention Maps

Before examining the heat-maps we note that higher AUC does *not* guarantee anatomically meaningful focus. Figure 4.2 therefore juxtaposes the saliency of the prostate-only model with that of the Three-Cancer model on identical cases. Visual inspection supports the quantitative findings: cross-cancer training produces sharper, tumour-centric attention, reinforcing the claim that morphological breadth aids feature localisation as well as classification accuracy.

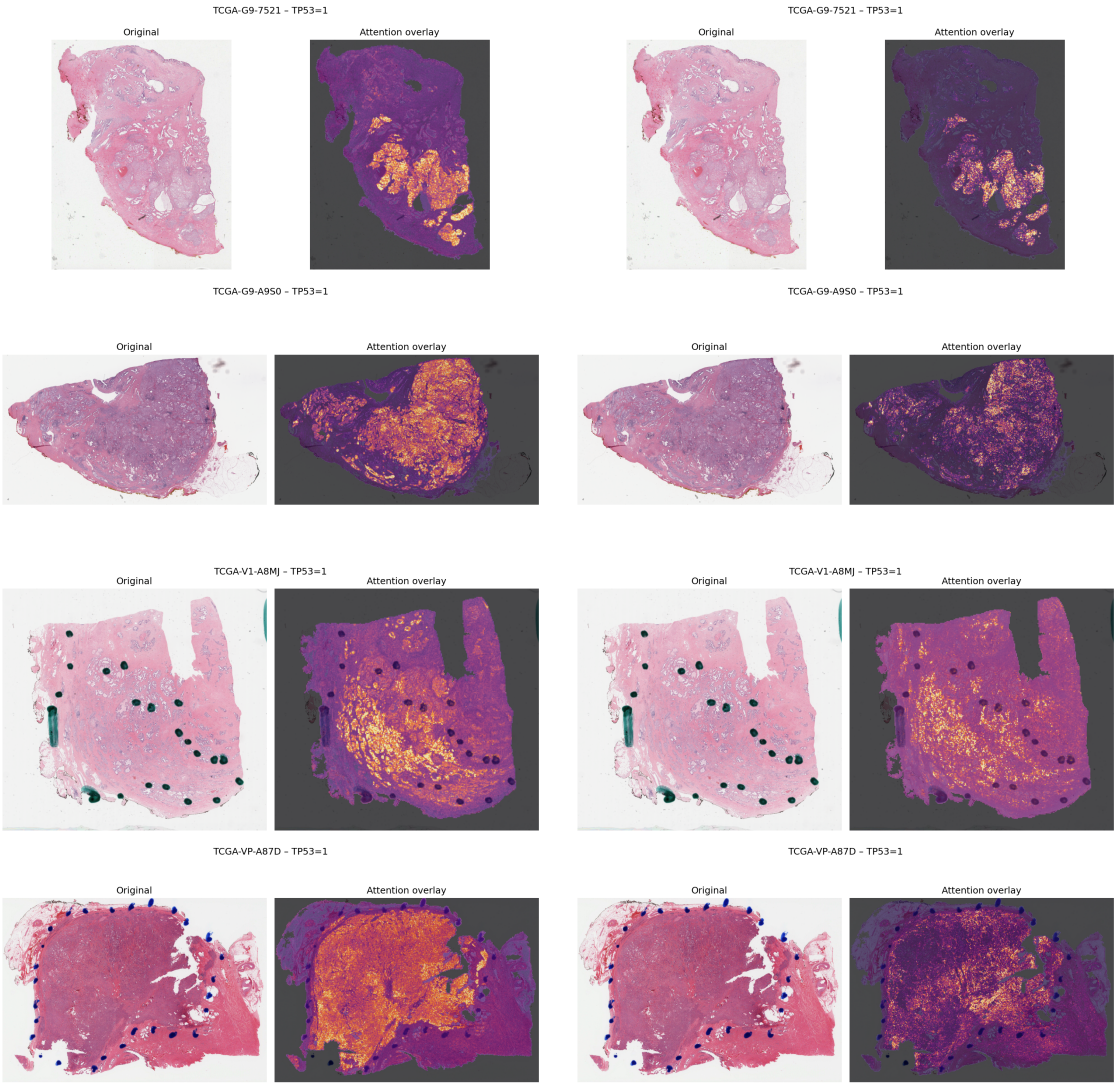


Figure 4.2: Each pair of images shows a prostate tissue sample and its attention heatmap. The left images use the three-cancer model, and the right images use the prostate-only model. The three-cancer model consistently attends more sharply to tumor regions. “Label one” = tumor; “Label zero” = non-tumor.

4.A3 Secondary Pipelines

To explore whether alternative embedding pipelines could surpass CLAM–UNI, we ran mAbMIL exclusively on the Three-Cancer dataset using Virchow2–PRISM embeddings and our two-stage MLP (attention-pooling) model. Virchow2–PRISM compresses each slide to a single 1×1280 vector, eliminating spatial variation. That proved detrimental in modeling, hence performance hovers at random (AUC ≈ 0.50), which translates into random geussing indicating no learned information. the results are presented in table 4.4, table 4.5 and table 4.6.

4.A3.1 mAbMIL + Virchow2–PRISM

Table 4.4 shows near-random AUC when using Virchow2–PRISM embeddings.As mentioned above, Virchow2–PRISM compresses each slide to a single 1×1280 vector, eliminating spatial variation. Without per-patch cues the MIL failed to learn meaningful patterns, hence performance hovers at random (AUC ≈ 0.50). The small size of the output proved insufficient to capture adequate information. As expected further regularization with auxiliary head and contrastive loss proved useless in mitigating the issue.

Table 4.4: AUC-ROC for mAbMIL on Three-Cancer with Virchow2–PRISM embeddings

Regime	Mean AUC	Max AUC	Min AUC	Std. Dev.
Three-Cancer	0.539	0.539	0.527	0.012

Table 4.5: AUC-ROC for mAbMIL on Three-Cancer with Virchow2–PRISM embeddings with auxiliary and contrastive loss

Regime	Mean AUC	Max AUC	Min AUC	Std. Dev.
Three-Cancer + Aux/Contrastive	0.506	0.507	0.505	0.001

4.A3.2 MAPT + Virchow2–PRISM

The structure of Virchow2–PRISM embeddings—single slide vectors of shape [1×1280] versus CLAM–UNI's [num_patches $\times$ 1024] matrix—suggested that a multiple-instance learning framework might not be an optimal choice. We therefore designed a two-stage, attention-based MLP classifier to account for this difference. Despite extensive hyperparameter tuning, this model on Virchow2-PRISM embeddings achieved only modest improvements in AUC. see (Table 4.6).

4.A3.3 MAPT + CLAM-UNI

We also tested our two-stage MLP on CLAM–UNI embeddings. The model performed better than the Virchow2-PRISM pipeline but performance remained inferior to mAbMIL (Table 4.7).

Table 4.6: AUC-ROC for attention-pooling MLP on Three-Cancer with Virchow2–PRISM embeddings. Despite architecture tuning, the MLP gains only 0.012 AUC over random, supporting the above explanation that spatial granularity is essential.

Regime	Mean AUC	Max AUC	Min AUC	Std. Dev.
Three-Cancer	0.519	0.542	0.486	0.024

Table 4.7: AUC-ROC for attention-pooling MLP on CLAM–UNI embeddings. MAPT benefits from CLAM–UNI's patch-level richness but still lags ≥ 0.19 AUC behind mAbMIL, confirming the superiority of the MIL aggregator for this task.

Configuration	Prostate-only	Prostate+BLCA	Three-Cancer
Mean AUC	0.365	0.469	0.585

5 Conclusion

This thesis asked whether whole-slide H&E images can predict *TP53* mutation status in prostate adenocarcinoma, — despite (i) the low prevalence of *TP53* mutations in non-advanced disease and (ii) the limited size of public prostate cohorts. To meet these challenges we pursued two main strategies: (a) leveraging a self-supervised Vision-Transformer foundation model to obtain richer patch embeddings and (b) adopting a pan-cancer training paradigm to increase morphological diversity. We also evaluated a two-stage transfer-learning pipeline, an auxiliary cancer-type head, and a supervised contrastive loss; none of these alternatives yielded any further improvements.

our main achievement was the state-of-the-art mAbMIL pipeline that replaces a lightweight CNN with a large Vision-Transformer and is trained across multiple cancer types. Across three cross-validation folds, the model achieved a pooled AUC of 0.80 (95% CI 0.72–0.88) on the held-out prostate test set, comfortably above the 0.71 reported previously.[19] Expanding the training set to five cancers did not result in further improvements, confirming that performance now lies on a shallow plateau. These results heavily indicates that TP53-specific morphological signatures are discernible in H&E slides and are potentially, at least partly conserved across tumors, which helps mitigating data scarcity in prostate cohorts. Although our experiments centered on TP53, the same training paradigm could be applied to other molecular alterations.

Our experiments show that *morphological breadth* is the single most influential driver of performance. Expanding training from a prostate-only cohort to a three-cancer cohort (PRAD, BLCA, BRCA) boosts mean AUC by $\approx$ 6.6 (0.713 $\rightarrow$ 0.780) and halves performance variance. Adding two further cancer types (LUAD, COAD) offers only a marginal gain (0.780 $\rightarrow$ 0.790), signalling a saturation point where the diversity already captured suffices for robust generalisation to held-out prostate slides. Similar marginal difference was observed when using auxiliary head and the compund loss.

Three non-exclusive explanations are plausible. First in case of the adding more data, increasing morphological heterogeneity across five cancers has potentially enlarged the hypothesis space and diluted TP53-specific signals. Second as mentioned earlier some cancers exhibit stronger TP53-associated morphology than others, so further cohorts contribute little additional useful signals. Lastly, as for the contrastive loss or the auxiliary head, the baseline

models may already capture the dominant TP53 features, so additional regularisation via these methods confers little extra benefit.

Equally important is the *alignment between representation and model*. Patch-level attention pooling (mAb-MIL on CLAM–UNI embeddings) consistently outperforms all alternatives. When patch granularity is removed (Virchow2–PRISM slide-level vectors), the AUC drifts towards chance, underscoring that fine-grained spatial cues—not just global slide summaries—are essential. Qualitative attention maps corroborate the quantitative trend: the three-cancer model sharply localises tumour foci, whereas the prostate-only model is more diffuse.

Finally, auxiliary cancer-type and supervised contrastive objectives act mainly as regularisers. They tighten the spread of results but do not exceed the baseline BCE loss with statistical significance once the dataset is morphologically rich.

Impact This work validates pan-cancer learning for mutation prediction and lays the groundwork for morphology-based biomarkers beyond prostate cancer and TP53. With further refinement, routine H&E slides could serve as a rapid, low-cost companion assay in precision oncology, while the tight confidence intervals observed on multi-cancer data underscore the practical reliability of this approach. promising direction for further exploration could include incorporating multi-institutional and prospective cohorts to test generalisability across scanners, stains, and populations or conducting a clinical-decision study to compare therapeutic choices made with and without model predictions.

Acknowledgments

Bibliography

[1] URL: https://my.clevelandclinic.org/health/body/23965-prostate.

[2] E. Baugh, H. Ke, A. Levine, et al. "Why are there hotspot mutations in the TP53 gene in human cancers?" In: *Nature* (2018).

[3] Freddie Bray et al. "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: A Cancer Journal for Clinicians* 74.3 (2024), pp. 229–263. DOI: https://doi.org/10.3322/caac.21834. eprint: https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21834. URL: https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21834.

[4] Alex Carriero et al. *The harms of class imbalance corrections for machine learning based prediction models: a simulation study*. 2024. arXiv: 2404.19494 [stat.ME]. URL: https://arxiv.org/abs/2404.19494.

[5] Richard J. Chen et al. "A General-Purpose Self-Supervised Model for Computational Pathology". In: *arXiv preprint arXiv:2308.15474* (2023).

[6] X. Chen, T. Zhang, W. Su, et al. "Mutant p53 in cancer: from molecular mechanism to therapeutic modulation." In: *nature* (2022).

[7] Shaikovski George et al. "PRISM: A Multi-Modal Generative Foundation Model for Slide-Level Histopathology". In: *arXiv preprint arXiv:2405.10254* (2024). URL: https://arxiv.org/abs/2405.10254.

[8] W. Gesztes, C. Schafer, D. Young, et al. "Focal p53 protein expression and lymphovascular invasion in primary prostate tumors predict metastatic progression". In: *nature* (2022).

[9] Donald F. Gleason, George T. Mellinger, and The Veterans Administration Cooperative Urological Research Group. "Prediction of Prognosis for Prostatic Adenocarcinoma by Combined Histological Grading and Clinical Staging". In: *journal of Urology* (1974). DOI: https://doi.org/10.1016/S0022-5347(17)59889-4.

[10] Hang Huang et al. “Significance of TP53 and immune-related genes to prostate cancer”. In: *Translational andrology and urology* 10 (Apr. 2021), pp. 1754–1768. DOI: 10.21037/tau-21-179.

[11] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. “Attention-based Deep Multiple Instance Learning”. In: *arXiv preprint arXiv:1802.04712* (2018). URL: https://arxiv.org/abs/1802.04712.

[12] M.D. Kenneth A. Iczkowski. URL: https://www.pathologyoutlines.com/topic/prostategrading.html.

[13] Prannay Khosla et al. *Supervised Contrastive Learning*. 2021. arXiv: 2004.11362 [cs.LG]. URL: https://arxiv.org/abs/2004.11362.

[14] Ming Y. Lu et al. “Data Efficient and Weakly Supervised Computational Pathology on Whole Slide Images”. In: *arXiv preprint arXiv:2004.09666* (2020). URL: https://arxiv.org/abs/2004.09666.

[15] Piccininni M et al. “Understanding random resampling techniques for class imbalance correction and their consequences on calibration and discrimination of clinical risk prediction models”. In: *biomed info* (2024). DOI: 155:104666.doi:10.1016/j.jbi.2024.104666.

[16] Tornesello ML. “TP53 mutations in cancer: Molecular features and therapeutic opportunities (Review)”. In: *Int J Mol Med* (2025). URL: doi:%2010.3892/ijmm.2024.5448.

[17] Hortobágyi T. Murnyák B. “Immunohistochemical correlates of TP53 somatic mutations in cancer”. In: *arXiv preprint arXiv:1802.04712* (2016).

[18] Javad Noorbakhsh et al. “Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images”. In: *Nature Communications* 11 (2020), p. 6367. DOI: 10.1038/s41467-020-20030-5.

[19] Marija Pizurica et al. “Whole Slide Imaging-Based Prediction of TP53 Mutations Identifies an Aggressive Disease Phenotype in Prostate Cancer”. In: *Cancer Research* (2023). DOI: 10.1158/0008-5472.CAN-22-3113.

[20] Scientific Figure on ResearchGate. *Automatic Diagnostic Tool for Predicting Cancer Grade in Bladder Cancer Patients Using Deep Learning*. Accessed: 2025-04-03. 2021. URL: https://www.researchgate.net/figure/WSI-images-are-stored-in-a-pyramidal-format-where-the-base-image-corresponds-to-the_fig2_353893643.

[21] The results <published or shown> here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

[22] Jiazhuo Wang, Jason Xu, and Xuejun Wang. *Combination of Hyperband and Bayesian Optimization for Hyperparameter Optimization in Deep Learning*. 2018. arXiv: 1801.01596 [cs.CV]. URL: https://arxiv.org/abs/1801.01596.

[23] Eric Zimmermann et al. "Virchow2: Scaling Self-Supervised Mixed Magnification Models in Pathology". In: *arXiv preprint arXiv:2408.00738* (2024). URL: https://arxiv.org/abs/2408.00738.