Progressive Fine-Tuning of Vision-Language Models for Urine Cytology-Based Bladder Cancer Staging

Pedro José Costa Oliveira
Master in Bioinformatics and Computational Biology
Department of Computer Science and Department of Biology
Faculty of Sciences of the University of Porto
2026

Progressive Fine-Tuning of Vision-Language Models for Urine Cytology-Based Bladder Cancer Staging

Pedro José Costa Oliveira

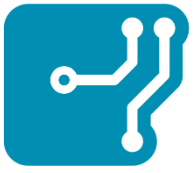
Dissertation carried out as part of the Master in Bioinformatics and Computational Biology

Department of Computer Science and Department of Biology

2026

Supervisor
Hélder Oliveira, Associate Professor, FCUP, INESC-TEC

Co-Supervisor
Raphaël Canadas, Principal Investigator, FMUP



INESCTEC
ASSOCIATE LABORATORY
PORTUGAL

Acknowledgements

I would like to express my gratitude to my supervisor, Dr. Helder Oliveira, and co-supervisor, Dr. Raphaël Canadas, for their guidance and clinical insights throughout this work. I am especially thankful to Dr. Hugo S. Oliveira for his technical mentorship, constant availability, and dedication throughout every stage of this thesis. His involvement from the early pipeline decisions through to the final revisions was central to bringing this work together, and his willingness to always engage deeply with the work no matter the stage made completing this thesis a far better experience than it might otherwise have been.

Funding

FCT via the project CELLo (Ref. 2022.05237.PTDC); and by “la Caixa” Foundation – CaixaImpulse Program under the project CELLECTED (LCF/TR/CI24/56010020) attributed to the co-supervisor Raphaël Canadas from the Faculty of Medicine of the University of Porto, which partially supported the costs of the developed work.

UNIVERSITY OF PORTO

Resumo

Faculty of Science of University of Porto
Department of Computer Science and Department of Biology
INESC TEC Visual Computing and Machine Intelligence Group

Mestrado em Bioinformatica e Biologia Computacional

Ajuste Fino Progressivo de Modelos Visão-Linguagem para Estadiamento do Cancro da Bexiga Baseado em Citologia Urinária

por Pedro OLIVEIRA

O estadiamento do cancro da bexiga a partir de imagens de citologia urinária é uma tarefa clinicamente relevante e computacionalmente desafiante, caracterizada por elevada similaridade visual entre classes, desequilíbrio significativo de classes e escassez de conjuntos de dados anotados. Esta tese investiga a adaptação de modelos de fundação visão-linguagem a este problema através de um framework de fine-tuning progressivo aplicado ao BiomedCLIP, um modelo contrastivo visão-linguagem pré-treinado em 15 milhões de pares imagem-texto biomédicos. O framework introduz incrementalmente componentes arquitecturais e de optimização, permitindo uma análise sistemática da contribuição de cada decisão de desenho para o desempenho global da classificação.

No conjunto de dados primário Cello, composto por imagens de microscopia de linhas celulares de cancro da bexiga, o pipeline atinge uma precisão balanceada final de 0.9404 ± 0.0002, com AUC próximo do perfeito em todos os cinco estadios tumorais. A selecção do backbone foi informada por análise qualitativa de mapas de atenção, que revelou que o BiomedCLIP mantém um foco espacial consistente em estruturas celulares morfologicamente relevantes, suportando a sua priorização face a uma alternativa com desempenho marginalmente superior mas menos interpretável. A avaliação cross-dataset no ISIC 2019 e Raabin-WBC demonstra que o pipeline generaliza eficazmente sob distribuição shift, atingindo precisão balanceada de 0.7546 e 0.9745 respectivamente, sem qualquer adaptação específica ao domínio. A profundidade de adaptação óptima varia entre domínios, indicando que o fine-tuning progressivo é flexível mas sensível ao contexto.

Os resultados estabelecem que o fine-tuning progressivo de modelos de fundação visão-linguagem biomédicos constitui um framework eficaz e interpretável para a classificação de imagens citológicas em diversos domínios de imagem, com a análise de mapas de atenção a confirmar um foco espacial clinicamente fundamentado ao longo do pipeline.

Palavras-chave **-** estadiamento do cancro da bexiga, citologia urinária, modelos visão-linguagem, BiomedCLIP, fine-tuning progressivo, aprendizagem contrastiva, classificação de imagem médica, mapas de atenção, avaliação cross-dataset

UNIVERSITY OF PORTO

Abstract

Faculty of Science of University of Porto
Department of Computer Science and Department of Biology
INESC TEC Visual Computing and Machine Intelligence Group

Masters in Bioinformatics and Computational Biology

Progressive Fine-Tuning of Vision-Language Models for Urine Cytology-Based Bladder Cancer Staging

by Pedro OLIVEIRA

Bladder cancer staging from urine cytology images is a clinically relevant yet computationally challenging task, characterised by high inter-class visual similarity, significant class imbalance, and the absence of large annotated datasets. This thesis investigates the adaptation of vision-language foundation models to this problem through a progressive fine-tuning framework applied to BiomedCLIP, a contrastive vision-language model pretrained on 15 million biomedical image-text pairs. The framework incrementally introduces architectural and optimisation components, enabling systematic analysis of the contribution of each design decision to overall classification performance.

On the primary Cello dataset of bladder cancer cell line microscopy images, the pipeline achieves a final balanced accuracy of 0.9404 ± 0.0002, with near-perfect AUC across all five tumour stages. Backbone selection was informed by qualitative attention map analysis, which revealed that BiomedCLIP maintains consistent spatial focus on morphologically relevant cellular structures, supporting its prioritisation over a marginally higher-performing but less interpretable alternative. Cross-dataset evaluation on ISIC 2019 and Raabin-WBC demonstrates that the pipeline generalises effectively under distribution shift, achieving balanced accuracy of 0.7546 and 0.9745 respectively without any dataset-specific adaptation. The optimal adaptation depth varies across domains, indicating that progressive fine-tuning is flexible but context-sensitive.

The results establish that progressive fine-tuning of biomedical vision-language models provides an effective and interpretable framework for cytological image classification across diverse imaging domains, with attention map analysis confirming clinically grounded spatial focus throughout the pipeline.

Keywords - bladder cancer staging, urine cytology, vision-language models, BiomedCLIP, progressive fine-tuning, contrastive learning, medical image classification, attention maps, cross-dataset evaluation

Contents

List of Figures

List of Tables

List of Abbreviations

CLIP Contrastive Language-Image Pre-training. 27, 31, 34, 40, 41, 42, 45, 46, 55

CNN Convolutional Neural Network. 11, 12, 13, 14, 27, 30

MIL Multiple Instance Learning. 13, 16, 17

TPS The Paris System. 19, 22

ViT Vision Transformer. 14, 24, 27, 30, 44

VLM Vision-Language Model. 11, 13, 27, 28, 31

Chapter 1

Introduction

1.1 Clinical Context and Problem Statement

Bladder cancer is one of the most common urological malignancies, ranking as the ninth most common cancer globally (Bray et al., 2024) and is associated with high recurrence rates and cost-intensive surveillance after initial treatment. Accurate and timely staging is critical: the distinction between non-muscle-invasive and muscle-invasive disease fundamentally governs the clinical pathway, with implications ranging from endoscopic resection to radical cystectomy. However, conventional staging methods rely heavily on cystoscopy and histopathological analysis, which are subject to inter-observer variability and lack objective, quantitative criteria. These methods are also invasive, which is particularly problematic during patient follow-up: given the high recurrence rates of bladder cancer, patients require repeated in-office cystoscopies over time, incurring high costs and procedural risk. Urine cytology offers a non-invasive alternative for diagnosis and surveillance, but current cytological assessment remains largely subjective and pathologist-dependent. Critically, existing cytology reporting frameworks are not quantitative and do not support fine-grained stratification of tumour severity: under the Paris System, the current standard for cytological reporting, a cytology-positive result is treated as high-grade by default, without further distinction. This lack of stratification is clinically significant, as it directly determines the intensity of the follow-up surveillance programme and the frequency of hospital visits required. This motivates a non-invasive, quantitative approach to tumour staging directly from cytological specimens. Figure 1.1 illustrates the staging of bladder cancer through the different layers of the bladder wall, highlighting the anatomical basis of tumour classification.

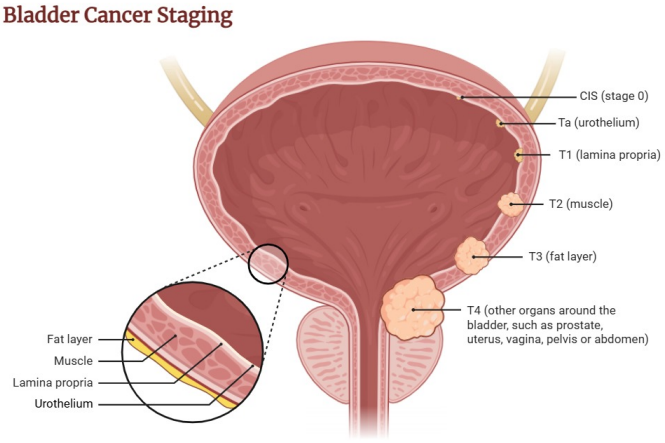


FIGURE 1.1: Section of bladder tumour stages according to the T grading of the TNM classification.

To address this gap, the Faculty of Medicine of the University of Porto (FMUP) developed a processing pipeline to produce high-resolution brightfield microscopy images of isolated urothelial cells using ImageStream technology, which combines flow cytometry with digital imaging to capture single-cell morphological profiles at scale. Each image corresponds to an individual cell captured from a urine-derived sample and is associated with a known tumour stage label derived from the originating cell line. The resulting dataset, referred to as CELLo, comprises cells from established bladder cancer cell lines spanning five tumour stages: NA, Ta, T2, T3, and T4, each exhibiting distinct morphological characteristics that reflect the degree of tumour invasion. The classification task addressed in this thesis is assigning each cell image to its corresponding T tumour stage (of the TNM - Tumour Node Metastases - classification), a problem that is both clinically relevant and computationally challenging due to high inter-class visual similarity and significant class imbalance across stages.

It is worth noting that bladder cancer classification encompasses several complementary systems beyond TNM staging, including tumour grade and the Paris System for cytological reporting. The T component addressed in this thesis specifically determines the depth of tumour invasion into the bladder wall, a distinction that is clinically critical as muscle invasion status is the primary determinant of patient survival probability and dictates the subsequent treatment pathway. While prior computational approaches to urine cytology have operated at the cell-patch, slide, or whole-slide image level (as detailed in Chapter 2), to the best of our knowledge, no prior work has addressed T-stage prediction

from a single, isolated cell. This thesis investigates whether such fine-grained morphological information, isolated at the level of the individual cell, is sufficient to support this clinically meaningful distinction.

1.2 AI for Urine Cytology and Bladder Cancer Diagnosis

Automated analysis of urine cytology has attracted growing interest, driven by the limitations of manual cytological assessment, including high interobserver variability and limited sensitivity for low-grade bladder cancer lesions.

From a cytological perspective, automated options have relied on coloured samples (e.g., H&E staining), which entail associated costs, processing time, and toxicity for the lab technician. Recent work has demonstrated that fully autonomous, clinical-grade cytopathology is achievable using staining-free optical imaging at scale, suggesting that the field is increasingly moving toward automated alternatives to traditional staining-based workflows (Nitta et al., 2026). From the computational side, early approaches relied on convolutional neural networks applied to individual cell images or small image patches. Custom augmentation strategies and domain-specific training procedures demonstrated that automated cell-level classification could match or exceed the performance of trained cytotechnologists on malignancy detection tasks (Kaneko et al., 2022). Further work explored the correlation between cytopathological image features and histopathological outcomes, demonstrating the feasibility of predicting downstream tissue diagnosis from cytology images alone (Liu et al., 2022). Despite these advances, CNN-based methods remained constrained by fixed label sets and single imaging modalities, primarily addressing binary or coarse diagnostic categories rather than fine-grained tumour staging.

Subsequent work extended classification to the slide level through weakly supervised multiple-instance learning frameworks, enabling diagnosis from whole digitised cytology slides without requiring cell-level annotation. Prospective multicentre validation demonstrated that such systems could achieve clinically meaningful sensitivity for detecting high-grade urothelial carcinoma (Tsuneki et al., 2022), with large-scale evaluation further confirming their clinical utility across multiple centres (Wu et al., 2024). However, these approaches predominantly target binary malignancy detection aligned with Paris System categories rather than the finer TNM staging granularity required for treatment planning.

More recently, foundation model approaches have begun to address the data scarcity and domain generalisation limitations of fully supervised methods. Self-supervised contrastive pretraining on large unlabelled pathology corpora has demonstrated improved transfer to cytological tasks compared to ImageNet-initialised baselines (Ciga et al., 2022). Large-scale vision transformers pretrained on diverse pathology corpora have further demonstrated strong transfer to cytological tasks with limited annotation (Bilal et al., 2025). The present work falls within this direction, extending foundation model adaptation to the specific problem of TNM tumour stage classification from microscopy images of urine-derived bladder cancer cell lines using a progressive fine-tuning framework.

1.3 Research Motivation

Vision-language foundation models (Radford et al., 2021; Zhang et al., 2023) represent a particularly promising direction for medical image classification. Unlike standard deep learning models that rely solely on visual features, these models are pretrained on large corpora of paired image-text data, allowing them to encode rich semantic representations that generalise across imaging domains. This alignment between visual and textual information is especially valuable in medical imaging, where class distinctions often correspond to nuanced morphological differences that benefit from semantic grounding through clinical descriptors.

Adapting these models to specialised tasks such as bladder cancer cell staging requires careful fine-tuning. Training from scratch is impractical given the limited availability of annotated medical data, making progressive fine-tuning of pretrained foundation models a natural and principled approach. Furthermore, interpretability is a critical requirement in clinical decision support systems — a model that achieves high classification accuracy but attends to non-diagnostic image regions cannot be safely deployed in a clinical workflow. This motivates the joint evaluation of classification performance and spatial attention quality as complementary criteria for model selection and validation.

1.4 Objectives

This work explores the application of vision-language foundation models to bladder cancer cell classification, developing and evaluating a progressive fine-tuning pipeline that

incrementally introduces architectural and optimisation components to improve classification performance. These considerations motivate the following research objectives, developed to address gaps identified in the literature and the clinical requirements of the target application.

- To evaluate candidate vision-language backbones through zero-shot classification and qualitative attention map analysis, identifying the most suitable model for downstream adaptation on bladder cancer cell microscopy images.
- To develop and evaluate a progressive fine-tuning pipeline across seven stages, isolating the contribution of each component, including frozen feature extraction, multi-scale fusion, partial backbone unfreezing, learnable temperature scaling, soft prompt tuning, and mixture-of-experts projection.
- To assess the generalisation capacity of the selected pipeline on two external benchmark datasets spanning different imaging modalities, evaluating robustness under distribution shift.
- To validate the interpretability of the final model through attention map analysis, confirming that predictions are grounded in morphologically relevant cellular structures.

1.5 Outline of the Thesis

The remainder of this thesis is organised as follows. Chapter 2 presents a literature review covering bladder cancer diagnosis, vision-language foundation models, and fine-tuning strategies for medical imaging. Chapter 3 describes the methodology employed throughout the project, including the datasets used, the multimodal architecture, the progressive fine-tuning pipeline, and the training and evaluation procedures. Chapter 4 presents the experimental results across all pipeline stages and datasets, discussing the findings in the context of the research objectives. Chapter 5 presents the conclusions drawn from this work, acknowledges its limitations, and outlines directions for future research.

Chapter 2

Literature Review

This chapter presents a comprehensive review of the current literature on bladder cancer diagnosis and staging, computational approaches to medical image analysis, vision-language foundation models and their biomedical adaptations, and strategies for fine-tuning and adapting these models to specialised imaging tasks. Together, these areas establish the theoretical and empirical foundations for the methodology proposed in this thesis.

2.1 Bladder Cancer - Clinical Overview

Bladder cancer is the ninth most common cancer globally and the sixth most common in men, with urothelial carcinoma accounting for approximately 90% of cases (Kamat et al., 2016; Bray et al., 2024; Cancer Research UK, 2026). Major risk factors include tobacco smoking, occupational exposure to aromatic amines, and exposure to arsenic in drinking water. Beyond its clinical impact, bladder cancer carries a substantial economic burden, representing one of the most expensive cancers to treat owing to high recurrence rates and the need for lifelong surveillance (Kamat et al., 2016).

Bladder cancer is broadly classified into two categories based on the depth of tumour invasion. Non-muscle-invasive bladder cancer (NMIBC), which constitutes approximately 70–75% of diagnoses, encompasses tumours confined to the urothelium (stage Ta) or lamina propria (stage T1), as well as carcinoma in situ (CIS). Muscle-invasive bladder cancer (MIBC) refers to tumours that have penetrated the detrusor muscle (stage T2) or extended beyond it (stages T3 and T4). This distinction is clinically important: NMIBC is

typically managed with transurethral resection of bladder tumour (TURBT) and intravesical therapy, whereas MIBC often requires radical cystectomy, potentially preceded by neoadjuvant chemotherapy (Babjuk et al., 2022). These categories correspond to the T component of the TNM staging system, which underpins clinical decision-making and is consistently used in major international guidelines, including those of the European Association of Urology (EAU) and the American Urological Association (AUA).

Despite well-established staging frameworks, accurate diagnosis and staging of bladder cancer remain challenging in clinical practice. White-light cystoscopy (WLC) is the current gold standard for initial detection, but has been reported to have a sensitivity of 71% and specificity of 72% (Mowatt et al., 2011), with misdiagnosis rates particularly high for subtle lesions such as carcinoma in situ. Histopathological assessment, while indispensable for confirming tumour grade and invasion depth, is subject to interobserver variability, particularly in borderline cases of NMIBC and MIBC, leading to inconsistent staging and suboptimal treatment decisions (Al-Sattar et al., 2026). Tosoni et al. demonstrated significant interobserver differences in both grading and staging of superficial bladder cancer, with 39% of biopsies showing disagreement in tumour grade between reviewing pathologists (Tosoni et al., 2000).

These limitations have motivated growing interest in computational approaches to bladder cancer diagnosis. AI-based models, particularly those leveraging convolutional neural networks and deep learning, have demonstrated promising results in tumour detection, segmentation, and grading from cystoscopic and histopathological images (Ferro et al., 2023; Ma et al., 2024). More recently, computational approaches have extended to urine cytology analysis, where deep learning pipelines have been developed to automate the identification and characterisation of atypical urothelial cells, addressing the inherent subjectivity and interobserver variability of manual cytological assessment (Awan et al., 2021). Despite these advances, challenges remain regarding standardisation, external validation, and integration into clinical workflows.

2.2 Computational Methods in Medical Image Analysis

Deep learning has fundamentally reshaped computational approaches to medical image analysis over the past decade. The rapid adoption of convolutional neural networks across tasks ranging from lesion detection to tumour grading and whole slide image classification has been widely documented, establishing that while deep learning methods

demonstrated strong performance across modalities, medical imaging presents distinct challenges compared to natural image tasks, including limited annotated data, significant class imbalance, and high inter-class visual similarity, which constrains the straightforward application of standard architectures (Litjens et al., 2017).

Within the specific imaging domains addressed in this thesis, supervised deep learning methods have achieved considerable results. On the ISIC 2019 skin lesion benchmark, ensembles of multi-resolution EfficientNet models with auxiliary metadata achieved a normalised balanced multiclass accuracy of 63.6%, placing first in the ISIC 2019 challenge and highlighting both the effectiveness of modern CNN architectures and the importance of addressing the dataset's severe class imbalance (Gessert et al., 2020). For white blood cell classification on the Raabin-WBC dataset, deep convolutional approaches have been shown to reliably distinguish between cell morphologies with high accuracy when sufficient labelled examples are available (Özcan and Öztürk, 2024). In the domain most directly relevant to this thesis, deep learning applied to urine cytology images demonstrated that digital cell profiles derived from convolutional networks can support risk stratification of bladder cancer patients from cytological specimens (Awan et al., 2021). Despite their strong in-distribution performance, these approaches share a common limitation: they rely on large pools of expert-annotated training data, are optimised for fixed label sets, and do not generalise to unseen classes or imaging domains without complete retraining.

The assumption that transfer learning from large natural image datasets such as ImageNet provides consistent benefits in medical imaging has been critically examined in prior work. Raghu et al. demonstrated that, in certain medical imaging settings, ImageNet pretraining offers limited performance improvements over training from scratch, and that in some cases lightweight models trained directly on medical data can achieve comparable performance to large ImageNet-pretrained architectures (Raghu et al., 2019). Their analysis attributes this to the substantial domain gap between natural and medical images, which limits the reuse of higher-level pretrained features, suggesting that the effectiveness of transfer learning depends heavily on the degree of domain similarity and task specificity.

Complementing this perspective, subsequent work has shown that model performance

is also strongly influenced by the choice of fine-tuning strategy. Kumar et al. demonstrated that full fine-tuning can lead to degradation of pretrained representations, particularly under distribution shift or in limited data regimes, and that adaptation strategy plays a critical role in determining downstream generalisation performance (Kumar et al., 2022). Staged approaches such as linear probing followed by fine-tuning can help mitigate this effect by preserving useful pretrained features while adapting to the target domain.

More recently, contrastive vision-language pretraining has introduced an alternative paradigm for medical image classification. Models such as CheXzero (Tiu et al., 2022) demonstrate that aligning images and text via contrastive objectives enables strong zero-shot transfer for medical imaging tasks, without requiring task-specific annotation, suggesting that vision-language models may offer greater flexibility under limited supervision, particularly in settings where label acquisition is expensive or inconsistent.

Together, these findings highlight two persistent challenges in the imaging domains addressed in this thesis, particularly in urothelial cell classification. First, supervised approaches remain dependent on large annotated datasets and fixed class vocabularies, constraining their scalability across clinical settings. Second, the task-dependent nature of transfer learning benefits, combined with the sensitivity of model performance to fine-tuning strategy, motivates the adoption of more principled adaptation approaches. These considerations also indicate a broader shift towards multimodal pretraining that reduces reliance on fixed label sets.

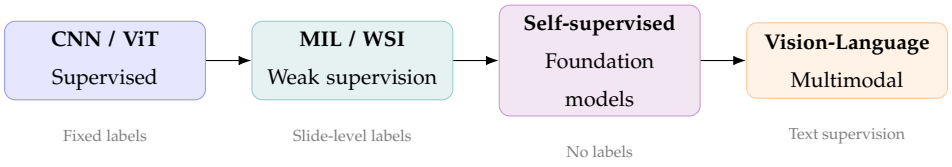


FIGURE 2.1: Evolution of computational approaches in medical image analysis, reflecting a progressive reduction in annotation dependency from fully supervised CNN methods to vision-language foundation models.

2.3 Literature Search Strategy

The literature reviewed in this chapter was selected to directly support the methodological decisions underlying the proposed pipeline. Rather than a large formal systematic review, papers were identified through targeted searches of PubMed, Google Scholar, IEEE Xplore, and arXiv using terms specific to each pipeline component — including bladder

cancer diagnosis, urine cytology analysis, vision-language pretraining, contrastive learning, partial backbone fine-tuning, prompt learning, and mixture of experts architectures. Additional works were then included on the basis of their foundational status, or their direct relevance to one or more components of the proposed pipeline, to the datasets evaluated, or to the clinical context addressed in this thesis. The review is intentionally focused rather than exhaustive, prioritising depth of coverage in areas directly informing the experimental design of bladder cell urine tasks over breadth across the wider medical imaging literature.

TABLE 2.1: Literature search strategy and inclusion criteria for urinary and bladder cancer-related studies.

Component	**Search criteria / strategy**
Databases	PubMed, Google Scholar, IEEE Xplore, and arXiv are primary sources for biomedical and machine learning literature.
Core clinical focus	Bladder cancer diagnosis, urothelial carcinoma, urine cytology analysis, and cystoscopy imaging, with emphasis on clinically relevant urinary tract pathology.
Methodological focus	Deep learning for medical imaging, including convolutional neural networks, vision-language models, contrastive learning, and CLIP-based pretraining approaches.
Model adaptation strategies	Partial backbone fine-tuning, layer-wise learning rate decay, prompt learning (CoOp / soft prompts), and mixture of experts (MoE) architectures with sparse routing.
Search strategy	Targeted keyword search per pipeline component, complemented with forward and backward citation tracing from key foundational works.
Inclusion criteria	Studies with direct relevance to at least one pipeline component, urinary/bladder cancer imaging tasks, or commonly used benchmark datasets in medical imaging.
Exclusion criteria	Theoretical-only contributions without empirical validation, non-transferable domains, and works not informing model design or experimental evaluation.
Review scope	Focused (non-systematic) literature selection prioritising methodological relevance over exhaustive coverage.

A final total of 29 studies were selected for additional discussion based on the specified inclusion criteria of Table 2.1. To facilitate systematic comparison across many genres of methods, the retrieved works were classified according to their core methodological nature and conceptual focus, creating a taxonomy. The chosen research was divided into five categories: CNN/ViT-based approaches, MIL/WSI-based approaches, semi-automated

scoring systems, foundation and self-supervised approaches, and vision-language models. This taxonomy makes it easier to understand how computational urine cytology and bladder cancer analysis have evolved, as well as how data-driven and pretraining-based paradigms have gradually replaced handcrafted feature engineering.

2.3.1 Previous Surveys

Prior to the emergence of foundation models and Vision-Language Models (VLMs), automated urine cytology analysis was primarily addressed using supervised deep learning approaches based on Convolutional Neural Networks (CNNs) models integrated into conventional computational cytology pipelines (Jiang et al., 2023; Nabiyouni and Chiou, 2026). These early methods focused on tasks such as urothelial cell classification, malignancy detection, atypical cell screening, and bladder cancer diagnosis from urine cytology specimens. More recently, weakly supervised learning and multiple-instance learning paradigms have been introduced to reduce annotation requirements and better exploit slide-level labels, particularly in digital pathology settings (Khoraminia et al., 2023; Superdock et al., 2026).

Collectively, these survey studies demonstrated the feasibility of developing an automated cytological assessment, with diagnostic performance comparable to that of experienced cytotechnologists and pathologists. Nevertheless, most existing approaches were designed for a narrowly defined subset of classification tasks, which often relied on limited quantities of annotated data, and were generally restricted to binary malignancy detection or coarse diagnostic categorisations rather than detailed tumour staging assessment.

Recent advances in artificial intelligence have been applied to pathology, shifting from traditional learning approaches towards self-supervised learning, foundation models, and large-scale pathology pretraining, enabling the extraction of transferable visual representations while reducing dependence on expert annotations (Khoraminia et al., 2023; Narimani et al., 2026; Pan et al., 2022).

This shift is inspired by the success of large-scale vision and vision-language pretraining, enabling the construction of fundational backbones to improve generalisation across staining protocols and different downstream diagnostic tasks. Despite this progress, the majority of published studies continue to focus on cancer detection and diagnostic classification rather than fine-grained tumour staging and risk stratification from urine cytology

images. Furthermore, systematic comparisons of different foundation-model adaptation strategies, including full fine-tuning, parameter-efficient tuning, prompt learning, and mixture-of-experts architectures, remain limited in recent literature.

Table 2.2 summarises representative survey studies in automated urine cytology, computational cytopathology, and bladder cancer image analysis, highlighting their methodological approaches, datasets, principal findings, and remaining limitations. The table also illustrates the evolution of the field from task-specific CNN-based models to foundation-model-driven approaches that leverage large-scale pretraining and more efficient adaptation strategies.

TABLE 2.2: Condensed summary of major surveys in urine cytology and bladder cancer computational pathology.

Survey	Method / Scope	Key Contributions / Limitations
(Jiang et al., 2023)	Computational cytology (multi-organ)	Comprehensive review of CNN-based and classical ML pipelines for cytology tasks. Strong coverage of DL paradigms, but not bladder-specific, and limited discussion of multimodal/foundation models.
(Khoraminia et al., 2023)	Bladder cancer digital pathology	Systematic review of AI in bladder cancer histopathology and cytology. Covers diagnosis, grading, prognosis, and workflow automation. Limitations include reliance on supervised learning and limited external validation.
(Nabiyouni and Chiou, 2026)	Urine cytopathology AI	Focused systematic review on urine cytology AI for urothelial carcinoma detection. Demonstrates strong performance in malignancy classification, but mostly in binary tasks and with high annotation dependency.
(Superdock et al., 2026)	Computational pathology (bladder cancer)	Scoping review of digital pathology pipelines, including weak supervision and MIL. Highlights scalable slide-level learning, but limited discussion of foundation models and tumour staging.

Table 2.2 continued...

Survey	Method / Scope	Key Contributions / Limitations
(Narimani et al., 2026)	Multimodal bladder cancer AI	Broad clinical review spanning pathology, cytology, and imaging. Emphasises the potential of multimodal AI across diagnostic workflows. Lacks systematic benchmarking and quantitative comparisons.

A clear observation emerging from Table 2.2 is that the narrow reviewed literature naturally clusters into three main methodological categories.

First, traditional computational cytology and machine learning approaches rely on handcrafted feature engineering and classical classifiers, typically applied to structured cytological descriptors and early digital imaging pipelines.

Second, supervised deep learning methods, particularly CNNs, still dominate modern digital pathology workflows in an end-to-end learning fashion for tasks such as classification, detection, and grading in both urine cytology and bladder cancer histopathology.

Finally, more recent work (Narimani et al., 2026) shows a clear shift toward foundation-model and multimodal paradigms, including self-supervised representation learning, VLM, weakly supervised Multiple Instance Learning (MIL), promoting multimodal clinical integration across pathology and imaging data.

This multimodal perspective motivates the exploration and systematic evaluation of modern foundation-model adaptation strategies for urine cytology, with particular emphasis on improving generalisation and enabling fine-grained bladder cancer staging.

2.4 Retrieved Works and Categorisation

To enable a structured synthesis of the literature, the selected studies were organised according to their primary methodological paradigm and level of model abstraction. This categorisation separates traditional machine learning approaches, supervised deep learning methods, and more recent foundation-model and multimodal strategies.

Traditional approaches primarily rely on handcrafted features and classical classifiers applied to cytological or histopathological descriptors. In contrast, supervised deep learning methods, particularly convolutional neural networks, have enabled end-to-end learning directly from imaging data, significantly improving performance in classification and

detection tasks. More recent studies explore self-supervised learning, weakly supervised multiple-instance learning, and vision-language foundation models, aiming to reduce annotation requirements and improve cross-domain generalisation.

This categorisation provides a coherent framework for analysing methodological trends in urine cytology and bladder cancer analysis, and it highlights the gradual transition towards scalable, pretraining-driven paradigms capable of supporting more fine-grained diagnostic and staging tasks.

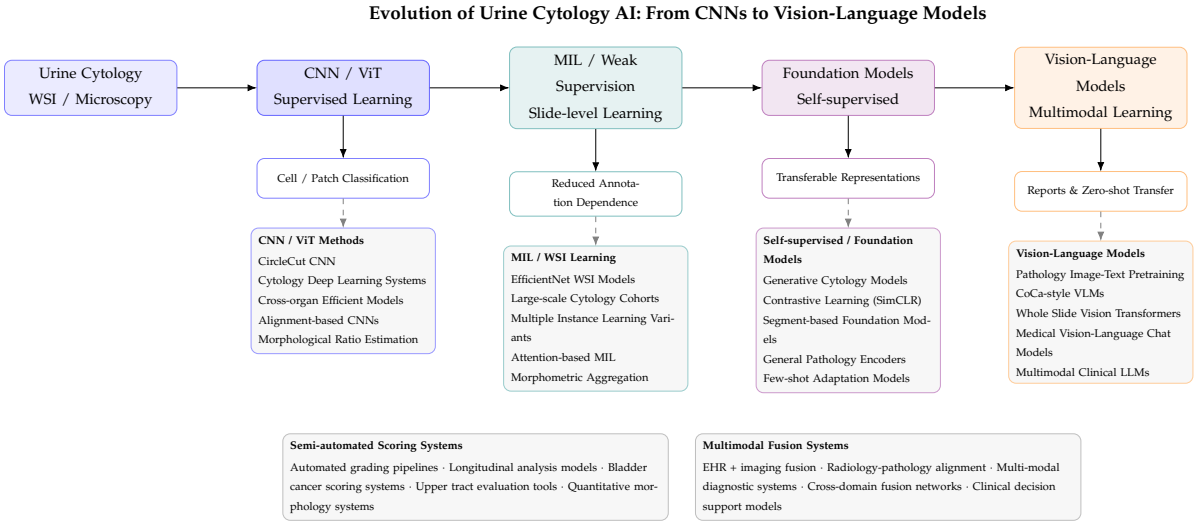


FIGURE 2.2: Methodological evolution of urine cytology analysis from CNN-based pipelines to vision-language foundation models. Categories represent major research directions, while cross-cutting themes capture semi-automated scoring and multimodal fusion strategies.

2.4.1 CNN/VIT Based Approaches

Early computational approaches in urological imaging, particularly in bladder cancer analysis and urine cytology, largely rely on fully supervised deep learning models. CNNs initially dominated this space due to their strong inductive bias for local texture extraction, which is critical for capturing nuclear atypia, chromatin irregularities, and cellular morphology.

More recent Vision Transformer (ViT) extend this paradigm by modelling global dependencies across entire cell fields, improving robustness to staining variability and heterogeneous slide conditions. In bladder cancer diagnosis, these methods are typically applied at the patch or cellular level, focusing on malignancy classification and grading,

as well as the detection of atypical cells. Despite strong performance, they remain constrained by the requirement of dense annotations and limited generalisation across acquisition protocols and institutions. A typical CNN/VIT-based approach pipeline contains:

1. **Cell / region detector** - rule-based segmentation or a lightweight detector (YOLO, RetinaNet) to crop candidate cells.
2. **CNN encoder / backbone** - ResNet-50/101, EfficientNet-B1/B6, InceptionV3; typically pretrained on ImageNet or a related cytology domain (lung, cervical).
3. **Data-augmentation layer** - rotation, flipping, stain normalisation (Macenko, Vahadane), CutMix, CircleCut.
4. **Classification head** - fully connected layers with softmax for binary (negative/positive) or multi-class TPS output.

Kaneko et al. (2022) introduce a three-class urothelial cell classification system using EfficientNet-B6 with ArcFace metric learning and a customised CutMix augmentation strategy termed CircleCut, achieving cell-level accuracy exceeding 0.90 across all subtypes and surpassing cytotechnologists in binary malignancy classification.

Liu et al. (2022) extends the cell-level paradigm toward cyto-histological correlation, training a ResNet-101 Faster R-CNN to detect malignant cells in urine cytopathological images and predict histopathological outcomes, demonstrating feasibility on both internal and external cohorts with AUCs of 0.90 and 0.93, respectively.

Yang et al. (2025) revisit the foundational N/C ratio criterion of the Paris System through AI-assisted quantification using the AIxURO platform, demonstrating that suspicious cells consistently exhibit a median N/C ratio of 0.66 across 106 WSIs, challenging the current TPS diagnostic threshold of >0.7 and motivating a data-driven recalibration of cytomorphological criteria.

TABLE 2.3: CNN/VIT Based Approach — CNN-based cell/patch-level systems.

Work	Backbone	Key contribution	Performance	Main limitations
Kaneko et al. (2022)	EfficientNet-B6 + ArcFace	CircleCut augmentation (customised CutMix); three-class urothelial cell classification (benign/atypical / malignant); performance surpasses cytotechnologists	AUC 0.99; accuracy 0.95; >0.90 across all subtypes	Single-centre; cell-level only; pre-segmented inputs required

Table 2.3 continued...

Work	Backbone	Key contribution	Performance	Main limitations
Liu et al. (2022)	ResNet-101 Faster R-CNN	Cyto-histo correlation: malignant cell detection used to predict histopathological outcome from cytopathological images	Internal AUC 0.90; external AUC 0.93	Retrospective; dense cell-level annotation required; no slide-level aggregation
Yang et al. (2025)	AIxURO (deep instance segmentation)	AI-assisted N/C ratio quantification across 106 WSIs challenges the TPS threshold of >0.7 for SHGUC/HGUC; proposes revised cutoff of 0.66	Suspicious cells: median N/C 0.66 vs. atypical 0.58 ($p < .001$); nuclear area adds discriminative value	Single cohort; industry-sponsored; threshold generalisability across labs unclear

Advantages:

- Well-understood architectures with readily available public pretrained weights;
- High cell-level accuracy (> 0.90) achieved on curated diagnostic datasets;
- Modest computational requirements during both training and inference phases;
- Cross-organ transfer learning capabilities that significantly reduce the downstream manual labelling burden.

Drawbacks:

- Limited to isolated cell-level decisions, making whole-slide-level aggregation largely ad-hoc;
- Highly fragile performance when subjected to variations in laboratory staining protocols and scanner hardware;
- Inherent inability to model critical spatial relationships and structural topologies between neighboring cells or clusters;
- Uninterpretable "black-box" predictions that severely limit utility and trust in deployment-phase clinical settings.

2.4.2 MIL/WSI Based Approaches

Weakly supervised learning frameworks based on MIL have become central in whole-slide imaging (WSI) analysis for urological pathology, where exhaustive pixel-level annotation is infeasible. In bladder cancer and urine cytology, MIL models treat each slide as

a bag of instances (cells or patches), learning discriminative signals from slide-level labels such as benign, atypical, or malignant.

Attention-based MIL variants further improve interpretability by assigning importance scores to diagnostically relevant regions, often highlighting rare malignant clusters within highly imbalanced cellular distributions. These methods are particularly effective in large-scale clinical datasets, where limited annotation granularity requires scalable learning while preserving diagnostic performance at the slide level. A typical MIL/WSI-based approach pipeline contains:

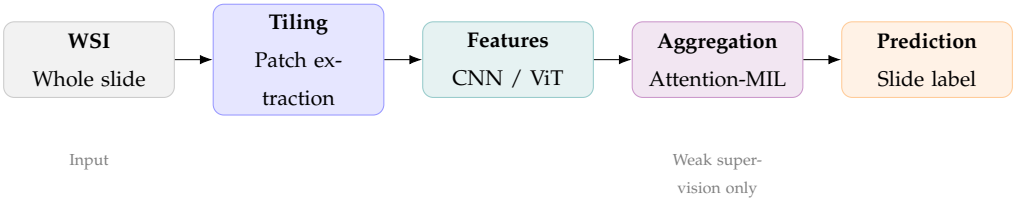


FIGURE 2.3: Typical MIL/WSI-based pipeline for urine cytology analysis. Whole slides are tiled into patches, features are extracted per patch using a pretrained encoder, and a learnable aggregation module produces a slide-level prediction from slide-level labels alone.

Tsuneki et al. (2022) investigates deep learning for binary neoplastic/non-neoplastic classification of urine LBC WSIs, comparing fully supervised and weakly supervised transfer learning from a uterine-cervix domain using EfficientNet-B1.

Trained on 786 SurePath WSIs from a single Japanese laboratory, the best model achieves ROC-AUCs of 0.984 and 0.990 on equal-balance and clinical-balance test sets, respectively, establishing a strong baseline for WSI-level urine cytology screening.

Wu et al. (2024) presents PUCAS, a three-stage pipeline (patch extraction, ResNet feature extraction, slide-level aggregation) validated prospectively across four Chinese hospitals on 5,376 patients, achieving sensitivity of 0.896–1.000 and NPV of 0.964, and demonstrating a reduction of 57.5% in unnecessary endoscopy referrals.

Shen et al. (2025) extends the PUCAS framework toward muscle-invasive urothelial carcinoma detection, integrating radiology context in a multi-centre prospective study, achieving an AUROC of 0.857, with the multimodal variant raising sensitivity to 0.833 for bladder cancer and 0.903 for upper urinary tract cases versus radiologists alone.

Zhao et al. (2024) proposes LESS, a label-efficient MIL framework combining variational positive-unlabeled learning for the patch feature extraction with a multi-scale cross-attention Transformer aggregator, reducing dependence on slide-level annotations and

achieving an accuracy of 0.848, AUC of 0.854, and sensitivity of 0.918 on a urine cytology WSI dataset of 130 samples.

TABLE 2.4: MIL/WSI Based Approach — MIL and WSI-level systems.

Work	Architecture	Key contribution	Performance	Main limitations
Tsuneki et al. (2022)	EfficientNet-B1; uterine-cervix domain transfer; FS + WS learning	DL screening of urothelial carcinoma in LBC WSIs; fully supervised vs. weakly supervised comparison; 786 train / 750 test slides	AUC 0.984 (equal-balance) / 0.990 (clinical-balance)	Single laboratory (Japan); SurePath only; no multi-site validation
Wu et al. (2024)	3-stage: patch extraction → ResNet features → aggregator	PUCAS: multicentre prospective diagnostic study; 5,376 patients across 4 Chinese hospitals; histopathology as ground truth	Sensitivity 0.896–1.000; NPV 0.964; reduces 57.5% unnecessary endoscopy	China-only cohort; LBC only; WSI scan quality dependency
Shen et al. (2025)	PUCAS extended + radiology integration (PUCAS-M)	First AI model for non-invasive detection of muscle-invasive UC from urine cytology; integrates imaging context	AUROC 0.857; multimodal variant raises sensitivity to 0.833 (BC) and 0.903 (UTUC) vs. radiologists alone	Prospective but single-country; radiology integration preliminary
Zhao et al. (2024)	Multi-scale MIL with VPU learning + CrossViT aggregator	Label-efficient WSI screening; variational positive-unlabeled learning uncovers hidden patch labels from slide-level supervision only	Accuracy 0.848; AUC 0.854; sensitivity 0.918 on 130 urine WSIs	Small urine cohort (130 slides); mixed cytology corpus; urine-specific generalisation unclear

Advantages:

- Enacts seamless end-to-end slide-level classification without requiring localized single-cell annotations;
- Enables weakly supervised training driven entirely by coarse, slide-level diagnosis labels alone;
- Learned attention maps offer partial clinical interpretability by highlighting diagnostic regions of interest;
- Demonstrates robust performance with an exceptional ROC-AUC range of 0.984–0.990 on large clinical validation cohorts.

Drawbacks:

- Imposes exceptionally large compute memory footprint and storage requirements for processing whole-slide image tile matrices;

- Standard MIL tile-independence assumptions frequently fail when modelling overlapping or cohesive cytological clusters;
- Diagnostic performance noticeably degrades when evaluating low-cellularity slides or specimens obscured by heavy background artefacts;
- Severe scanner and laboratory staining heterogeneity limits cross-site out-of-distribution generalisation.

2.4.3 Semi-Automated Scoring Systems

Semi-autonomous tools quantify The Paris System (TPS)-defined morphological criteria such as N/C ratio, chromatin texture, nuclear membrane irregularity (circularity, solidity), and cluster atypia burden, presenting them as structured scores to a human-in-the-loop cytopathologist. These systems occupy a distinct niche between fully automated classification and manual reporting, combining computational objectivity with pathologist oversight.

AutoParis and its successor AutoParis-X are the most mature open-source tools in this category, while the AIxURO platform represents a parallel commercial semi-autonomous approach. Both lineages are grounded in TPS criteria and target the reduction of inter-observer variability, particularly for the diagnostically ambiguous atypical urothelial cell category. A typical Semi-Automated Scoring pipeline contains:

1. **Instance segmentation** - QuPath scripts or deep learning (HoVer-Net, CellViT/SAM-based) to isolate nuclei and cytoplasm.
2. **Morphometric feature extraction** - N/C ratio; nuclear area, circularity, solidity, perimeter; chromatin intensity and coarseness; cluster-level size and arrangement.
3. **Atypia Burden Score (ABS)** - logistic regression or gradient-boosted trees mapping morphometric features to TPS categories.
4. **Longitudinal module** - ABS time-series across surveillance visits for recurrence risk modelling.
5. **Interactive viewer** - open-source web interface for pathologist review and case-level override.

The AutoParis-X series by Levy et al. (2023a) represents the most mature open-source semi-autonomous scoring system for urine cytology.

Levy et al. (2023a) presents a large-scale retrospective validation of AutoParis-X on 1,252 ThinPrep specimens, using a pipeline of UroNet (cell detection), UroSeg (U-Net

nucleus/cytoplasm segmentation), and AtyNet (ResNet-18 atypia scoring) to produce a slide-level Atypia Burden Score (ABS), demonstrating close alignment with TPS categories and introducing a cluster border modelling via BorderDet to improve the N/C ratio accuracy in overlapping clusters.

Levy et al. (2023b) extends this framework longitudinally, applying AutoParis-X across surveillance visits in a cohort of 159 patients (1,259 slides) to predict bladder cancer recurrence, achieving a C-index of 0.77 and outperforming standard cytological assessment, with predictive performance strongest in specimens collected within the first six months post-diagnosis.

Levy et al. (2023c) further uncovers additional predictors from voided urothelial cell clusters by applying a deep learning-based cluster separator prior to morphometric extraction, identifying cell border overlap, cluster smoothness, and total cluster count as markers significantly associated with specimen atypia and urothelial carcinoma diagnosis.

Liu et al. (2024) evaluates the commercial AIxURO platform on 116 urine cytology WSIs reviewed by one cytopathologist and two cytotechnologists, demonstrating that AIxURO increases the AUC-category sensitivity from 0.250–0.306 to 0.639 and NPV from 0.913–0.916 to 0.953, while reducing screening time by 52.3 to 83.2% relative to microscopy.

Chang et al. (2024) apply AIxURO to 185 upper urinary tract cytology specimens, identifying a 20% discrepancy rate with conventional reporting and demonstrating improved diagnostic confidence for atypical cases in a single Taiwanese centre.

Chen et al. (2025) assesses AIxURO for post-nephroureterectomy UTUC surveillance across 296 slides from 113 patients, reporting earlier detection of intravesical recurrence and a reduction in the AUC designation rate relative to conventional cytology reporting.

Hoshino et al. (2025) propose a computer-assisted scatter-plot method plotting cell and cluster area against the maximum nuclear area across 192 cases (52 negative, 140 positive), finding significant differences between negative and non-invasive groups ($p = 0.028$) using area features alone without deep learning.

Xiong et al. (2026) develops an interpretable morphometric framework trained on 10,230 expert-annotated urothelial cells, extracting 20 quantitative features aligned with Paris System criteria and training ordinal logistic regression and random forest classifiers that achieve $>$0.90 accuracy for three-class cell categorisation (normal/atypical / suspicious), with slide-level risk scores validated on 247 cases, showing effective stratification

across all TPS groups ($p < 0.0001$).

TABLE 2.5: Semi-Automated Scoring Systems — quantitative scoring tools.

Work	Scoring model	Key contribution	Performance	Main limitations
Levy et al. (2023a)	UroNet + UroSeg (U-Net) + AtyNet (ResNet-18); slide-level ABS	Large-scale validation of AutoParis-X; BorderDet improves N/C accuracy in overlapping clusters; open-source interactive web viewer	ABS closely predicts TPS categories; cluster-border modelling improves N/C accuracy	Retrospective; single institution (Dartmouth); ThinPrep only
Levy et al. (2023b)	AutoParis-X longitudinal ABS time-series	ABS extracted across surveillance visits to predict bladder cancer recurrence; 159 patients, 1,259 slides	C-index 0.77; outperforms standard cytological assessment; strongest in first 6 months post-diagnosis	Observational; single institution; unmeasured confounders; incomplete treatment data
Levy et al. (2023c)	DL cluster separator + morphometric ML	Deep learning cluster separator uncovers additional predictors from voided urothelial cell clusters; identifies border overlap, smoothness and total cluster count as atypia markers	Cluster-level features significantly associated with UC diagnosis	Single institution; unadjusted for specimen preparation variability
Liu et al. (2024)	AIxURO instance segmentation + TPS quantification	Evaluation of AIxURO on 116 WSIs; sensitivity and NPV increase for AUC+ cases; screening time reduced by 52 to 83%	AUC sensitivity: 0.250–0.306 $\rightarrow$ 0.639; NPV: 0.913–0.916 $\rightarrow$ 0.953	Industry tool; US-only cohort; potential conflict of interest
Chang et al. (2024)	AIxURO (upper-tract specimens)	AIxURO applied to 185 upper urinary tract cytology samples; 20% discrepancy rate with conventional reporting	Improved diagnostic confidence for atypical cases; identifies under-diagnosed upper-tract cases	Small cohort; single centre (Taiwan); no comparative performance metric
Chen et al. (2025)	AIxURO + TPS quantitative output	AIxURO for post-nephroureterectomy UTUC surveillance; 296 slides, 113 patients; earlier intravesical recurrence detection	Lower AUC designation rate; earlier recurrence detection vs. conventional cytology	Retrospective; single country; no randomised comparison
Hoshino et al. (2025)	Computer-assisted scatter-plot (cell/cluster area + max nuclear area)	Scatter-plot of cell area vs. maximum nuclear area distinguishes UC from benign; 192 cases (52 negative / 140 positive)	Significant difference between negative and non-invasive groups ($p = 0.028$)	Manual scatter-plot method; limited to area features; no deep learning

Table 2.5 continued...

Work	Scoring model	Key contribution	Performance	Main limitations
Xiong et al. (2026)	Ordinal logistic regression + random forest on 20 morphometric features	Interpretable morphometric scoring of 10,230 annotated urothelial cells; slide-level risk scores from aggregated cell probabilities; TPS-aligned	>0.90 accuracy (three-class); risk scores stratify all TPS groups ($p < 0.0001$) in 247-case validation	External cross-country validation pending; no deep feature extraction

Advantages:

- Provides highly interpretable, TPS-grounded metrics directly aligned with standardized clinical guidelines;
- Significantly reduces intra- and inter-observer diagnostic variability when evaluating borderline AUC and SHGUC cases;
- Readily available open-source implementations lower adoption barriers for low-resource clinical laboratories;
- Longitudinal tracking capabilities enable precise patient monitoring and recurrence risk stratification over time;
- A computationally lightweight architecture enables rapid inference, completing slide evaluation in seconds.

Drawbacks:

- Relies heavily on flawless upstream cellular segmentation, meaning that complex, overlapping cell clusters remain a major operational challenge;
- System calibration drifts noticeably when processing varying specimen preparation methods (e.g., ThinPrep versus SurePath liquid-based cytology);
- Functions as an isolated vision parser that does not incorporate crucial clinical covariates or longitudinal molecular data;
- Initial boundary and threshold settings require extensive, institution-specific cross-validation prior to clinical deployment.

2.4.4 Foundation/SSL-Based Approaches

Scarcity of labelled cytological data and the domain gap from natural images have motivated data synthesis, self-supervised pretraining, and large-scale foundation models in

computational pathology. In bladder cancer and urine cytology, self-supervised pretraining on large unlabelled histology or cytology corpora enables the extraction of generalisable morphological representations.

Methods such as contrastive learning and masked image modelling allow models to capture fine-grained cellular structures, including nuclear contours and texture distributions, without manual labels. These pretrained encoders can then be fine-tuned for downstream tasks such as cancer detection, grading, and subtype classification. This paradigm is particularly relevant in clinical domains where annotation cost is high and inter-observer variability is significant. A typical Foundation/SSL-based approach pipeline contains:

1. **Generative models** - StyleGAN3 for realistic cell-image synthesis; conditional GANs for class-conditioned augmentation (HGUC vs. negative).
2. **Self-supervised backbone** - SimCLR, MoCo, DINO, iBOT, or MAE pretrained on unlabelled cytology or histopathology patches; fine-tuned on small labelled subsets.
3. **Foundation models** - large ViT (ViT-B/L/H) pretrained on hundreds of thousands of pathology images; zero-shot or few-shot transfer to urine cytology downstream tasks.
4. **Downstream head** - linear probe, *k*-NN, or light MIL aggregator on the frozen or fine-tuned encoder.

McAlpine et al. (2022) train a StyleGAN3 model on 1,000 malignant urothelial cytology images to generate novel synthetic examples of malignant urine cytology, demonstrating that realistic and morphologically diverse synthetic images can be produced using computational resources within reach of most pathology departments, with the stated goal of augmenting cytology training in settings where real-world examples of urothelial neoplasia are rare.

McAlpine et al. (2023) follow this with a formal comparative assessment in which 12 pathology personnel evaluated a balanced 60-image survey of real and conditional StyleGAN3-generated synthetic urine cytology images, finding no significant difference in diagnostic error rates or subjective image quality scores between real and synthetic images, validating the perceptual fidelity of GAN-generated cytology for training purposes.

Ciga et al. (2022) pretrain a SimCLR self-supervised contrastive learning model on 60 unlabelled histopathology datasets and demonstrating that the resulting features outperform ImageNet-pretrained initialisations on a range of downstream pathology tasks, including cytological classification, establishing the value of domain-specific self-supervised pretraining over general-purpose encoders.

Hörst et al. (2024) introduce CellViT, a vision transformer nuclear instance segmentation model using a SAM-H encoder and UNETR-style decoder trained on PanNuke (approximately 200,000 annotated nuclei across 19 tissue types), achieving state-of-the-art segmentation performance with strong out-of-domain generalisation applicable to cytological specimens, though cytology-specific fine-tuning is still required.

Ivezić et al. (2025) introduces CytoFM, the first cytology-specific self-supervised foundation model, pretraining a ViT-Base using the iBOT framework (combining masked image modelling and self-distillation) on a diverse collection of cytology datasets, and demonstrating that CytoFM outperforms both histopathology-pretrained (UNI) and natural image-pretrained (iBOT-ImageNet) models on two out of three downstream cytology tasks despite a relatively small pretraining corpus.

Alfasly et al. (2024) critically evaluate general-purpose vision-language foundation models (PLIP, BiomedCLIP) for histopathological classification across eight datasets, including four internal Mayo Clinic cohorts, finding that domain-specific non-foundational models such as DinoSSLPath and KimiaNet consistently outperform general-purpose foundation models, and concluding that high-quality curated training data with validated image-text alignment remains essential for clinical-grade performance.

Al-Asi et al. (2026) provides a comprehensive review of pathology foundation model evolution covering UNI, Virchow, CONCH, GigaPath and beyond, explicitly identifying cytopathology as an underserved domain in current pretraining corpora and outlining the technical and clinical challenges that must be addressed for foundation models to achieve broad cytological applicability.

TABLE 2.6: Foundation/SSL Based Approaches — generative, self-supervised, and foundation-model methods.

Work	Model type	Key contribution	Performance	Main limitations
McAlpine et al. (2022)	StyleGAN3	StyleGAN3 trained on 1,000 malignant urothelial cytology images; generates morphologically diverse synthetic malignant cell images for education and training augmentation	Realistic synthetic images produced with a modest dataset and standard compute	Single morphological class (malignant only); no downstream classifier training experiment
McAlpine et al. (2023)	StyleGAN3 (evaluation study)	Comparative assessment of real vs. synthetic urine cytology images; 60-image survey completed by 12 pathology personnel	No significant difference in diagnostic error rates or subjective quality between real and synthetic images	Perception study only; no downstream ML training experiment
Ciga et al. (2022)	SimCLR self-supervised contrastive learning	Self-supervised pretraining on 60 unlabelled histopathology datasets; features outperform ImageNet initialisation on downstream cytological and pathology tasks	Consistent improvement over ImageNet-pretrained models across multiple downstream tasks	Histology-focused; urine cytology is one sub-domain among many
Hörst et al. (2024)	CellViT: SAM-H ViT encoder + UNETR-style decoder	Vision-transformer nuclear instance segmentation trained on PanNuke (≈200,000 annotated nuclei, 19 tissue types); strong out-of-domain generalisation to cytological specimens	State-of-the-art on PanNuke; strong generalisation across tissue types	Trained primarily on H&E histology; cytology fine-tuning needed
Ivezić et al. (2025)	CytoFM: iBOT ViT-Base (masked image modelling + self-distillation)	First cytology-specific self-supervised foundation model; pretrained on diverse cytology datasets; evaluated on breast cancer classification and cell type identification	Outperforms UNI and iBOT-ImageNet on 2 out of 3 downstream cytology tasks	Preprint; small pretraining corpus; not evaluated on urine cytology specifically
Alfasly et al. (2024)	Evaluation: PLIP, Biomed-CLIP vs. DinoSSLPath, KimiaNet	Critical evaluation of foundation models for histopathology across 8 datasets (4 internal Mayo Clinic); domain-specific models consistently outperform general VLMs	DinoSSLPath: MV@5 F1 0.63 (colorectal), 0.74 (liver); KimiaNet: 0.56 (breast), 0.70 (skin)	Histopathology focus; no cytology datasets; no urine specimens

Table 2.6 continued...

Work	Model type	Key contribution	Performance	Main limitations
Al-Asi et al. (2026)	Survey: UNI, Virchow, CONCH, GigaPath and beyond	Comprehensive review of pathology FM evolution; cytopathology explicitly identified as an underserved domain in the current pretraining corpora	Documents SOTA benchmark performance across 20+ diagnostic and prognostic tasks	Review paper; no new experimental data

Advantages:

- Generates synthetic images to augment rare morphological variants and support privacy-preserving data sharing across institutions;
- Self-supervised frameworks learn rich, robust pixel representations without requiring expensive, expert-level visual annotations;
- Large-scale foundation models transfer seamlessly across diverse tissue types with minimal target-domain fine-tuning;
- Cytology-specific pretraining drastically reduces the inherent domain gap left by traditional histopathology-centric encoders.

Drawbacks:

- Generative adversarial network (GAN) architectures may encode subtle visual hallucinations undetectable by human pathologists;
- Large-scale foundation models require massive, specialised computational infrastructure for the initial pretraining phases;
- Cytopathology structures remain significantly underrepresented within the current public pathology foundation model corpora;
- The formal regulatory status and validation pipelines of synthetic-data generation systems remain largely unresolved.

2.4.5 Vision-Language Models

Multimodal and vision-language models represent the most recent shift in urological AI, integrating visual histopathology or cytology data with natural language descriptions. In bladder cancer workflows, these models enable joint reasoning over microscopic images and clinical narratives, supporting tasks such as diagnosis explanation, report generation, and structured staging prediction.

By aligning visual embeddings with textual semantics, VLMs improve interpretability and enable zero-shot or few-shot generalisation across clinical settings. Recent developments further explore conversational pathology systems, where models assist clinicians by answering questions about cellular morphology and diagnostic criteria, marking a transition toward interactive AI-assisted urological diagnosis. A typical Vision-Language Model pipeline contains:

1. **Image encoder** - ViT or CNN backbone pretrained on large pathology corpora; outputs patch or slide-level visual embeddings.
2. **Text encoder** - transformer-based language model (BERT, GPT-style) encoding natural language descriptions, reports, or diagnostic prompts.
3. **Alignment module** - contrastive learning (CLIP-style), CoCa, or the BLIP-2 framework, aligns image and text embedding spaces.
4. **Downstream interface** - zero-shot or few-shot classification; text-to-image or image-to-text retrieval; visual question answering; report generation.

Huang et al. (2023) introduces PLIP, the first pathology-specific VLM, fine-tuned from Contrastive Language-Image Pre-training (CLIP) on OpenPath, a dataset of 208,414 pathology image-text pairs sourced from medical Twitter, achieving zero-shot classification F1 scores of 0.565 to 0.832 across four external datasets and enabling image and text-based case retrieval for knowledge sharing.

Ikezogwo et al. (2023) curate Quilt-1M, the largest histopathology vision-language dataset to date, comprising one million image-text pairs sourced from educational YouTube videos, research papers, and Twitter, and use it to train QuiltNet, a CLIP-based model that outperforms state-of-the-art methods on both zero-shot and linear probing tasks across 13 patch-level datasets spanning eight sub-pathologies.

Lu et al. (2024) presents CONCH, a CoCa-based vision-language foundation model for computational pathology, pretrained on over 1.17 million histopathology image-caption pairs, evaluated on a suite of 14 diverse benchmarks and achieving state-of-the-art performance on histology images classification, segmentation, captioning, and both text-to-image and image-to-text retrieval, substantially outperforming concurrent systems.

Ahmed et al. (2024) develop PathAlign, a BLIP-2-based VLM for whole slide images trained on over 350,000 de-identified WSI and diagnostic text pairs spanning diverse diagnoses, procedure types, and tissue types, enabling WSI-level report generation and

text-based case retrieval; pathologists rated model-generated reports as accurate without clinically significant error or omission for 78% of WSIs on average.

Dai et al. (2025) proposes PathologyVLM, a large vision-language model for pathology image understanding using a two-stage training strategy combining domain-specific alignment and visual question answering fine-tuning, with a PLIP-based visual encoder and a scale-invariant connector to preserve spatial information across image resolutions, achieving best overall performance among models of comparable scale on both supervised and zero-shot VQA benchmarks, including PathVQA and PMC-VQA.

(Zhang et al., 2023) presents BiomedCLIP, pretrained on approximately 15 million biomedical image-text pairs from PubMed Central spanning diverse imaging modalities, serves as a recurring point of reference across comparative evaluations in the field, representing the broader biomedical pretraining paradigm beyond pathology-specific corpora.

Recent conceptual reviews have begun to explore the potential of VLM-based copilots alongside immersive technologies in cytopathology workflows, identifying applications in the report generation, workflow support, and diagnostic decision guidance (Giansanti, 2026). Crucially, these reviews confirm that VLM application to cytopathology remains largely conceptual and preclinical, with database searches across PubMed, Web of Science and Scopus returned a limited number of studies addressing AI-assisted cytopathology imaging — directly reinforcing the research gap motivating this thesis.

TABLE 2.7: Vision-Language Models — pathology image-text alignment and multimodal reasoning.

Work	Model type	Key contribution	Performance	Main limitations
Huang et al. (2023)	PLIP: CLIP fine-tuned on OpenPath (208,414 Twitter image-text pairs)	First pathology-specific VLM; zero-shot classification and image/text retrieval across diverse pathology tasks	Zero-shot F1 0.565 to 0.832 across 4 external datasets; outperforms prior CLIP models (F1 0.030 to 0.481)	Histopathology only; Twitter-sourced data may have quality variance; no cytology evaluation
Ikezogwo et al. (2023)	QuiltNet: CLIP fine-tuned on Quilt-1M (1M YouTube + Twitter + paper image-text pairs)	Largest histopathology vision-language dataset; zero-shot and linear probing across 13 patch-level datasets spanning 8 subpathologies	Outperforms SOTA on zero-shot and linear probing across all 13 datasets	No cytology datasets; patch-level only; no WSI-level evaluation

Table 2.7 continued...

Work	Model type	Key contribution	Performance	Main limitations
Lu et al. (2024)	CONCH: CoCa-based VLM pretrained on 1.17M histopathology image-caption pairs	State-of-the-art on 14 benchmarks; classification, segmentation, captioning, and retrieval;	SOTA across 14 diverse computational pathology benchmarks	Histopathology-centred pretraining; cytology underrepresented
Ahmed et al. (2024)	PathAlign: BLIP-2 framework on 350,000+ WSI-text pairs	WSI-level VLM enabling report generation and text-based case retrieval; diverse diagnoses and tissue types	78% of generated reports rated accurate by pathologists; strong retrieval performance	Preprint; no cytology specimens; report accuracy varies by tissue type
Dai et al. (2025)	PathologyVLM: PLIP encoder + scale-invariant connector; two-stage training (alignment + VQA fine-tuning)	Domain-specific large VLM for pathology VQA; preserves spatial information across resolutions via a scale-invariant connector	Best overall on PathVQA and PMC-VQA among comparable-scale models; strong zero-shot VQA performance	No cytology evaluation; VQA focus limits its direct diagnostic applicability
Zhang et al. (2023)	BiomedCLIP: CLIP pretrained on PMC-15M (15M figure-caption pairs from 4.4M PubMed Central articles)	Broad biomedical vision-language pretraining spanning diverse image types including radiology, pathology, and microscopy; recurring reference baseline in comparative evaluations	SOTA across biomedical retrieval, classification, and VQA benchmarks	Scientific figure-caption pairs lack clinical annotation depth; not pathology-specific
Giansanti (2026)	Conceptual narrative review on VLMs and immersive tools in cytopathology	First review explicitly addressing VLM-based copilots in cytopathology; identifies report generation, workflow support, and decision guidance as key applications; confirms VLM application to cytopathology remains largely conceptual and preclinical	Limited empirical studies on AI-assisted cytopathology imaging were confirmed across PubMed, Web of Science, and Scopus	Conceptual only; no experimental validation; not specific to urine cytology or classification tasks

Advantages:

- Robust zero-shot and few-shot generalisation capabilities drastically reduce downstream visual annotation requirements;
- Native natural language interfaces significantly improve clinical interpretability and trust for deployment pathologists;
- Cross-modal retrieval capabilities facilitate seamless case-matching and knowledge sharing across distinct medical institutions;
- Automated report generation pipelines natively support highly structured and efficient clinical documentation workflows.

Drawbacks:

- Public pretraining data remains overwhelmingly histopathology-focused, leaving cytopathology highly underrepresented;
- Large-scale medical image-text pair curation remains exceptionally expensive, labour-intensive, and domain-specific;
- General-purpose VLMs consistently underperform domain-specific supervised models on highly specialised diagnostic tasks (Alfasly et al., 2024);
- Inherent text generation hallucination risks in automated clinical reports demand rigorous, expert-level human validation.

2.5 Summary and Research Gaps

The literature reviewed in this chapter establishes the clinical and technical foundations motivating this thesis. Bladder cancer remains a significant clinical challenge, with current diagnostic workflows affected by interobserver variability and limited sensitivity, particularly in early-stage and low-grade disease (Tosoni et al., 2000; Mowatt et al., 2011). Urine cytology, while non-invasive and widely used for surveillance, is constrained by subjective morphological assessment and high inter-observer disagreement, especially for the diagnostically ambiguous atypical urothelial cell category under the Paris System (Babjuk et al., 2022).

Computational approaches have progressively attempted to address these limitations, yet each methodological generation has introduced new constraints alongside its advances. Supervised CNN and ViT-based methods demonstrated that deep learning can reliably capture nuclear atypia and morphological features relevant to malignancy detection (Kaneko et al., 2022; Liu et al., 2022; Yang et al., 2025), but their dependence on

densely annotated datasets and fixed label vocabularies limits scalability across clinical settings.

MIL-and WSI-based frameworks reduced this annotation burden by learning from slide-level labels alone (Wu et al., 2024; Zhao et al., 2024), yet introduced sensitivity to scan quality and staining variability, which constrain cross-institutional deployment. Semi-automated scoring systems such as AutoParis-X (Levy et al., 2023a,b) and AIxURO (Liu et al., 2024; Chang et al., 2024) brought quantitative TPS-grounded outputs closest to clinical deployment, but remain dependent on accurate upstream segmentation and institution-specific calibration. More recently, self-supervised and foundation model approaches have begun to reduce reliance on expert annotation through large-scale pretraining (Ciga et al., 2022; Hörst et al., 2024), though cytopathology remains substantially underrepresented in existing pretraining corpora (Al-Asi et al., 2026). Vision-language models such as PLIP (Huang et al., 2023), and BiomedCLIP (Zhang et al., 2023) extend this further by enabling zero-shot and few-shot transfer through image-text alignment, yet have been evaluated almost exclusively on histopathology benchmarks, leaving their applicability to cytological imaging largely unexamined.

Despite this progress, the literature points to several areas that remain underexplored. Systematic comparison of vision-language model adaptation strategies, spanning frozen feature extraction, partial fine-tuning, prompt learning, and mixture-of-experts architectures, has not been conducted within a single controlled framework on cytological data, making it difficult to assess the relative contribution of each component. Relatedly, whether domain-specific biomedical pretraining offers meaningful advantages over general vision-language pretraining for fine-grained cytological classification remains poorly established, as existing comparisons are largely confined to histopathology benchmarks.

Beyond adaptation strategies, CLIP-based models have not been applied to the morphological discrimination required for TPS-aligned staging, leaving open the question of whether contrastive image-text alignment can support structured cytological classification beyond binary malignancy detection. Evaluation of bladder cancer cell line microscopy also remains largely absent from the literature, despite its value as a controlled setting for pipeline development and experimental validation. A recent conceptual review explicitly confirmed that the application of VLMs to cytopathology remains largely preclinical, with systematic searches across major databases yielding only a limited number of empirical studies on VLM-assisted cytopathological imaging (Giansanti, 2026).

This observation reflects not a failure of the approach but an absence of exploration, underscoring the novelty of applying vision-language model adaptation to cytological classification in a controlled experimental framework. Collectively, these gaps motivate a principled investigation of vision-language model adaptation for cytological bladder cancer analysis.

The following table goes over all the papers mentioned in this literature review.

TABLE 2.8: Condensed summary of works mentioned in this literature review.

Authors/Year	Method / Contributions	Relevance / Findings
Kamat et al. (2016)	Clinical review of bladder cancer	Epidemiology and disease burden
Bray et al. (2024)	Global cancer statistics	Incidence and staging context
Babjuk et al. (2022)	EAU clinical guidelines	TNM staging and treatment protocols
Mowatt et al. (2011)	Systematic review	Diagnostic sensitivity limitations of WLC
Al-Sattar et al. (2026)	Multimodal AI analysis	Interobserver variability in staging
Tosoni et al. (2000)	Pathology observer study	Grading disagreement in histopathology
Ferro et al. (2023)	AI survey	Deep learning for bladder cancer diagnosis
Ma et al. (2024)	AI review	Computational diagnosis and treatment
Awan et al. (2021)	Deep learning cytology	Risk stratification from urine cytology
Nitta et al. (2026)	Autonomous cytopathology pipeline	Clinical-grade staining-free cytology automation
Cancer Research UK (2026)	Clinical reference (CRUK)	Urothelial carcinoma prevalence statistic
Litjens et al. (2017)	Medical imaging DL survey	Data scarcity and imbalance challenges
Gessert et al. (2020)	EfficientNet ensemble	Strong CNN baseline under imbalance
Özcan and Öztürk (2024)	CNN WBC classification	High accuracy on Raabin-WBC dataset
Raghu et al. (2019)	Transfer learning study	Limited ImageNet benefit in medical domain
Tiu et al. (2022)	CheXzero VLM	Zero-shot transfer in medical imaging
Kaneko et al. (2022)	CNN/ViT cell classification	Urothelial cell recognition
Liu et al. (2022)	Cyto-histo models	Cytology–histopathology prediction
Yang et al. (2025)	N/C ratio estimation	TPS threshold recalibration
Tsuneki et al. (2022)	MIL WSI screening	Bladder cancer detection
Wu et al. (2024)	PUCAS study	Multicentre validation

Authors/Year	Method / Contributions	Relevance / Findings
Shen et al. (2025)	Radiology fusion MIL	Muscle-invasive UC detection
Zhao et al. (2024)	Label-efficient MIL	Weakly supervised WSI learning
Levy et al. (2023a)	AutoParis-X	ABS scoring and clustering
Levy et al. (2023b)	Longitudinal ABS	Recurrence prediction (C-index 0.77)
Levy et al. (2023c)	Cluster predictors	Morphological atypia modelling
Liu et al. (2024)	AIxURO system	Faster screening, higher sensitivity
Xiong et al. (2026)	Morphometric ML	Interpretable risk scoring
Chang et al. (2024)	AIxURO upper tract	Single-centre upper urinary tract evaluation
Chen et al. (2025)	AIxURO UTUC surveillance	Post-nephroureterectomy recurrence detection
Hoshino et al. (2025)	Scatter-plot morphometry	Area-based UC discrimination without DL
McAlpine et al. (2022)	StyleGAN3 synthesis	Synthetic cytology generation
McAlpine et al. (2023)	GAN evaluation	Real vs synthetic indistinguishable
Ciga et al. (2022)	SimCLR SSL	Domain SSL > ImageNet
Hörst et al. (2024)	CellViT	ViT-based segmentation
Ivezić et al. (2025)	CytoFM	Cytology foundation model
Al-Asi et al. (2026)	FM survey	Gap in cytopathology corpora
Alfasly et al. (2024)	Foundation model evaluation	Domain-specific models outperform general VLMs
Huang et al. (2023)	PLIP	First pathology VLM
Ikezogwo et al. (2023)	Quilt-1M	1M image-text pairs
Lu et al. (2024)	CONCH	SOTA pathology VLM
Ahmed et al. (2024)	PathAlign	WSI report generation
Dai et al. (2025)	PathologyVLM	Scalable VQA training
Zhang et al. (2023)	BiomedCLIP	Large-scale biomedical VLM
Giansanti (2026)	VLM review in cytopathology	VLM application remains largely conceptual; limited empirical validation on cytological tasks

Chapter 3

Methodology

This chapter outlines the methodology adopted in this project, encompassing the datasets selected for evaluation, the multimodal architecture employed, and the training procedures applied consistently across all experiments.

3.1 Methodology Overview

Our approach focuses on an architecture that integrates pre-trained image-text encoders built on Contrastive Language-Image Pre-training (CLIP)-based vision-language models to develop an effective image classification model. This work is structured into three key stages designed to explore different aspects of the classification problem:

- **Backbone Selection and Baseline Analysis:** The initial phase involves evaluating candidate vision-language backbones through zero-shot classification on the target dataset, identifying the most suitable models for further development.
- **Progressive Fine-Tuning:** The main stage applies a sequential fine-tuning pipeline across seven stages, incrementally introducing architectural and optimisation components to improve classification performance and quantify the contribution of each design decision.
- **Cross-Dataset Evaluation:** The final stage evaluates the generalisation capacity of the pipeline on two external benchmark datasets spanning different imaging modalities, assessing robustness under distribution shift.

3.2 Datasets

This study utilises three datasets for evaluation: CELLo as the primary dataset for training and main performance assessment, and ISIC 2019 Codella et al. (2019) and Raabin-WBC Kouzehkanan et al. (2022) as external datasets for cross-domain generalisation. This setup enables a comprehensive analysis of model robustness under distribution shift and its ability to generalise beyond the primary training domain. Each dataset is detailed in the subsequent sections.

3.2.1 CELLo Dataset Description

The CELLo project aims to investigate a non-invasive, complementary approach to cystoscopy for the early and personalised diagnosis of bladder cancer, based on urine cellular images in a staining-free fashion. Given the disease's high recurrence rate and the limitations of current invasive cystoscopy and urine cytology, CELLo focuses on isolating and concentrating cancer cells from urine. This non-invasive strategy seeks to enable staining-free, precise, and patient-specific diagnostics, advancing the field of personalised cancer care.

In this context, the CELLo dataset provides a high-quality, curated collection of microscopic images of bladder cancer cells, including urine-derived cells from patients and healthy controls, specifically designed to support diagnostic research in this domain. The images were acquired using imaging flow cytometry under brightfield microscopy with standardised acquisition settings to ensure reproducibility. The collected images (Figure 3.1) often show isolated or overlapping cells with circular morphology. Some images may include background variability and debris, accurately reflecting the complexities of real clinical samples.

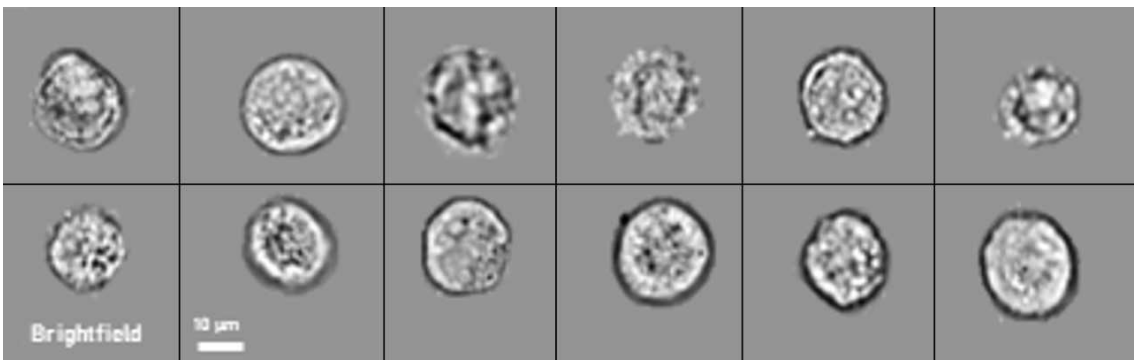


FIGURE 3.1: Representative cell images from the CELLo dataset.

Captured at a total magnification of 60×, the cell images are provided in a greyscale brightfield spectrum. This heterogeneity requires models to effectively handle noisy, low-contrast images while ensuring consistent performance when extracting representative features. High-grade tumours are characterised by highly pleomorphic cells that exhibit aggressive growth and carry a significantly higher risk of invasion and metastasis (Figure 3.2). Patient groups are categorised, encompassing multiple fields of view or time points, enabling patient-level analysis and preventing data leakage during model training.

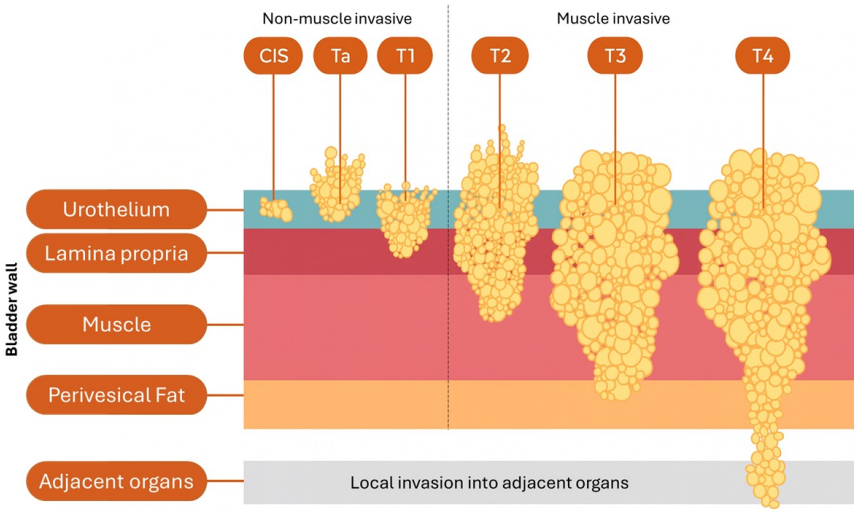


FIGURE 3.2: 2D overview of the classification.

Additionally, the single-cell images are linked to patient-level metadata, including diagnostic labels indicating whether the sample is from a healthy individual or a patient with cancer (i.e., healthy vs cancerous). For cancer-positive samples, details about tumour stage and tumour grade are provided. Tumour stage (T) indicates the extent of tumour invasion and spread through and beyond the bladder wall, with stages ranging from non-muscle invasive (Ta to T1) to muscle-invasive (T2 to T4) (Figure 3.2). T4 is associated with tumours growing from the bladder to surrounding tissues/organs, such as the kidney, prostate or urothelial tract. Tumour grade relates to the appearance of tumour morphology and cancer cells under the microscope, with low-grade (LG) tumours showing cells that resemble normal bladder tissue (epithelium), grow slowly, and have a lower risk of spread.

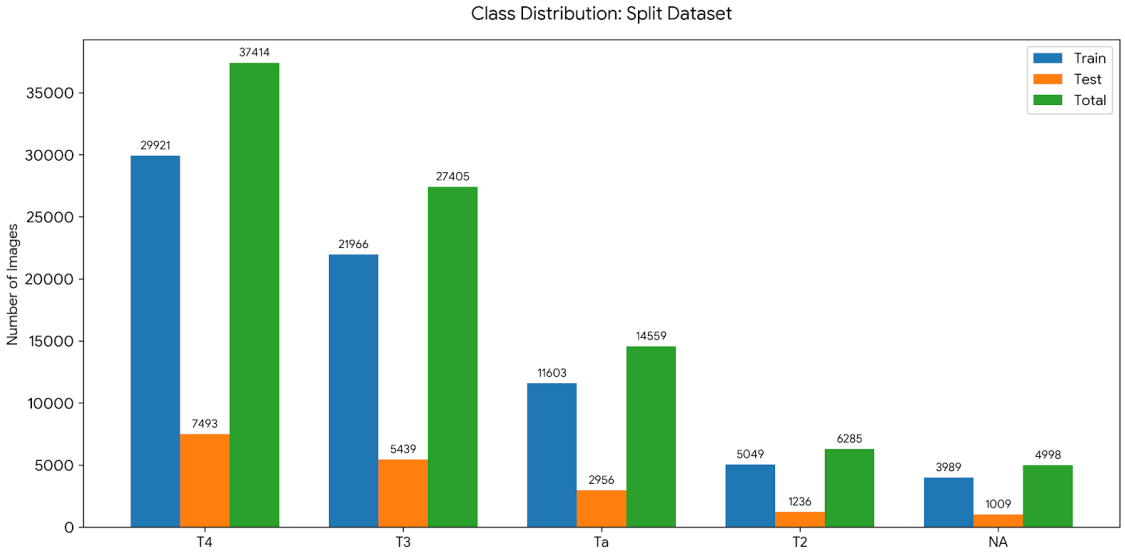


FIGURE 3.3: CELLo class distribution.

High-grade (HG) tumours are associated with morphologically altered bladder epithelial cells, which are associated with advanced cancer. "Na" corresponds to no evidence of disease or simply a normal case sample. The focus of this work is on distinguishing cells by tumour invasiveness stage, avoiding unnecessary complexity, thereby allowing a sufficient cell sample size to train robust models. Cells from CIS (carcinoma in situ (a flat, superficial neoplastic lesion)) tumours were excluded due to their precancerous nature and the difficulty in obtaining them, to avoid biased results.

The SV-HUC-1 uroepithelial non-tumorigenic control cell line was included due to its similarities with normal uroepithelium, but may also exhibit features associated with early-stage carcinoma, requiring careful comparison. Nevertheless, it was included as a control case because this cell line represents an abnormal yet non-tumorigenic phenotype. A normal epithelial cell line from the human bladder was also included.

3.2.2 ISIC 2019 Benchmark Dataset

The ISIC 2019 dataset is a publicly available dermoscopy image dataset released as part of the ISIC 2019 Skin Lesion Analysis challenge (Figure 3.4) (Codella et al., 2019), built upon multiple dermoscopic datasets including HAM10000 (Tschandl et al., 2018) and BCN20000 (Combalia et al., 2019).

The ISIC 2019 dataset was used as an additional benchmark to evaluate the generalisation performance of the proposed pipeline. It consists of 25,331 dermoscopic images across eight skin lesion categories: melanoma, melanocytic nevus, basal cell carcinoma,

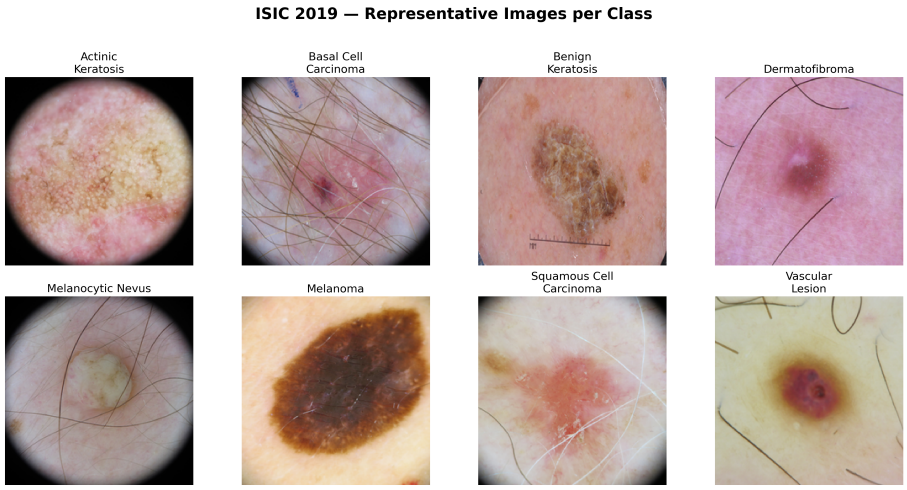


FIGURE 3.4: Sample images from the ISIC 2019 dataset across the eight skin lesion categories.

actinic keratosis, benign keratosis, dermatofibroma, vascular lesion, and squamous cell carcinoma.

The dataset exhibits severe class imbalance (Figure 3.5), with melanocytic nevus as the majority class and rare categories such as dermatofibroma and vascular lesion containing substantially fewer samples.

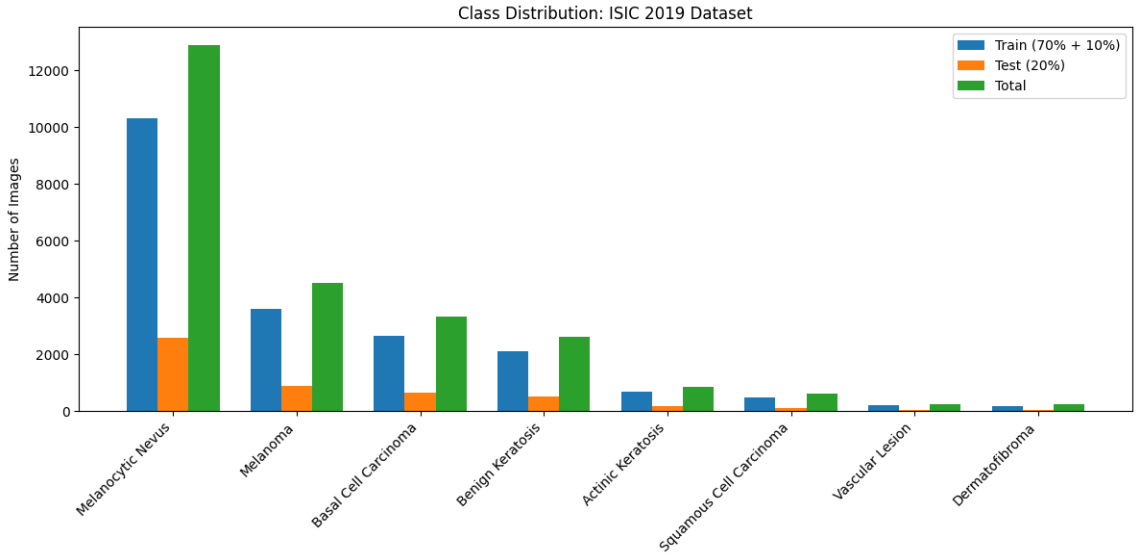


FIGURE 3.5: ISIC 2019 class distribution.

The dataset was used without modification to assess model robustness under realistic clinical conditions. A stratified split of 70%, 10%, and 20% was applied for training, validation, and testing, respectively.

3.2.3 Raabin-WBC: Benchmark Dataset

The Raabin-WBC dataset (Kouzehkanan et al., 2022) is a large-scale white blood cell microscopy dataset comprising peripheral blood images across five cell categories: Neutrophil, Lymphocyte, Eosinophil, Monocyte, and Basophil. The dataset provides an official Train/Test-A split, with the training partition containing 10,175 images and Test-A containing 4,339 images. Figure 3.6 illustrates representative microscopy images from the Raabin-WBC dataset, covering the five major white blood cell categories. The samples highlight the morphological diversity among Neutrophils, Lymphocytes, Eosinophils, Monocytes, and Basophils, which is essential for training and evaluating robust classification models under realistic haematological conditions.

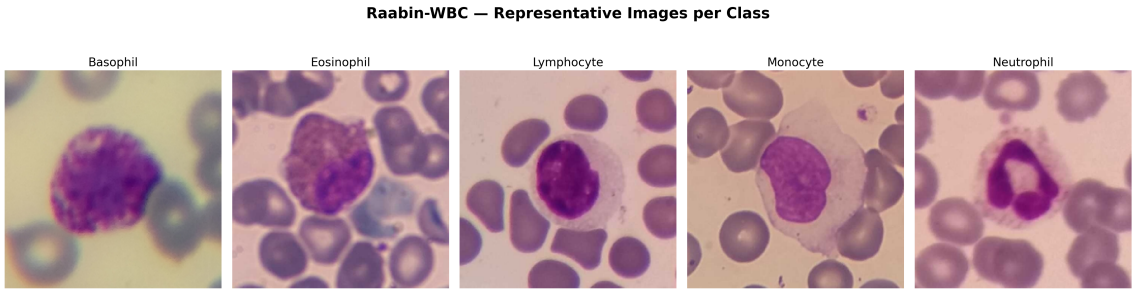


FIGURE 3.6: Representative microscopy images from the Raabin-WBC dataset across the five white blood cell categories.

This official split was respected in this work, with the validation set carved from the training partition, yielding an approximate 60%/10%/30% overall split. The dataset exhibits class imbalance, with Neutrophil representing the dominant class and Basophil the minority, as shown in Figure 3.7.

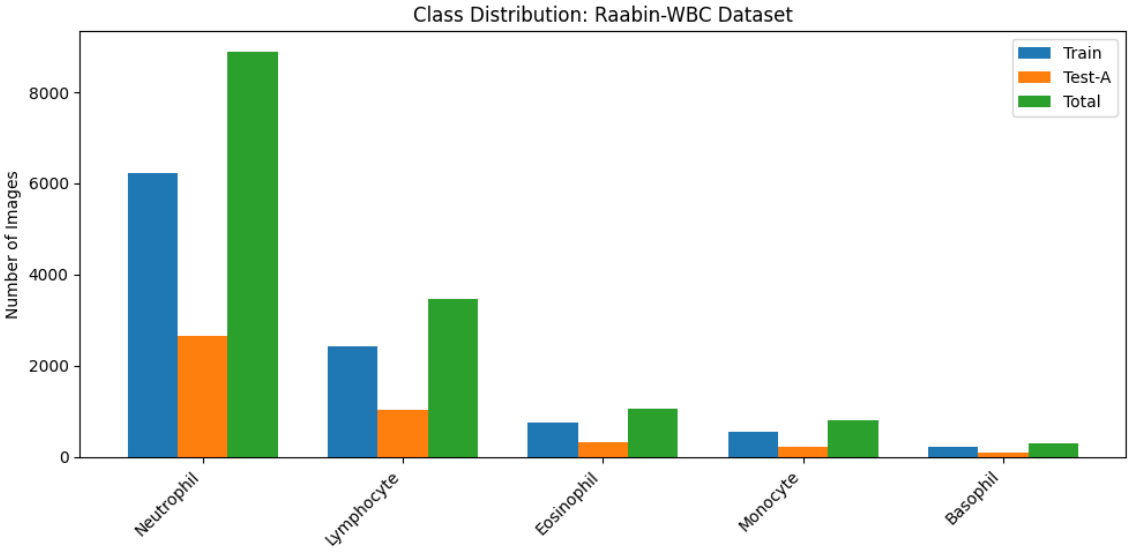


FIGURE 3.7: Raabin-WBC dataset class distribution.

3.3 Multimodal Architecture Definition

Multimodal models are machine learning systems designed to process and relate information from multiple data types, such as text and images. Unlike single-modality models, multimodal models learn shared representations that align different data sources in a common embedding space. In this work, this is achieved through the CLIP framework, in which the model learns to associate images with their textual descriptions while distinguishing them from unrelated pairs.

3.3.1 CLIP

CLIP introduced a paradigm shift in vision-language learning by replacing traditional supervised classification with large-scale contrastive pretraining on image-text pairs (Radford et al., 2021). CLIP employs a dual-encoder architecture consisting of an image encoder f_I and a text encoder f_T, which project visual and textual inputs into a shared embedding space. Training maximises the similarity of matching image-text pairs while minimising the similarity of non-matching pairs within each mini-batch through a symmetric contrastive objective.

For a batch of N image-text pairs, the training loss is defined as:

$$\mathcal{L} = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{\exp(\tau\cdot\mathrm{sim}(f_I(x_i),f_T(t_i)))}{\sum_{j=1}^{N}\exp(\tau\cdot\mathrm{sim}(f_I(x_i),f_T(t_j)))}+\log\frac{\exp(\tau\cdot\mathrm{sim}(f_T(t_i),f_I(x_i)))}{\sum_{j=1}^{N}\exp(\tau\cdot\mathrm{sim}(f_T(t_i),f_I(x_j)))}\right], \tag{3.1}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes the cosine similarity between L2-normalized embeddings and τ is a learnable temperature parameter.

Trained on approximately 400 million image-text pairs collected from the internet, CLIP demonstrated remarkable zero-shot transfer capabilities, enabling image classification through natural language prompts without task-specific retraining. This established contrastive vision-language pretraining as a general-purpose framework for multimodal representation learning (Figure 3.8).

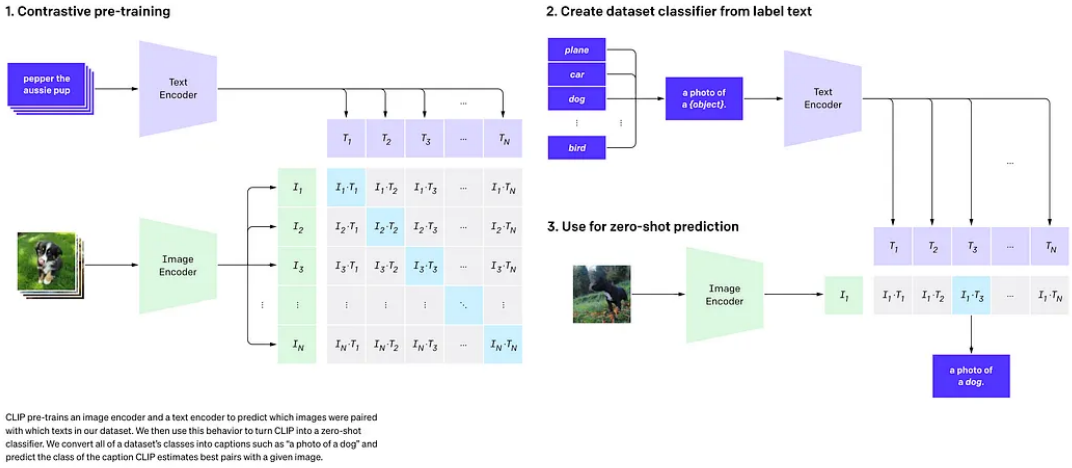


FIGURE 3.8: Overview of the CLIP training objective. An image encoder and text encoder are jointly optimised to align image-text representations in a shared embedding space using contrastive learning. Figure adapted from Radford et al. (2021).

During inference, classification is performed by comparing the image embedding against text embeddings generated from class-specific prompts:

$$p(y=c|x) = \frac{\exp(\tau\cdot\mathrm{sim}(f_I(x),f_T(t_c)))}{\sum_{c'}\exp(\tau\cdot\mathrm{sim}(f_I(x),f_T(t_{c'})))}, \tag{3.2}$$

where t_c denotes the textual description associated with class c. The predicted label corresponds to the prompt with the highest similarity score.

Building upon CLIP, OpenCLIP provided an open-source reproduction trained on publicly available datasets, enabling reproducible research in large-scale vision-language pretraining (Cherti et al., 2023). Scaling studies demonstrated predictable performance

improvements with increasing model and dataset sizes while highlighting the importance of data distribution in downstream performance.

The success of CLIP and OpenCLIP motivated the development of domain-specific variants for medical imaging, where radiological images and clinical reports substantially differ from web-scale image-text data. Consequently, several studies have adapted contrastive vision-language pretraining to medical datasets, learning aligned representations from image-report pairs to support downstream tasks such as disease classification, report generation, image retrieval, and medical visual question answering.

3.3.2 Candidate Backbones

Six vision-language models were evaluated as candidate backbones, spanning general and biomedical pretraining domains. The general domain candidates comprise three OpenCLIP variants (ViT-B/16, ViT-L/14, and ViT-H/14) trained on the LAION dataset (Cherti et al., 2023). The biomedical candidates comprise BiomedCLIP (Zhang et al., 2023), PubMedCLIP (Eslami et al., 2023), and PLIP (Huang et al., 2023).

Among the biomedical candidates, BiomedCLIP introduces several concrete differences from standard CLIP on the text side. The general-domain GPT-2-based text encoder is replaced with PubMedBERT (Gu et al., 2021), a BERT-based encoder-only transformer pretrained on PubMed abstracts and full-text PubMed Central articles. This substitution brings three practical differences: the tokeniser is switched from byte-pair encoding to WordPiece with a biomedical vocabulary, reducing out-of-vocabulary issues with specialised clinical terms; the maximum context length is extended from 77 to 256 tokens, accommodating longer biomedical descriptions, and the text representation shifts from the final `[EOS]` token to the `[CLS]` token, consistent with BERT-style encoding. Together, these changes ground the text encoder in biomedical terminology, with the joint vision-language alignment trained on biomedical image-text pairs, calibrating the shared embedding space for biomedical semantics rather than general web content. A summary of the employed backbones is described in Table 3.1.

TABLE 3.1: Overview of vision-language models used, spanning general and biomedical domain pretraining approaches.

Model	Pretraining Data	Scale	Focus
OpenCLIP	LAION (web-scale image-text pairs)	Very Large	General visual recognition across diverse domains
PubMedCLIP	ROCO dataset (PubMed figure captions)	Moderate	Biomedical domain adaptation of general CLIP
PLIP	∼208k pathology image-text pairs from medical Twitter	Small	Histopathology classification and retrieval
BiomedCLIP	∼15M biomedical image-text pairs from PubMed Central	Large	Broad biomedical coverage across multiple modalities

Following the zero-shot evaluation, the strongest-performing model and BiomedCLIP are taken forward for a preliminary evaluation. BiomedCLIP is included as a fixed domain-specific candidate due to its large-scale pretraining on diverse biomedical image-text pairs spanning multiple imaging modalities, making it the most comprehensive biomedical pretraining among the candidate models reviewed.

3.4 Fine-Tuning Strategies for Vision-Language Models

Adapting pre-trained vision-language models to downstream medical imaging tasks requires careful consideration of how much of the pretrained knowledge to preserve and how much to adapt. A range of strategies has been proposed, spanning from minimal adaptation with frozen backbones to progressive unfreezing, prompt-based tuning and mixture-of-experts approaches.

3.4.1 Feature Extraction and Linear Probing

The simplest form of adaptation involves keeping the entire backbone frozen and training only a lightweight classification head on top of the extracted features. This approach, known as linear probing, preserves the full pretrained vision-language alignment while adapting the output to the target task.

Preserving pretrained features through frozen backbones can be particularly advantageous under a distribution shift, where full fine-tuning risks degrading generalisation (Kumar et al., 2022). In the medical domain, keeping both image and text encoders frozen while training only a lightweight adapter module has been shown to achieve competitive

performance while reducing trainable parameters by over 90% compared to end-to-end fine-tuning approaches (Qin et al., 2024). This knowledge motivates the use of frozen features extraction as the baseline stage of the pipeline, establishing what the pretrained backbone can achieve with minimal supervision before any architectural adaptation is introduced.

3.4.2 Backbone Fine-Tuning

While frozen feature extraction preserves pretrained representations, it limits the model's ability to adapt its visual features to domain-specific characteristics. Partial fine-tuning addresses this by selectively unfreezing a subset of the backbone layers while keeping the remainder frozen. Partial freezing of Vision Transformer (ViT) has been shown to achieve remarkable performance, particularly under limited sample sizes, outperforming various common fine-tuning methods while preserving foundational pretrained knowledge (Zheng et al., 2023). These findings motivate adopting partial backbone adaptation as a core component of the progressive fine-tuning pipeline explored in this work.

3.4.3 Multi-Scale Feature Extraction

Standard ViT-based encoders produce a single pooled representation from the final transformer block, which captures high-level semantic information but discards intermediate representations that encode lower-level structural and texture features. Multi-scale approaches address this by extracting representations from multiple points in the network hierarchy (Figure 3.9). Integrating multi-scale feature representations from ViT architectures, providing substantial improvements in classification performance on medical imaging benchmarks, including ISIC 2018 and ChestX-ray14, with F1-score improvements of up to 4.84% over standard ViT baselines (Alqahtani et al., 2025).

This principle is consistent with findings in hierarchical vision architectures more broadly, where combining features at different depths captures complementary information at early (local structure and texture), mid-level (shape and pattern), and late (semantic and global context) stages of the encoder (Ji et al., 2025). These motivate adopting a multi-scale [CLS] token representation, combined with a feature fusion strategy to accommodate intermediate representations in this work.

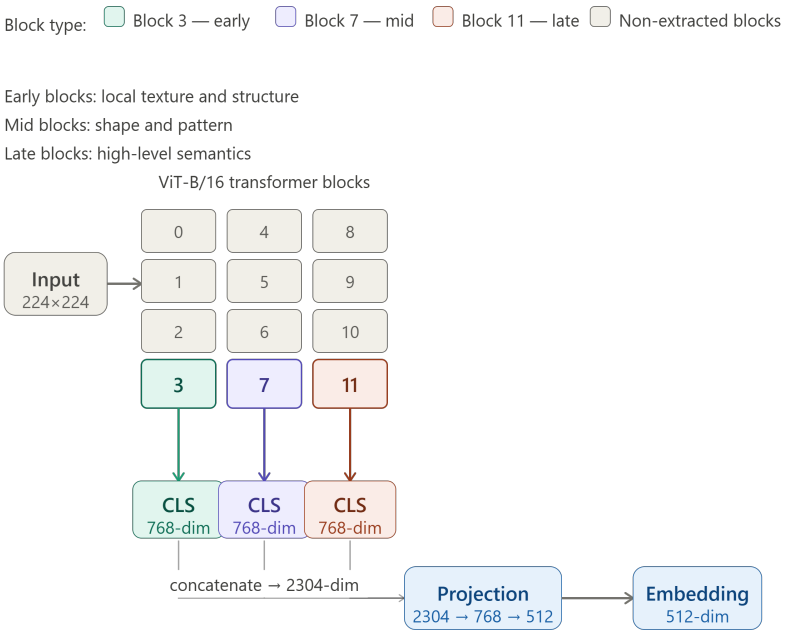


FIGURE 3.9: Multi-scale CLS token extraction from a ViT-B/16 visual encoder. CLS token representations are extracted from the transformer blocks 3, 7, and 11, capturing early structural, mid-level pattern, and high-level semantic features, respectively. The three 768-dimensional tokens are concatenated to form a 2304-dimensional multi-scale representation, which is then projected to a 512-dimensional embedding space via a two-layer projection head, aligning with the shared vision-language embedding space of the pretrained CLIP model.

3.4.4 Temperature Calibration

In CLIP based classification, similarity logits are scaled by a learnable temperature parameter referred to as the logit scale, which is jointly optimised during contrastive pretraining to control the sharpness of the similarity distribution (Radford et al., 2021).

When backbone embeddings shift during partial fine-tuning, the pretrained logit scale may no longer be appropriately calibrated to the adapted embedding distribution, as the geometric relationship between image and text embeddings changes with each unfrozen layer. Fine-tuning CLIP improves task performance, but can alter the calibration properties of the pretrained model (Kumar et al., 2022), and recent work has shown that CLIP models repurposed for downstream tasks exhibit overconfident logit scores irrespective of true image-language pair compatibility (Wang et al., 2024).

These findings motivate evaluating the logit scale as a learnable parameter after partial backbone adaptation, where embedding drift from unfreezing the upper transformer blocks is most pronounced.

3.4.5 Prompt Learning and Soft Prompts

Prompt tuning is a parameter-efficient adaptation strategy that replaces full weight updates with a small set of learnable continuous tokens appended to the text encoder input. This mechanism enables task adaptation by shifting the textual embedding space while keeping the backbone parameters fixed.

Context Optimisation (CoOp) formalises this paradigm in CLIP-based architectures by substituting manually designed prompts with optimised context vectors learned end-to-end from downstream labelled data (Figure 3.10). This formulation consistently improves performance over fixed prompt templates across multiple image recognition benchmarks (Zhou et al., 2022). Extensions based on class-specific context (CSC) further decouple the prompt representation per class, allowing independent context vectors that better capture intra-class variability and improving robustness in settings with high inter-class diversity.

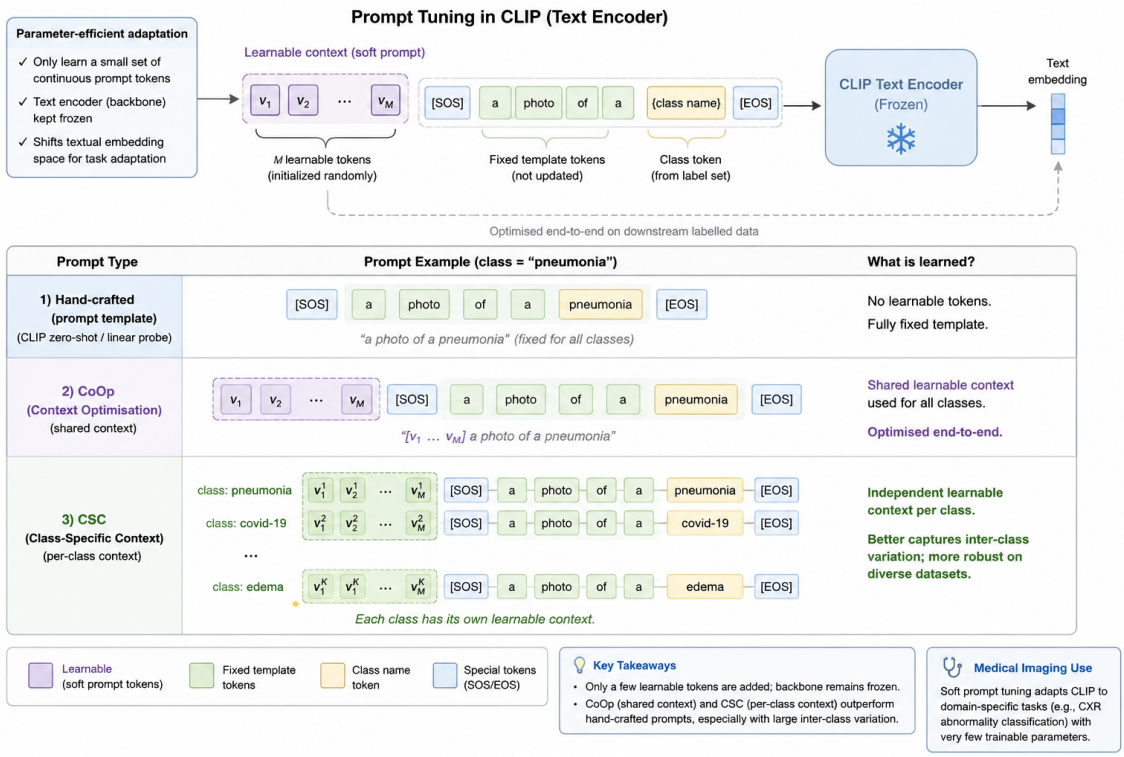


FIGURE 3.10: Prompt tuning strategies in CLIP-based models. The diagram illustrates parameter-efficient adaptation via learnable soft prompt tokens appended to the text encoder input while keeping the CLIP backbone frozen. It contrasts (1) hand-crafted fixed templates, (2) Context Optimisation (CoOp) with shared learnable context vectors, and (3) Class-Specific Context (CSC) with independent per-class prompt embeddings. This formulation enables task adaptation by shifting the textual embedding space without modifying backbone weights, improving performance in downstream recognition tasks, including medical imaging scenarios under limited data regimes.

In medical imaging applications, soft prompt tuning has been shown to effectively adapt vision-language models to domain-specific tasks while maintaining a frozen backbone, achieving competitive performance under strict parameter constraints (Chen et al., 2023). These properties motivate evaluating CoOp-based prompt learning as a lightweight text-side adaptation component within the proposed framework.

3.4.6 Mixture of Experts

Mixture of Experts (MoE) architectures implement conditional computation by routing each input through a subset of specialised sub-networks, selected via a learned gating function. This mechanism enables adaptive allocation of computational resources conditioned on input characteristics, thereby improving efficiency and scalability (Figure 3.11).

Vision Mixture of Experts (V-MoE), a sparse adaptation of Vision Transformers, incorporates sparsely gated MoE layers within attention blocks, demonstrating that conditional routing can maintain competitive performance while substantially reducing inference cost compared to dense models (Riquelme et al., 2021). In the medical imaging domain, MoE extensions of Vision Transformers have been investigated for segmentation, where input-dependent expert selection improves feature discrimination and consistently outperforms standard dense baselines (Ou et al., 2022).

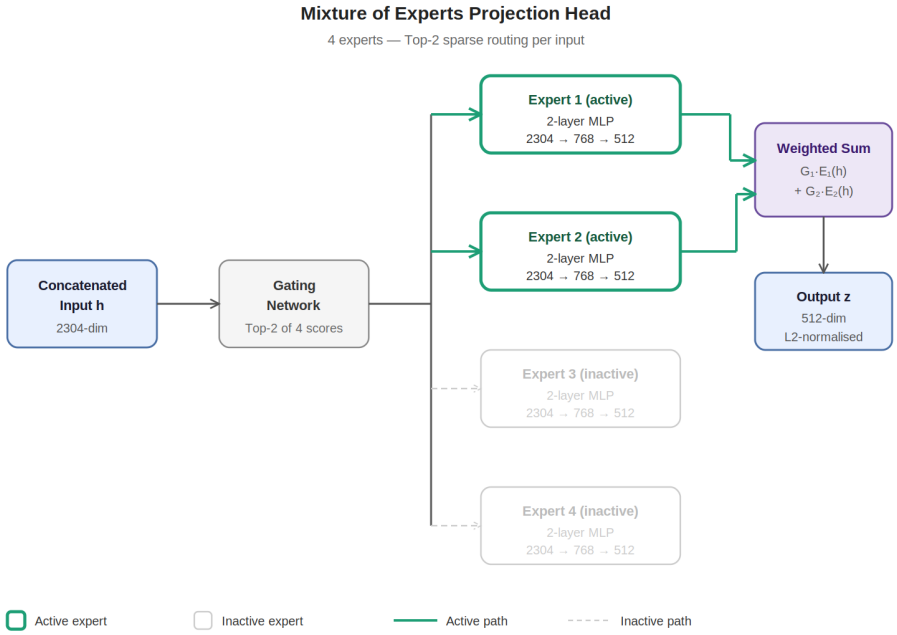


FIGURE 3.11: Sparse Mixture of Experts projection head. The concatenated multi-scale input h is processed by a gating network that selects the top 2 experts from a pool of 4. Only the selected experts compute outputs, which are combined via a weighted sum to produce the final 512-dimensional L2-normalised embedding.

More recently, M4oE introduces a multimodal medical framework that integrates modality-specific experts with a learned routing network, enabling adaptive fusion of heterogeneous representations and improving both generalisation across imaging modalities and interpretability (Jiang and Shen, 2024). Collectively, these findings motivate incorporating sparse MoE-based routing mechanisms as an architectural component within the projection head of the proposed framework.

3.5 Progressive Fine-Tuning Pipeline

The proposed framework follows a progressive design, where architectural and optimisation complexity is introduced incrementally. Each stage builds on the previous one, enabling controlled evaluation of individual components and isolating their contribution to performance. The pipeline comprises seven main configurations (Ce1–Ce7), followed by a final hyperparameter refinement step.

TABLE 3.2: Summary of progressive fine-tuning pipeline stages, describing the architectural change introduced at each stage and its primary objective.

Stage	Name	Change Introduced	Objective
Ce1	Zero-Shot Baseline	No training	Establish out-of-the-box performance of pretrained backbone
Ce2	Single-Scale Extraction	Projection head on frozen backbone	Assess task-specific adaptation without backbone modification
Ce3	Multi-Scale CLS Fusion	CLS tokens from blocks 3, 7, 11	Capture complementary structural and semantic features
Ce4	Partial Fine-Tuning	Top 3 blocks unfrozen with LLRD	Enable domain-specific adaptation of upper backbone layers
Ce4A	Extended Fine-Tuning	Top 6 blocks unfrozen with LLRD	Assess benefit of deeper backbone adaptation
Ce4B	Sequential Unfreezing	Two-phase 3+3 block unfreezing	Assess whether gradual unfreezing improves over simultaneous
Ce5	Learnable Temperature	Logit scale made learnable	Recalibrate similarity distribution after backbone adaptation
Ce6	Soft Prompt Tuning	16 learnable context tokens per class	Assess text-side adaptation as complement to visual adaptation
Ce7	MoE Projection Head	4-expert sparse MoE with top-2 routing	Assess conditional computation over shared projection head

3.5.1 Ce1: Zero-Shot Baseline

The first stage establishes a zero-shot baseline by applying the pretrained model directly to the target datasets without any task-specific training. Image embeddings are produced by the frozen visual encoder and compared against text embeddings derived from task-specific prompts using the classification formulation described in Section 3.3.1. No parameters are updated at this stage. This baseline quantifies the out-of-the-box capability of the pretrained vision-language model on the target domain prior to any adaptation.

3.5.2 Ce2: Single-Scale Feature Extraction

The second stage introduces a lightweight two-layer projection head trained on top of frozen backbone features. The visual encoder remains fully frozen, and the standard pooled output of the visual encoder is used as the image representation, producing a 512-dimensional embedding via `encode_image()`. The projection head maps this embedding to the shared vision-language space via:

$$z = W_2(\text{ReLU}(W_1(f_I(x)))) \tag{3.3}$$

where $W_1 \in \mathbb{R}^{512\times512}$ and $W_2 \in \mathbb{R}^{512\times512}$ are learnable projection matrices. The output is L2-normalised before classification. This stage isolates the contribution of task-specific adaptation of the output representation without modifying the pretrained backbone features.

3.5.3 Ce3: Multi-Scale CLS Fusion

The third stage replaces the single-scale representation with a multi-scale CLS token fusion strategy. Rather than using only the final pooled output, CLS token representations are extracted from intermediate transformer blocks 3, 7, and 11, capturing early structural, mid-level pattern, and high-level semantic features respectively. The three 768-dimensional tokens are concatenated and projected to 512 dimensions via a two-layer projection head:

$$z = W_2(\text{ReLU}(W_1([h_3; h_7; h_{11}]))) \tag{3.4}$$

where $h_i \in \mathbb{R}^{768}$ denotes the CLS token from block i, $W_1 \in \mathbb{R}^{768\times2304}$ and $W_2 \in \mathbb{R}^{512\times768}$ are learnable projection matrices. The output is L2-normalised before classification:

$$\hat{z} = \frac{z}{\|z\|} \tag{3.5}$$

where $\hat{z}$ replaces $f_I(x)$ in the classification equation. The visual backbone remains fully frozen at this stage, isolating the contribution of multi-scale feature fusion from backbone adaptation.

3.5.4 Ce4: Partial Backbone Fine-Tuning

The fourth stage introduces the selective unfreezing of the upper three transformer blocks of the visual encoder, allowing the backbone to adapt its representations to the target domain. The remaining backbone layers are kept frozen via `requires_grad=False`. Layer-wise learning rate decay (LLRD) is applied with a decay factor of 0.80, such that block 11 receives the full backbone learning rate of 1×10^{-5} while lower unfrozen blocks receive progressively reduced rates:

$$\mathrm{lr}_i = \mathrm{lr}_{\text{backbone}} \times \delta^{(11-i)} \tag{3.6}$$

where $\delta = 0.80$ is the decay factor and i denotes the block index. This stage builds on the multi-scale CLS fusion architecture from Ce3. This decay value was selected to apply moderate attenuation to lower backbone layers while preserving the majority of pretrained representations, consistent with common practice in partial fine-tuning of vision transformers.

3.5.5 Ce4A and Ce4B: Ablation Variants

Two ablation variants explore the effect of the number and sequencing of unfrozen blocks. Ce4A extends Ce4 by unfreezing the top six transformer blocks simultaneously, doubling the adaptation capacity while maintaining the same LLRD schedule applied over six blocks. Ce4B instead adopts a sequential two-phase unfreezing strategy: Phase 1 unfreezes the top three blocks as in Ce4, and Phase 2 subsequently unfreezes the next three blocks while retaining the Phase 1 weights. This sequential approach investigates whether gradual unfreezing offers benefits over simultaneous unfreezing of the same total number of blocks.

3.5.6 Ce5: Learnable Temperature Scaling

The fifth stage extends the best Ce4 by making the logit scale parameter τ learnable, optimised as a separate parameter group at a learning rate of 1×10^{-6}. The logit scale is constrained to the interval $[0, \ln(100)]$ to ensure numerical stability. This stage investigates whether recalibrating the temperature parameter following the backbone adaptation improves classification performance, motivated by the observation that embedding drift during partial fine-tuning may render the pretrained logit scale suboptimal for the adapted embedding distribution.

3.5.7 Ce6: Soft Prompt Tuning

The sixth stage introduces text-side adaptation via CoOp-based soft prompt tuning (Zhou et al., 2022), extending the Ce5 version as an independent branch. Sixteen learnable context tokens are optimised per class, following the original CoOp recommendation as a stable mid-range context length across recognition tasks, while the text encoder weights

remain frozen. The context tokens are initialised from the existing prompt token embeddings, which are prepended to the class name token embeddings before being passed to the frozen text encoder. This stage investigates whether learnable text-side adaptation provides complementary benefits to the visual adaptation achieved in Ce5.

$$t_c = [v_1, v_2, \ldots, v_M, \text{cls}_c] \tag{3.7}$$

where $v_1, \ldots, v_M \in \mathbb{R}^d$ are the $M = 16$ learnable context vectors and cls_c is the class name token embedding for class c.

3.5.8 Ce7: Mixture of Experts Projection Head

The seventh stage replaces the standard projection head with a sparse Mixture of Experts (MoE) projection head, extending the Ce5 version as an independent branch. Four expert networks are defined, each consisting of a two-layer MLP projection, with a learned gating network selecting the top-2 experts for each input via sparse routing:

$$z = \sum_{k \in \text{Top-2}(G(h))} G_k(h) \cdot E_k(h) \tag{3.8}$$

where $h = [h_3; h_7; h_{11}]$ is the concatenated multi-scale representation, $G(h)$ is the gating network output, and E_k denotes the k-th expert network. This stage investigates whether conditional computation via expert routing provides benefits over a single shared projection head for the multi-scale feature fusion task. Top-2 routing was adopted as a balance between the representational diversity of multiple experts and the computational overhead of dense routing, consistent with standard sparse MoE practice (Shazeer et al., 2017).

3.6 Training Procedure

All experiments follow a unified optimisation protocol to ensure fair comparison across pipeline stages.

3.6.1 Optimiser and Scheduling

Training uses AdamW with weight decay 1×10^{-2}. A cosine annealing schedule with $T_{\max} = 100$ is applied.

Differential learning rates are used:

- Projection head: 1×10^{-4}
- Backbone (unfrozen layers): 1×10^{-5}
- When applicable, LLRD with decay factor 0.80

Early stopping with a patience of 10 epochs is used, selecting the checkpoint with the best validation balanced accuracy. This patience was chosen to allow sufficient time for validation performance to recover from temporary plateaus without permitting excessive overfitting. Batch size is set to 32, and all experiments use seed 42 for reproducibility.

3.6.2 Loss Function

The class-weighted cross-entropy loss was used as the optimisation objective, incorporating inverse-frequency class weighting computed from the training data. Specifically, each class weight was defined as

$$w_c = \frac{1}{n_c}, \tag{3.9}$$

where n_c denotes the number of samples in class c. These weights were normalised so that their sum equals the number of classes, thereby increasing penalisation for misclassifying minority classes and improving robustness to class imbalance.

3.6.3 Data Splits

Dataset splits were applied consistently across all experiments. For the ISIC 2019 and CELLo datasets, a stratified 70%/10%/20% split was used for training, validation, and testing, respectively, preserving class distributions and patient split among the split sets. For the Raabin-WBC dataset, the official Train/Test-A split provided by the dataset authors was followed, while a validation subset was further extracted from the training partition, resulting in an approximate 60%/10%/30% overall split. Stratified sampling is used to maintain a consistent distribution of classes among the sets.

3.6.4 Embedding and Regularisation

Both image and text embeddings were L2-normalised prior to similarity computation. Classification was performed using scaled cosine similarity between normalised image and text embeddings, leveraging each model's pretrained temperature parameter. Text

embeddings were precomputed once prior to training using a frozen text encoder and reused throughout training, except in configurations involving soft prompt tuning, where embeddings were recomputed at each forward pass due to the adaptive nature of the text encoder.

3.6.5 Prompt Design

Dataset-specific prompts are used:

- CELLo: "a microscopy image of {class} bladder cancer cells"
- ISIC 2019: "a dermoscopy image of {class}"
- Raabin-WBC: "a microscopy image of {class} white blood cell"

3.6.6 Training Configuration Summary

Table 3.3 summarises the fixed training parameters applied consistently across all pipeline stages and datasets.

TABLE 3.3: Fixed training configuration applied across all pipeline stages.

Parameter	Value
Optimiser	AdamW
Weight decay	1×10^{-2}
Scheduler	CosineAnnealingLR ($T_{\max} = 100$)
Maximum epochs	100
Early stopping	Patience 10 (best validation BACC)
Batch size	32
Seed	42
Loss function	Class-weighted cross-entropy
Embedding norm	L2 normalisation

3.6.7 Evaluation Metrics

The primary evaluation metric used across all experiments was balanced accuracy, defined as the mean per-class recall. This metric is particularly appropriate for imbalanced datasets as it treats all classes equally, regardless of sample size, providing a more reliable measure of model performance than standard accuracy. Macro AUC was reported

as a supplementary metric, measuring the model's discriminative ability across all classes independently of the classification threshold. Macro precision and macro F1-score were also reported to provide granular insight into performance across individual classes.

Attention map visualisation was employed to assess the spatial focus of the model's predictions. Attention weights were extracted from the final transformer block of the visual encoder. The `[CLS]` token attention to all patch tokens was retrieved, averaged across attention heads, and upsampled to the original image resolution via bilinear interpolation. Maps were generated for representative correct and incorrect predictions per class to assess whether the model attended to biologically meaningful regions. For pipeline stages involving partial backbone fine-tuning, attention maps were extracted from the fine-tuned model to capture the adapted spatial attention patterns. For frozen backbone stages, the pretrained backbone was used directly.

3.7 Chapter Summary

This chapter presented the methodological framework underpinning the proposed approach. The overall pipeline was introduced, leveraging pre-trained vision-language models based on CLIP for multimodal image classification. The three main stages were outlined: backbone selection, progressive fine-tuning, and cross-dataset evaluation.

The three datasets were described in detail. The CELLo dataset was introduced as the primary biomedical dataset for bladder cancer cell classification, while ISIC 2019 and Raabin-WBC were employed as external benchmarks to evaluate cross-domain generalisation under distribution shift.

The multimodal architecture was then presented, covering the CLIP framework, the candidate backbone models, and their key architectural differences. The progressive fine-tuning pipeline was described across seven stages, from a zero-shot baseline to soft prompt tuning and a mixture of experts. The training procedure was detailed, covering the optimisation strategy, loss function, data splitting protocol, regularisation, prompt design, and evaluation metrics.

Chapter 4

Experiments and Results

This chapter presents the experimental setup and reports the results obtained from evaluating the proposed approach across multiple datasets and configurations.

4.1 Experimental Setup

All experiments follow the training configuration, dataset splits, and evaluation protocol described in Chapter 3. This chapter reports the results obtained under that configuration across all pipelines stages and datasets, along with the specific changes made at every step, and how the pipeline evolves.

4.1.1 Backbone Selection

To select the best visual encoder for downstream adaptation, an initial backbone selection step was performed due to the growing variety of vision-language and foundation models available for biomedical imaging. Without task-specific fine-tuning, candidate models were initially evaluated in a zero-shot setting to assess their inherent representation quality and alignment with bladder cancer cell morphology. To assess generalisation across several biological imaging domains, external benchmarks were considered alongside performance on the target CELLo dataset. Table 4.1 summarises the obtained results.

TABLE 4.1: Zero-shot backbone comparison across datasets. BACC = Balanced Accuracy; Macro AUC = macro-averaged area under the ROC curve.

Model	CELLo		RAABIN-WBC		ISIC	
	BACC	Macro AUC	BACC	Macro AUC	BACC	Macro AUC
BiomedCLIP	0.2033	0.4427	0.3088	0.6423	**0.2331**	**0.6633**
PubMedCLIP	0.2002	0.4490	0.1952	0.3954	0.1182	0.5779
PLIP	0.2070	0.4646	0.1786	0.4632	0.1716	0.5426
OpenCLIP ViT-B/16	**0.2796**	**0.5977**	0.2352	0.5647	0.1439	0.5656
OpenCLIP ViT-L/14	0.2014	0.4924	**0.3781**	**0.6694**	0.1170	0.6199
OpenCLIP ViT-H/14	0.1577	0.4662	0.2947	0.6407	0.1509	0.6462

On the primary CELLo dataset, OpenCLIP ViT-B/16 achieved the strongest zero-shot balanced accuracy (0.2796), outperforming all biomedical domain-specific models despite their specialised pretraining. BiomedCLIP and the remaining candidates achieved balanced accuracy values around 0.20, close to random chance for a five-class classification problem. These results suggest that, despite their biomedical pretraining, the models were insufficiently aligned with the visual characteristics of bladder cancer cell-line microscopy images to support meaningful zero-shot classification. This observation is consistent with the fact that their pretraining corpora, although biomedical in nature, do not specifically capture urothelial cell morphology or tumour staging patterns. Across the external generalisation benchmarks, BiomedCLIP demonstrated competitive performance on Raabin-WBC and achieved the highest balanced accuracy on ISIC 2019 among all evaluated models. These findings indicate broader alignment across diverse imaging modalities within the biomedical domain and provide an initial assessment of each model's generalisation capability before fine-tuning, a question further explored in the cross-dataset analysis. Based on these results, two models were selected for evaluation. OpenCLIP ViT-B/16 was chosen as the strongest zero-shot performer on the target dataset. In contrast, BiomedCLIP (referred to as BIOCE) was retained as a domain-specific comparator due to its established performance across a wide range of biomedical imaging tasks reported in the literature. The remaining candidates were excluded due to consistently lower performance across datasets.

4.1.2 Backbone Selection Analysis

Following backbone selection, both were taken through a preliminary version of the pipeline very similar to the one discussed in this work, and although OpenCLIP ViT-B/16 attained slightly higher balanced accuracy (+1%) than BiomedCLIP, qualitative analysis of attention maps revealed systematic differences in spatial focus between the two backbones that informed the final selection decision.

Attention weights were extracted from the final transformer block of each model following preliminary fine-tuning, capturing high-level spatial saliency patterns most directly associated with the classification decision. Figures 4.1 and 4.2 show representative attention activations for T3 and NA tumour stage samples, respectively.

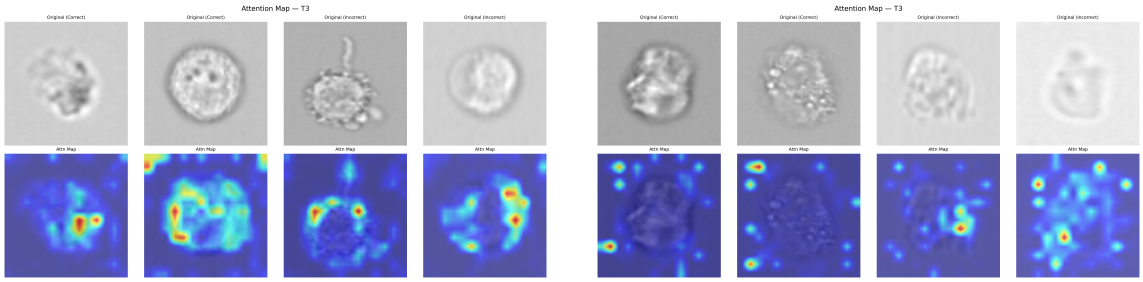


FIGURE 4.1: Attention maps for representative T3 class samples from preliminary backbone evaluation experiments. Left: BiomedCLIP. Right: OpenCLIP ViT-B/16. Each panel shows two correctly classified samples, followed by two misclassified samples, with the original image above and the corresponding attention map below.

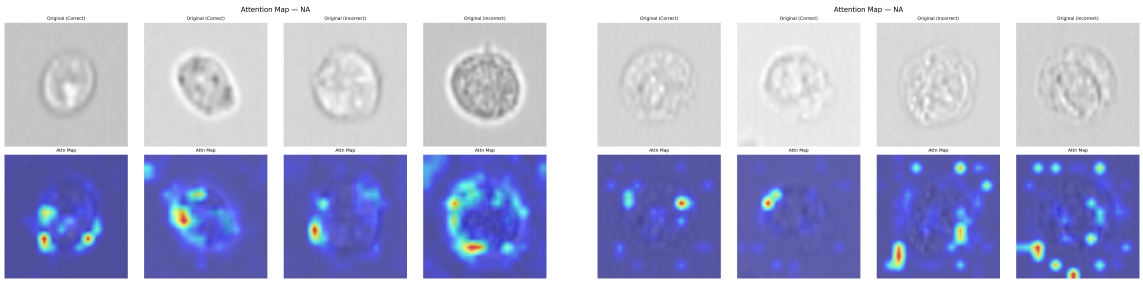


FIGURE 4.2: Attention maps for representative NA class samples from preliminary backbone evaluation experiments. Left: BiomedCLIP. Right: OpenCLIP ViT-B/16.

For BiomedCLIP, attention consistently concentrated on the cell body and internal structures, including nuclear and cytoplasmic regions, across both correct and incorrect predictions. For OpenCLIP ViT-B/16, while attention was also directed towards the cell in many cases, a higher incidence of strong activations in background regions was observed,

suggesting partial reliance on non-cellular cues for prediction. This behaviour was consistent across tumour stages and not limited to misclassified cases, indicating a systematic difference in spatial focus rather than isolated prediction artefacts.

In a clinical staging context where interpretability and spatial grounding are critical alongside classification performance, these findings support selecting BiomedCLIP as the primary visual encoder for all subsequent experiments. It should be noted that this selection does not guarantee BiomedCLIP is the optimal backbone in absolute performance terms. However, the systematic attention patterns observed with OpenCLIP raised concerns regarding the clinical reliability of its predictions, motivating the prioritisation of interpretability in the final selection decision to run only BiomedCLIP in the final version of the pipeline.

Selected Image Backbone The visual backbone used throughout all subsequent pipeline stages is BiomedCLIP's ViT-B/16 encoder, pretrained on approximately 15 million biomedical image-text pairs from PubMed Central spanning radiology, pathology, and microscopy modalities. The encoder produces 512-dimensional image embeddings, which are L2-normalised prior to similarity computation.

Selected Text Backbone BiomedCLIP's text encoder is PubMedBERT, pretrained on 21 billion words from PubMed abstracts and full-text articles from PubMed Central. It produces 512-dimensional text embeddings compatible with the image embedding space. The text encoder remains frozen throughout all pipeline stages, except where soft prompt tuning is applied, in which case the input token embeddings are modified while the encoder weights remain fixed.

4.2 Progressive Fine-Tuning Pipeline

This section outlines the procedures for progressively fine-tuning the model pipeline component by component.

4.2.1 Architectural Model Modifications

Following backbone selection, a progressive fine-tuning pipeline was evaluated on BiomedCLIP, introducing one architectural change at a time to isolate the contribution of each component to the overall classification performance.

The following table presents the results obtained from stages Ce1–Ce4B, highlighting the architectural changes introduced at each stage and their impact on performance. The results are subsequently explained and discussed stage by stage.

TABLE 4.2: Progressive fine-tuning pipeline results on the CELLo test set. BACC = balanced accuracy (primary metric), all other metrics are macro-averaged. Δ denotes absolute improvement over the previous stage.

Stage	Description	BACC	Δ BACC	AUC	Precision	F1
Ce1	Zero-shot baseline	0.2033	—	0.4427	0.1486	0.1314
Ce2	Frozen single-scale	0.8407	+0.6374	0.9669	0.7640	0.7933
Ce3	Frozen multi-scale	0.9071	+0.0664	0.9897	0.8593	0.8796
Ce4	Partial fine-tuning (3 blocks)	0.9267	+0.0196	0.9944	0.8818	0.9008
Ce4A	6 blocks (vs Ce4)	0.9317	+0.0050	0.9952	0.8842	0.9040
Ce4B	Sequential 3+3 (vs Ce4A)	0.9312	−0.0005	0.9945	0.8846	0.9040

4.2.2 Zero-Shot Baseline (Ce1)

The zero-shot baseline corresponds to the backbone selection evaluation presented in Section 4.1.1, where no training is performed, and classification relies entirely on the pretrained vision-language alignment of the BiomedCLIP backbone. As shown in Figure 4.3, the model predicts the vast majority of samples as T4 regardless of true class, indicating that the pretrained text-image alignment does not support meaningful tumour stage discrimination in the zero-shot setting. This establishes a clear motivation for task-specific adaptation.

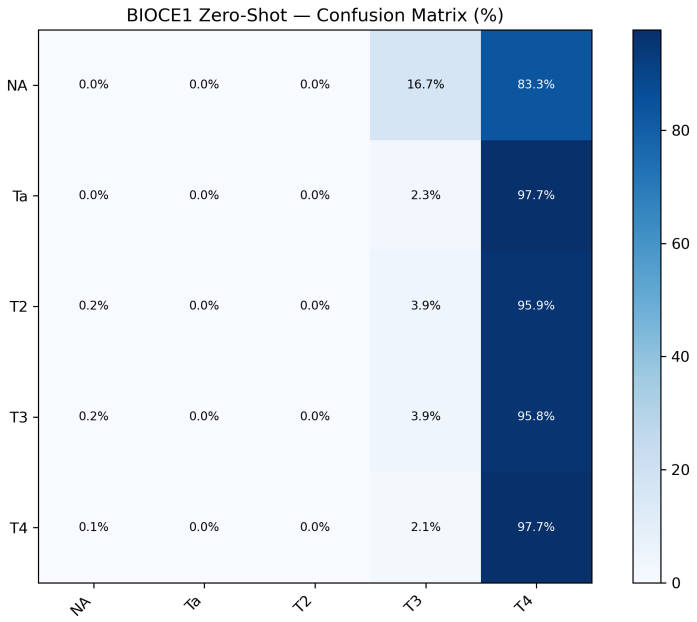


FIGURE 4.3: Confusion matrix for Ce1 on the CELLo test set. The model collapses predictions almost entirely to T4, indicating no meaningful class discrimination in the zero-shot setting.

4.2.3 Single-Scale Feature Extraction (Ce2)

The introduction of a two-layer projection head trained on frozen backbone features produced a substantial improvement over the zero-shot baseline, with balanced accuracy rising from 0.2033 to 0.8407, representing an absolute gain of 0.6374. With the backbone remaining fully frozen, the projection head learned a task-specific mapping from the pretrained feature space to the classification space, suggesting that the backbone representations contain sufficient structure for effective downstream classification even without domain adaptation. The macro AUC of 0.9669 further indicates strong class separability in the learned embedding space.

Figure 4.4 shows the per-class confusion matrix for Ce2. T2 achieved the highest recall at 0.909, while T3 showed the lowest at 0.734, with notable confusion towards T4 (0.122). This pattern is clinically consistent, as T3 and T4 represent adjacent muscle-invasive stages with overlapping morphological characteristics. NA and Ta also performed well at 0.850 and 0.872, respectively. The primary source of error across all classes is confusion between adjacent invasive stages, a pattern that will be revisited as backbone adaptation is introduced in subsequent stages.

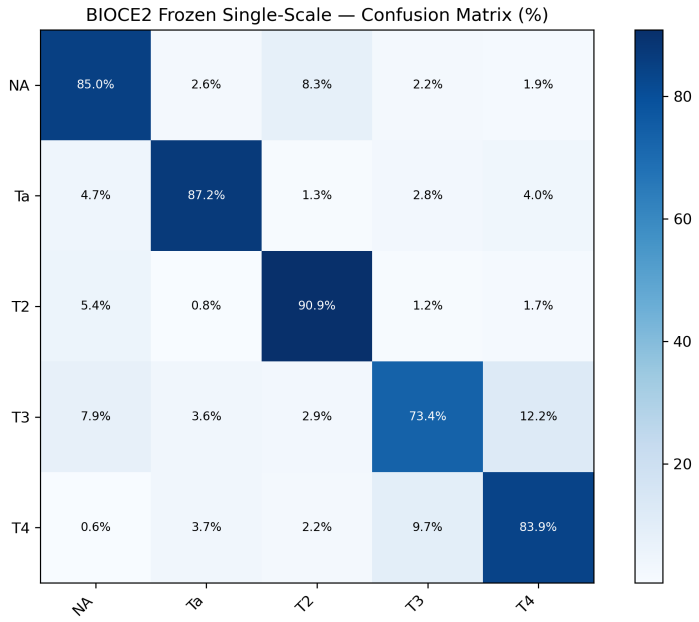


FIGURE 4.4: Confusion matrix for Ce2 on the CELLo test set. Values represent the percentage of samples per class. T3 shows the highest confusion, primarily with T4, consistent with the morphological similarity between adjacent invasive tumour stages.

Configuration Backbone frozen. Projection head trained at $\mathrm{LR}_{\mathrm{head}} = 1 \times 10^{-4}$. All other parameters are as in Table 3.3.

4.2.4 Multi-Scale CLS Fusion (Ce3)

Replacing the single-pooled representation with multi-scale [CLS] token fusion yielded a further improvement in balanced accuracy from 0.8407 to 0.9071, an absolute gain of 0.0664 over Ce2. By extracting [CLS] token representations from transformer blocks 3, 7, and 11, the projection head receives complementary information from early structural, mid-level pattern, and high-level semantic stages of the encoder, rather than relying solely on the final pooled output. The macro AUC improved to 0.9897, suggesting stronger class separability across all tumour stages. Notably, these gains were achieved with the backbone remaining fully frozen, indicating that intermediate representations of the pre-trained model encode discriminative information not fully captured by the final layer alone.

The confusion matrix in Figure 4.5 reflects consistent improvements across all classes relative to Ce2. T3 showed the most notable gain, improving from 0.734 to 0.853, while T4 improved from 0.839 to 0.938 and Ta reached 0.948. The pattern of confusion between

adjacent invasive stages persists, though at reduced magnitude, suggesting that multi-scale feature fusion better captures the morphological differences between tumour stages without any backbone adaptation.

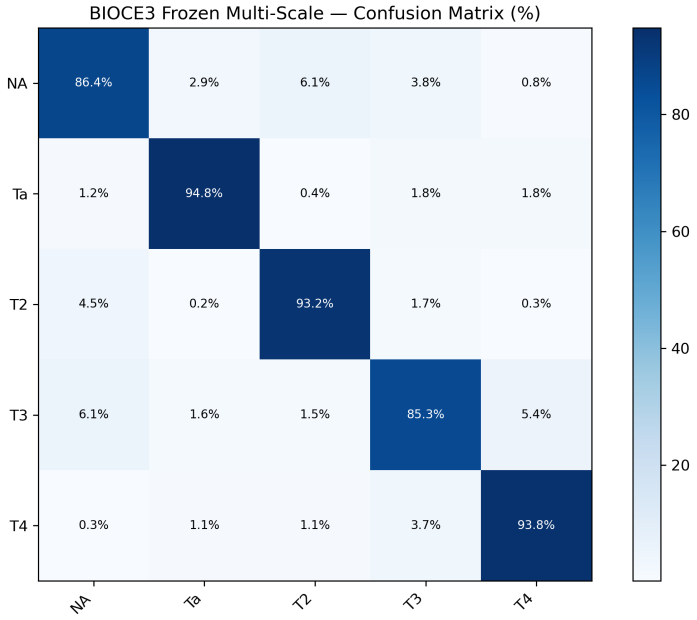


FIGURE 4.5: Confusion matrix for Ce3 on the CELLo test set. Consistent improvements across all classes relative to Ce2, with T3 showing the most notable gain from 0.734 to 0.853.

Figure 4.6 shows the training and validation loss curves for Ce3. Both losses decrease together in the initial epochs before gradually diverging, with the validation loss oscillating throughout training while the training loss continues to decrease smoothly. The growing gap between training and validation loss is consistent with a frozen backbone regime, where the projection head progressively fits the training data, but the fixed representations limit further generalisation improvement. Early stopping was triggered at approximately epoch 47.

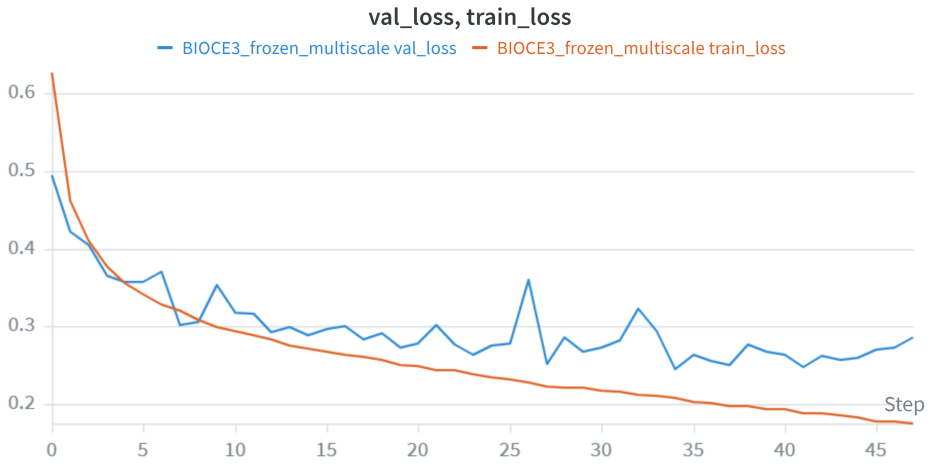


FIGURE 4.6: Training and validation loss curves for Ce3. The validation loss oscillates, while the training loss decreases smoothly, with a gradually widening gap, consistent with frozen-backbone training.

Configuration Backbone frozen. Multi-scale [CLS] extraction from blocks 3, 7, and 11. Projection head trained at $\mathrm{LR}_{\mathrm{head}} = 1 \times 10^{-4}$. All other parameters are as in Table 3.3.

4.2.5 Partial Backbone Fine-Tuning (Ce4, Ce4A, Ce4B)

Introducing selective backbone unfreezing consistently improved over the frozen multi-scale baseline. Ce4, unfreezing the top three transformer blocks with LLRD, achieved a balanced accuracy of 0.9267, an absolute gain of 0.0196 over Ce3. Allowing the backbone to adapt its upper-layer representations to the target domain brought measurable benefit, suggesting that the pretrained features, while already useful, benefit from domain-specific refinement at higher semantic levels.

Ce4A, extending unfreezing to the top six blocks simultaneously, achieved 0.9317, a modest improvement of 0.0050 over Ce4. Ce4B, which adopted a sequential two-phase strategy unfreezing three blocks at a time, achieved 0.9312, marginally below Ce4A. The differences between Ce4A and Ce4B are small and within the range of expected run-to-run variance, suggesting that sequential unfreezing offers no clear advantage over simultaneous unfreezing of the same number of blocks. Ce4A was therefore adopted as the base for subsequent stages, as it achieves comparable performance to Ce4B with a simpler single-phase training procedure.

Figure 4.7 shows the confusion matrix for Ce4A. Backbone adaptation produced the most notable gains in Ta (0.948 to 0.985) and T4 (0.938 to 0.971), suggesting that upper-layer adaptation particularly benefits the more morphologically distinct classes. T3 remained the most challenging class at 0.873. Whereas at earlier stages its residual confusion fell mainly on T4, by Ce4A the largest share had shifted to NA (0.059), with T4 accounting for 0.031.

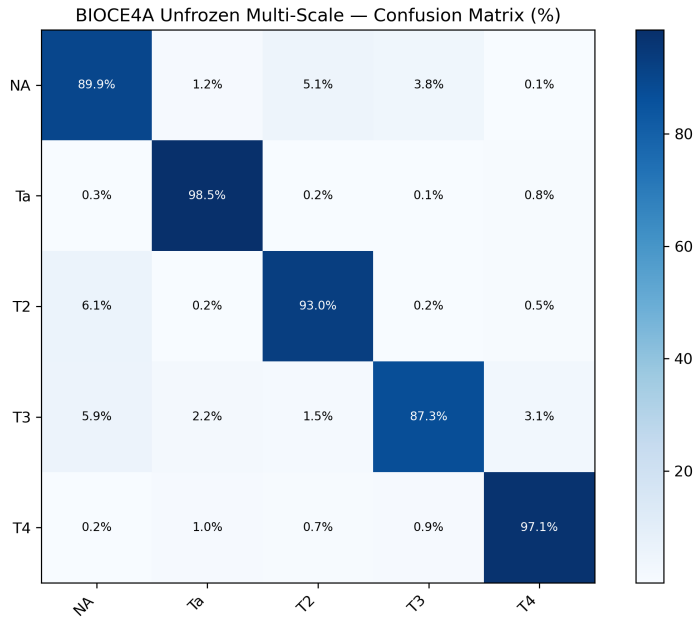


FIGURE 4.7: Confusion matrix for Ce4A on the CELLo test set. Backbone adaptation produced clear gains in Ta and T4, while T3 remained the most challenging class, with residual confusion falling mainly on NA.

Figure 4.8 shows the training and validation loss curves for Ce4A. Compared to Ce3, convergence is substantially faster, with both losses dropping steeply in the first four epochs. The training loss decreases as backbone adaptation enables the model to better fit the training data. The validation loss stabilises at a lower floor than Ce3, reflecting the performance improvement from backbone adaptation, though the larger gap between the training and validation losses indicates greater model capacity. Early stopping was triggered at approximately epoch 22.

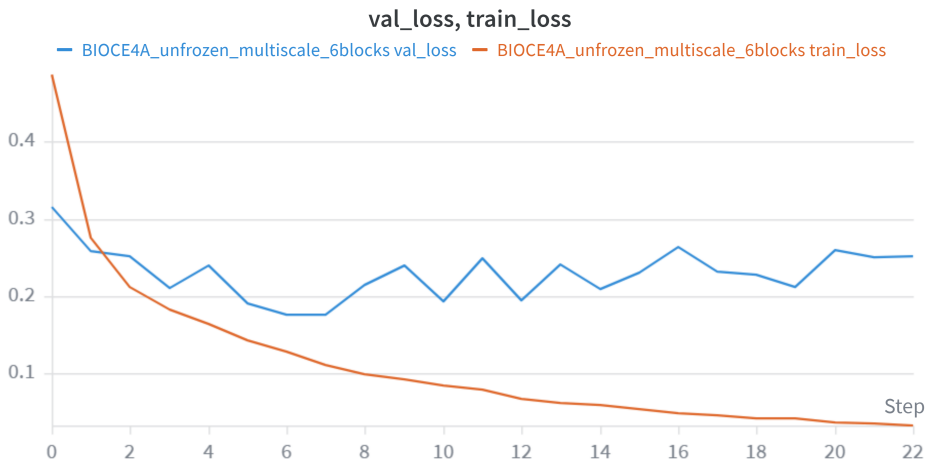


FIGURE 4.8: Training and validation loss curves for Ce4A. Backbone unfreezing produces faster convergence and a lower validation loss floor than Ce3, though with a larger gap between training and validation losses.

Configuration Ce4: top 3 blocks unfrozen. Ce4A: top 6 blocks unfrozen. Ce4B: sequential 3+3 unfreezing. All variants use $\text{LR}_{\text{backbone}} = 1 \times 10^{-5}$ with LLRD factor 0.80. All other parameters as in Table 3.3.

4.2.6 Calibration and Architectural Addons

The following stages build on Ce4A, introducing logit-scale calibration and two independent architectural extensions. Table 4.3 summarises the results under the initial configuration ($\text{LR}_{\text{backbone}} = 1 \times 10^{-5}$), evaluated like-for-like. At this configuration, Ce5 improves slightly over Ce4A, while Ce6 and Ce7, both built from Ce5, provide no meaningful improvement. As no architectural extension produced a meaningful gain over logit-scale calibration alone, Ce5 was selected as the base branch on grounds of parsimony: it adds a single trainable parameter, whereas Ce7 introduces four expert networks and a gating mechanism to reach a statistically equivalent result. This branch was subsequently refined (Section 4.2.10).

TABLE 4.3: Results for calibration and architectural extension stages on the CELLo test set, all evaluated under the initial configuration ($LR_{backbone} = 1 \times 10^{-5}$) for a like-for-like comparison. Ce4A is included as the reference baseline. BACC = balanced accuracy (primary metric), all other metrics are macro-averaged. Δ denotes absolute change relative to Ce4A for Ce5 and Ce5 for Ce6 and Ce7.

Stage	Description	BACC	Δ BACC	AUC	Precision	F1
Ce4A	Unfrozen 6 blocks (baseline)	0.9317	—	0.9952	0.8842	0.9040
Ce5	Learnable logit scale	0.9334	+0.0017	0.9951	0.8935	0.9101
Ce6	Soft prompt tuning (vs Ce5)	0.9325	−0.0009	0.9957	0.8983	0.9113
Ce7	MoE projection head (vs Ce5)	0.9335	+0.0001	0.9955	0.9028	0.9162

4.2.7 Learnable Temperature Scaling (Ce5)

The introduction of a learnable logit-scale parameter improved balanced accuracy from 0.9317 to 0.9334, extending Ce4A by a single trainable parameter. Allowing the temperature to recalibrate following partial backbone adaptation enabled the model to better control the sharpness of the similarity distribution, consistent with the observation that embedding drift during fine-tuning can render the pretrained logit scale suboptimal.

The confusion matrix in Figure 4.9 shows a mixed per-class picture relative to Ce4A: T3 improves markedly from 0.873 to 0.913, while Ta (0.985 to 0.964) and T4 (0.971 to 0.965) show slight regressions and NA remains stable at 0.908, for a small overall gain.

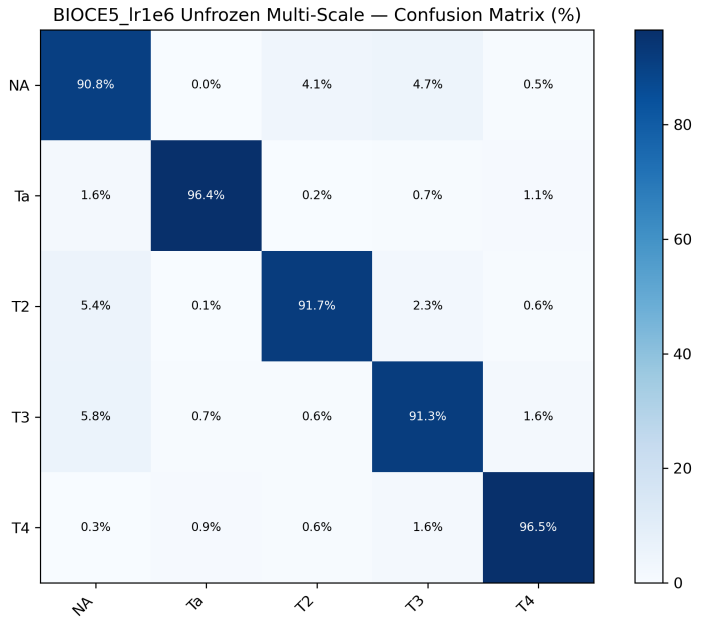


FIGURE 4.9: Confusion matrix for Ce5 on the CELLo test set. T3 improves relative to Ce4A, while Ta and T4 show slight regressions, resulting in a small net gain.

Configuration Ce4A backbone with top 6 blocks unfrozen. Logit scale made learnable at $\mathrm{LR}_{\mathrm{temp}} = 1 \times 10^{-6}$, constrained to $[0, \ln(100)]$. All other parameters as in Table 3.3.

4.2.8 Soft Prompt Tuning (Ce6)

The introduction of CoOp-based soft prompt tuning as an independent branch from Ce5 did not improve upon the Ce5 baseline, achieving a balanced accuracy of 0.9325 compared to 0.9334 for Ce5. While the macro AUC of 0.9957 remained competitive, the text-side adaptation via learnable context tokens provided no additional benefit beyond the visual adaptation already achieved in Ce5 on this dataset. This suggests that the fixed prompt templates used in earlier stages were already sufficiently descriptive for the classification task, and that the additional flexibility introduced by learnable context vectors did not capture information beyond what the visual backbone adaptation provided.

Configuration Ce5 backbone with 16 learnable context tokens per class optimised at $\mathrm{LR}_{\mathrm{prompt}} = 3 \times 10^{-4}$. Text encoder weights remain frozen. All other parameters are as in Table 3.3.

4.2.9 Mixture of Experts Projection Head (Ce7)

Replacing the standard projection head with a sparse Mixture of Experts projection head, also as an independent branch from Ce5, achieved a balanced accuracy of 0.9335, statistically equivalent to Ce5 (0.9334) under the same configuration. The macro F1 of 0.9162 was the highest among the same-configuration branch variants, suggesting that the MoE routing mechanism may improve per-class precision. However, this came at the cost of substantially greater architectural complexity (four expert networks and a gating mechanism) for no meaningful gain in balanced accuracy over the single-parameter Ce5. Conditional expert routing, therefore, provides limited practical benefit over the standard projection head under the evaluated configuration, and Ce5 was retained on grounds of parsimony.

Configuration Ce5 backbone with MoE projection head comprising 4 expert networks and top-2 sparse routing. All other parameters are as in Table 3.3.

4.2.10 Optimal Learning Rate Selection

Following the identification of Ce5 as the strongest configuration, a targeted learning rate search was conducted over the projection head and backbone learning rates to assess whether performance could be further improved under optimised hyperparameters. The search was conducted on the CELLo dataset using the Ce5 architecture, evaluating seven combinations of head and backbone learning rates while keeping all other parameters fixed.

TABLE 4.4: Learning rate grid search results for Ce5 on the CELLo test set. All runs use $\mathrm{LR}_{\mathrm{temp}} = 1 \times 10^{-6}$ and the fixed configuration in Table 3.3. Best performance shown in **bold**.

LR Head	LR Backbone	BACC	AUC	Precision	F1
1×10^{-3}	1×10^{-4}	0.9382	0.9956	0.8997	0.9160
1×10^{-3}	1×10^{-5}	0.9357	0.9946	0.8898	0.9077
1×10^{-4}	1×10^{-4}	**0.9404**	**0.9956**	**0.9008**	**0.9173**
1×10^{-4}	1×10^{-5}	0.9334	0.9951	0.8935	0.9101
1×10^{-4}	1×10^{-6}	0.9286	0.9939	0.8768	0.8972
1×10^{-5}	1×10^{-5}	0.9362	0.9955	0.8988	0.9159
1×10^{-5}	1×10^{-6}	0.9322	0.9941	0.8839	0.9043

The configuration with $LR_{head} = 1 \times 10^{-4}$ and $LR_{backbone} = 1 \times 10^{-4}$ achieved the highest balanced accuracy of 0.9404, a modest improvement over the initial Ce5 configuration of $LR_{backbone} = 1 \times 10^{-5}$, which achieved 0.9334. The results demonstrate low sensitivity to learning rate selection within the evaluated range, with performance varying by less than 1% across all configurations, suggesting that the Ce5 architecture is robust to moderate hyperparameter variation.

4.2.11 Final Configuration Analysis

The selected Ce5 configuration was evaluated across three independent random seeds to assess its stability and reproducibility. Table 4.5 summarises the performance across all three seeds.

TABLE 4.5: Ce5 stability evaluation across three independent seeds. BACC = balanced accuracy; all other metrics are macro-averaged.

Seed	BACC	AUC	Precision	F1
42	0.9404	0.9956	0.9008	0.9173
123	0.9407	0.9962	0.9082	0.9215
7	0.9402	0.9962	0.8974	0.9114
Mean ± Std	0.9404 ± 0.0002	0.9960 ± 0.0003	0.9021 ± 0.0045	0.9167 ± 0.0043

The results demonstrate strong reproducibility across seeds, with balanced accuracy varying by less than 0.03 percentage points. This low variance is notable given that individual training runs observed throughout the experimental pipeline exhibited higher run-to-run variance, attributable to the non-deterministic nature of GPU operations. The stability results, therefore, reflect consistent convergence behaviour under the selected configuration rather than a single favourable outcome.

The confusion matrix for seed 42 is shown in Figure 4.10. Ta achieved the highest per-class recall at 0.988, followed by T4 at 0.974 and T2 at 0.928. T3 and NA show the lowest per-class recall, both at 0.906, with the largest residual confusion occurring between T3 and NA (0.060 and 0.043 respectively). The overall class-level pattern is consistent with the progression observed across earlier pipeline stages, though the dominant source of confusion has shifted from adjacent invasive stages to the T3–NA boundary.

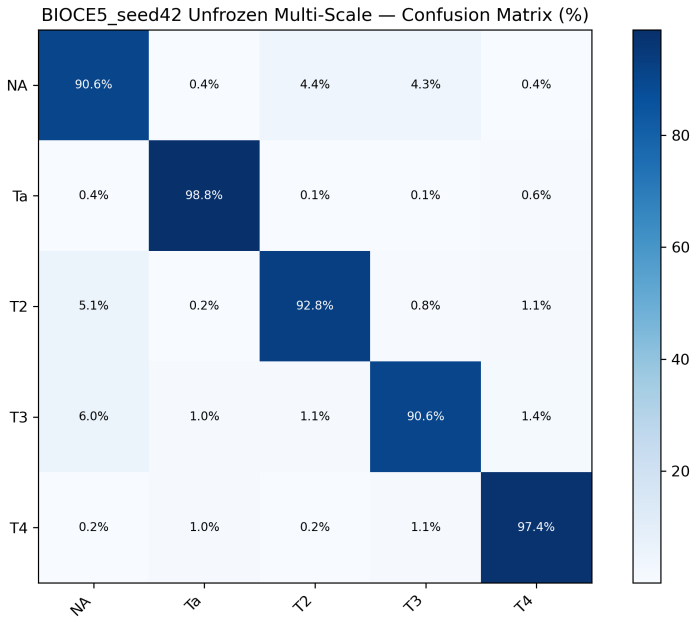


FIGURE 4.10: Confusion matrix for Ce5 (seed 42) on the CELLo test set. Ta and T4 achieve the highest per-class recall, while T3 and NA show the lowest at 0.906, with the largest residual confusion occurring between T3 and NA (0.060 and 0.043 respectively).

The ROC curves in Figure 4.11 confirm strong class separability across all tumour stages, with Ta, T2, and T4 achieving AUC of 1.00 and NA and T3 achieving 0.99. All classes demonstrate near-perfect discrimination, indicating that the model assigns well-calibrated confidence scores across the full spectrum of tumour stages.

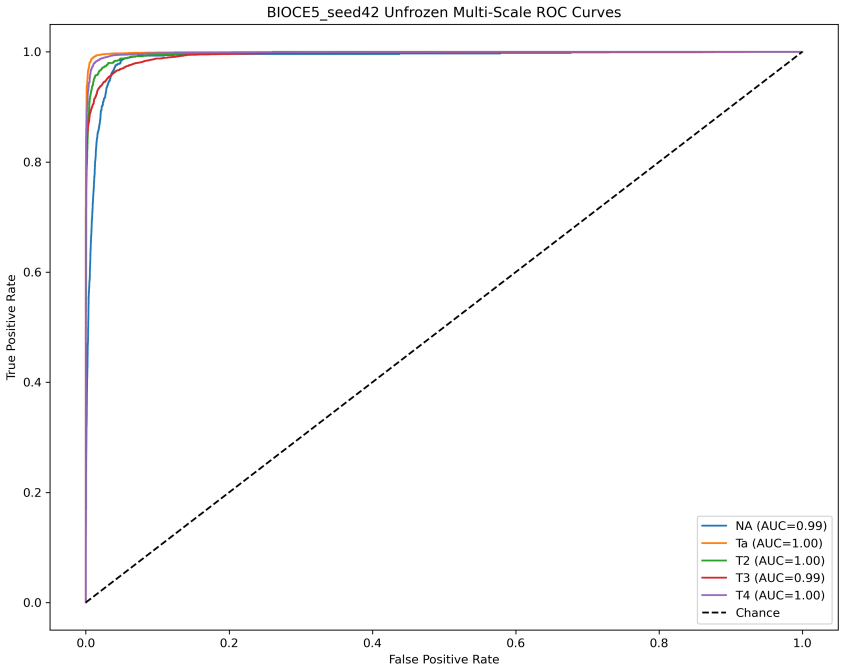


FIGURE 4.11: ROC curves for Ce5 (seed 42) on the CELLo test set. Ta, T2, and T4 achieve an AUC of 1.00, while NA and T3 achieve 0.99, indicating strong class separability across all tumour stages.

Attention maps for all five tumour stage classes are shown in Figures 4.12 through 4.14. Across all classes, the model consistently focuses attention on the cell body and internal structures, including nuclear and cytoplasmic regions. Correctly classified samples show tighter, more focused activation patterns centred on morphologically relevant cellular features. Misclassified samples tend to exhibit more diffuse or peripherally distributed attention, suggesting that classification errors are associated with ambiguous morphological presentations rather than systematic reliance on non-cellular regions.

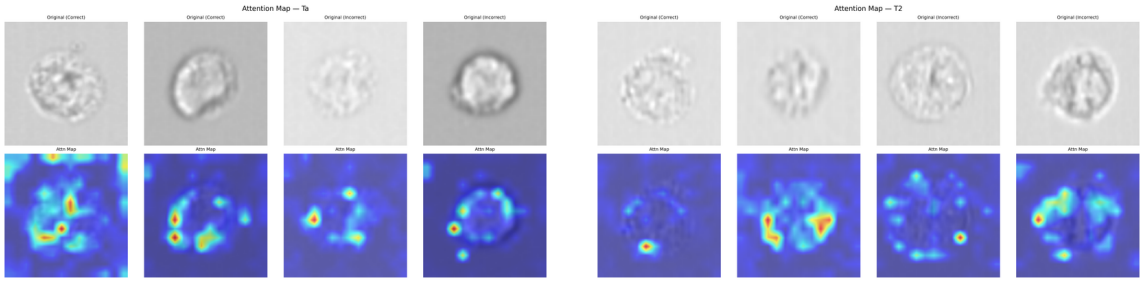


FIGURE 4.12: Attention maps for representative Ta (left) and T2 (right) class samples for Ce5 (seed 42). Each panel shows two correctly classified samples followed by two misclassified samples, with the original image above and the corresponding attention map below.

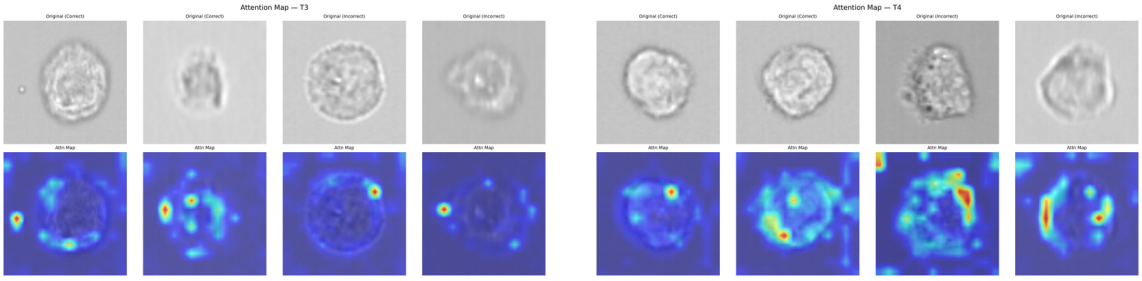


FIGURE 4.13: Attention maps for representative T3 (left) and T4 (right) class samples for Ce5 (seed 42).

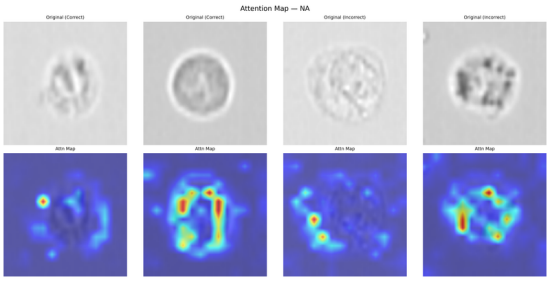


FIGURE 4.14: Attention maps for representative NA class samples for Ce5 (seed 42).

The training and validation loss curves for seed 42 are shown in Figure 4.15. The training loss decreases smoothly and consistently throughout training, while the validation loss drops rapidly in the initial epochs before stabilising with minor oscillations. The gap between training and validation loss is moderate and stable, indicating that the model generalises well without severe overfitting. Early stopping was triggered at approximately epoch 23, consistent with the patience of 10 epochs applied throughout the pipeline.

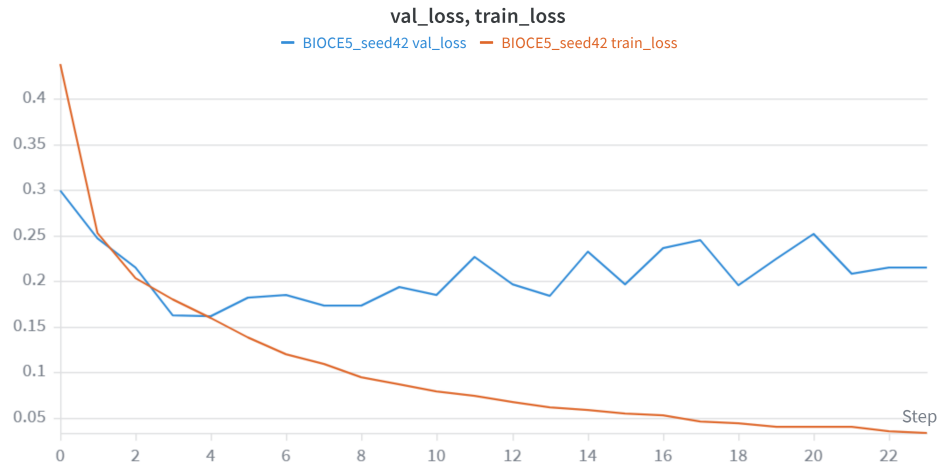


FIGURE 4.15: Training and validation loss curves for Ce5 (seed 42). The training loss decreases consistently while the validation loss stabilises after the initial convergence phase, with early stopping triggered at approximately epoch 23.

Model Complexity The BiomedCLIP backbone consists of a ViT-B/16 visual encoder with 86.19 million parameters and a PubMedBERT text encoder with 109.71 million. Together they account for 195.90 million pretrained parameters, occupying 747.5 MB on disk. Adding the multi-scale projection head brings the full Ce5 model to 198.06 million parameters, of which 44.69 million are trainable. These comprise 42.53 million in the top six visual transformer blocks, 2.16 million in the projection head, and the learnable logit scale. The lower six visual blocks, the patch and positional embeddings, and the full text encoder remain frozen, accounting for the remaining 153.37 million parameters. Note that all parameters are still required at inference, since the frozen weights participate in the forward pass. A single 224×224 image through the visual encoder requires 33.72 GFLOPs (16.85 GMACs), and approximate GPU VRAM usage during inference was 4.6 GB.

4.2.12 Model Configuration Ablation

Several training configuration choices were evaluated during pipeline development. Early experiments used a uniform backbone learning rate across all unfrozen blocks, without layer-wise decay. Introducing LLRD with a decay factor of 0.80 produced more stable training and improved final performance, consistent with established practice in ViT fine-tuning. A linear warmup schedule was also evaluated but removed after showing no consistent benefit under the evaluated conditions. Weight decay was tested at both 1×10^{-4}

and 1×10^{-2}, with the latter producing more consistent results across pipeline stages. These configurations were not reported individually as they preceded the final pipeline design and used inconsistent intermediate configurations.

4.3 Cross-Dataset Evaluation

To assess the generalisation capacity of the proposed pipeline beyond the primary training domain, the full progressive fine-tuning pipeline was evaluated on two external benchmark datasets — ISIC 2019 and Raabin-WBC. All experiments used the same pipeline configuration established on CELLo, without any dataset-specific hyperparameter optimisation, providing a direct assessment of the model performance under distribution shift.

4.3.1 ISIC 2019

Table 4.6 presents the progressive fine-tuning results on the ISIC 2019 dataset. The pipeline demonstrates consistent improvement from the zero-shot baseline through to partial backbone fine-tuning, with balanced accuracy rising from 0.2331 at Ce1 to 0.7546 at Ce4A. While the largest single gain comes from the zero-shot to trained transition (Ce1→Ce2), the subsequent +0.0889 improvement at Ce3 isolates the contribution of multi-scale feature fusion, suggesting it is particularly beneficial for the greater visual diversity of dermoscopic images. Ce4A achieved the strongest performance at 0.7546, with Ce4B and subsequent stages showing marginal degradation. Unlike CELLo, where Ce5 improved upon Ce4A, the learnable logit scale and architectural extensions did not provide additional benefit on ISIC, suggesting that the pretrained logit scale is already well calibrated for this domain or that the performance ceiling under the current configuration has been reached. Notably, Ce4A achieves a higher BACC than Ce4 but a lower F1 (0.6848 vs 0.7222), indicating that the additional unfrozen blocks improve overall class balance at the cost of per-class precision on some categories.

TABLE 4.6: Progressive fine-tuning pipeline results on the ISIC 2019 test set. BACC = balanced accuracy (primary metric), all other metrics are macro-averaged. Δ denotes absolute improvement over the previous stage. Best performance shown in **bold**.

Stage	Description	BACC	Δ BACC	AUC	F1
Ce1	Zero-shot baseline	0.2331	—	0.6663	0.1227
Ce2	Frozen single-scale	0.6198	+0.3867	0.9190	0.5212
Ce3	Frozen multi-scale	0.7087	+0.0889	0.9432	0.6383
Ce4	Partial fine-tuning (3 blocks)	0.7310	+0.0223	0.9582	0.7222
Ce4A	6 blocks (vs Ce4)	**0.7546**	+0.0236	**0.9589**	0.6848
Ce4B	Sequential 3+3 (vs Ce4A)	0.7293	−0.0253	0.9545	0.6948
Ce5	Learnable logit scale	0.7336	−0.0210	0.9580	0.6864
Ce6	Soft prompt tuning (vs Ce5)	0.7290	−0.0046	0.9542	0.6710
Ce7	MoE projection head (vs Ce5)	0.7120	−0.0216	0.9617	**0.7233**

Figure 4.16 shows the confusion matrix for Ce4A on ISIC 2019. Vascular lesion achieved the highest per-class recall at 0.960, while melanoma (0.640) and actinic keratosis (0.682) showed the lowest. Notable confusion was observed between melanoma and melanocytic nevus, and between actinic keratosis and benign keratosis, patterns that are clinically consistent given the visual similarity between these adjacent lesion categories. Dermatofibroma achieved 0.708 despite being a minority class, suggesting the model learned discriminative features for this category.

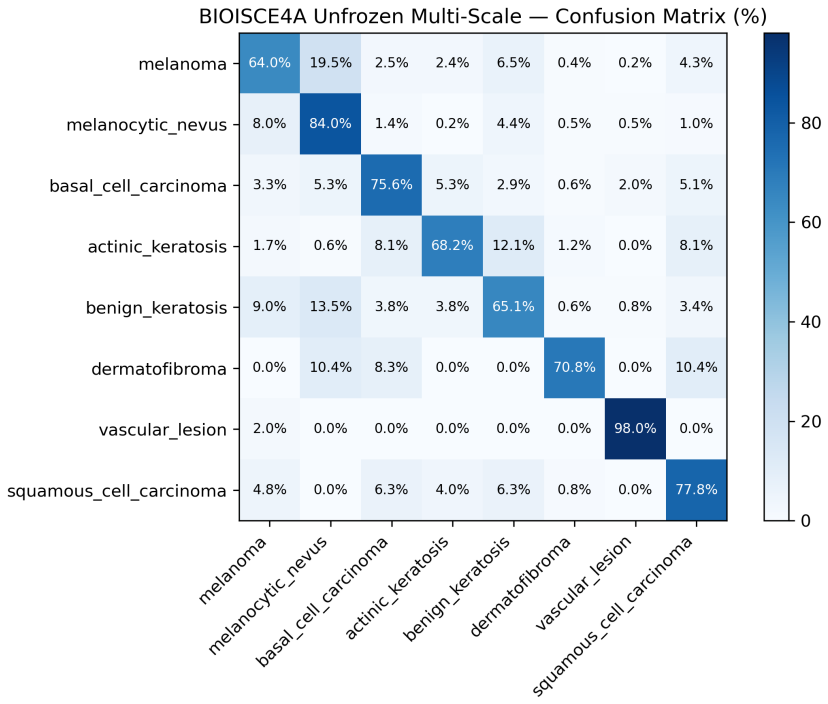


FIGURE 4.16: Confusion matrix for Ce4A on the ISIC 2019 test set. Vascular lesion achieves the highest per-class recall, while melanoma and actinic keratosis show the highest confusion with visually similar categories.

4.3.2 Raabin-WBC

Table 4.7 presents the progressive fine-tuning results on the Raabin-WBC dataset. The pipeline achieved strong performance across all stages, with balanced accuracy rising from 0.3088 at Ce1 to 0.9745 at Ce7. Unlike ISIC and CELLo, performance continued to improve consistently across all stages, including Ce5, Ce6, and Ce7, with Ce7 achieving the highest balanced accuracy across all three datasets. Notably, Ce4B outperformed Ce4A on Raabin (0.9653 vs 0.9554), the only dataset where sequential unfreezing provided a clear advantage over simultaneous unfreezing, suggesting that the gradual adaptation strategy may better suit the morphological characteristics of white blood cell classification. The consistent improvement across Ce6 and Ce7 further distinguishes Raabin from other datasets, suggesting that both text-side adaptation and expert routing provide meaningful benefits in this domain.

TABLE 4.7: Progressive fine-tuning pipeline results on the Raabin-WBC test set. BACC = balanced accuracy (primary metric), all other metrics are macro-averaged. Δ denotes absolute improvement over the previous stage. Best performance shown in **bold**.

Stage	Description	BACC	Δ BACC	AUC	F1
Ce1	Zero-shot baseline	0.3088	—	0.6423	0.1538
Ce2	Frozen single-scale	0.9343	+0.6255	0.9959	0.8908
Ce3	Frozen multi-scale	0.9477	+0.0134	0.9973	0.8930
Ce4	Partial fine-tuning (3 blocks)	0.9507	+0.0030	0.9978	0.8909
Ce4A	6 blocks (vs Ce4)	0.9554	+0.0047	0.9982	0.9045
Ce4B	Sequential 3+3 (vs Ce4A)	0.9653	+0.0099	0.9985	0.9288
Ce5	Learnable logit scale	0.9661	+0.0008	0.9987	0.9180
Ce6	Soft prompt tuning (vs Ce5)	0.9669	+0.0008	0.9987	0.9291
Ce7	MoE projection head (vs Ce5)	**0.9745**	+0.0076	**0.9989**	**0.9383**

Figure 4.17 shows the confusion matrix for Ce7 on Raabin-WBC. Performance is strong across all five cell categories, with Basophil achieving 1.000 recall and Eosinophil reaching 0.988. Neutrophils showed the lowest per-class recall at 0.955, with minor confusion with Lymphocytes, consistent with the morphological overlap between these cell types. The near-perfect classification across all classes suggests that the pipeline effectively captures the morphological differences between white blood cell types.

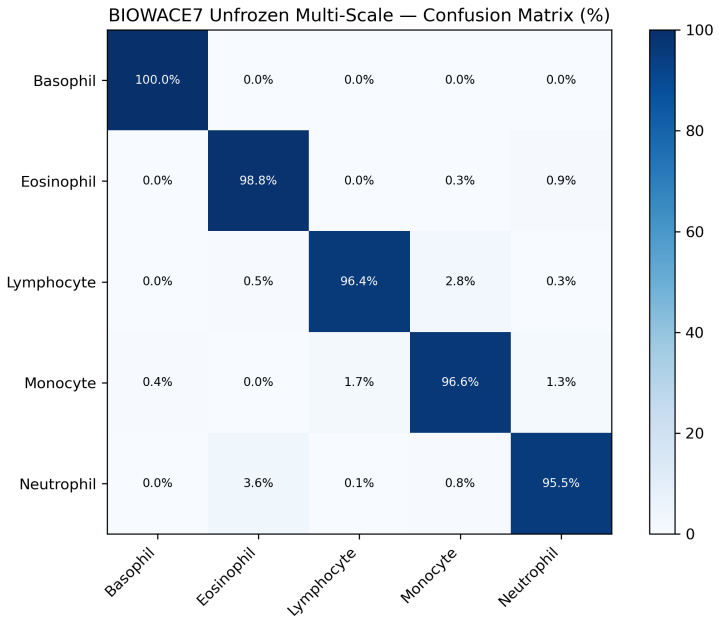


FIGURE 4.17: Confusion matrix for Ce7 on the Raabin-WBC test set. All five cell categories achieve recall above 95%, with Basophil reaching 100% and Neutrophil showing the lowest recall at 95.5%.

4.3.3 Cross-Dataset Generalisation Discussion

The cross-dataset results reveal distinct patterns of pipeline behaviour across the three evaluated domains. On Raabin-WBC, the final model consistently performed strongly across all stages, with progressive improvements from Ce1 through Ce7, and the final Ce7 configuration achieved 0.9745 BACC, the highest across all three datasets. On ISIC 2019, performance plateaued at Ce4A (0.7546 BACC) with subsequent stages showing marginal degradation, suggesting that the visual complexity and severe class imbalance of dermoscopic classification limit the benefit of further adaptation under the current configuration. On CELLo, Ce5 yielded the best configuration, with improvement driven primarily by logit-scale recalibration.

The differences in which pipeline stage achieved peak performance across datasets, Ce4A on ISIC, Ce5 on CELLo, and Ce7 on Raabin, indicate that the optimal adaptation depth is domain-dependent. Additionally, Ce4B outperformed Ce4A on Raabin but not on CELLo or ISIC. This single case suggests sequential unfreezing may benefit certain domains. These findings suggest that the progressive fine-tuning framework generalises effectively across imaging domains, though the extent and nature of adaptation required vary with the visual characteristics of the target task.

4.4 Pipeline Summary and Component Analysis

This section provides an aggregated summary of the experimental findings, consolidating the contribution of each pipeline component on the primary CELLo dataset and the generalisation performance across all three evaluated domains.

4.4.1 Component Contributions on CELLo

Table 4.8 summarises the absolute contribution of each architectural component introduced across the progressive fine-tuning pipeline on the primary CELLo dataset, measured as the gain in balanced accuracy over the previous stage.

TABLE 4.8: Summary of architectural component contributions to balanced accuracy on the CELLo test set. Δ BACC denotes absolute gain over the previous stage. The final Ce5 configuration uses the optimised learning rate ($LR_{backbone} = 1 \times 10^{-4}$).

Component	Description	BACC	Δ BACC
Zero-shot baseline (Ce1)	No training; pretrained alignment only	0.2033	—
Projection head (Ce2)	Task-specific head on frozen backbone	0.8407	+0.6374
Multi-scale CLS fusion (Ce3)	CLS tokens from blocks 3, 7, 11	0.9071	+0.0664
Partial fine-tuning (Ce4A)	Top 6 blocks unfrozen with LLRD	0.9317	+0.0246
Logit scale calibration (Ce5)	Learnable temperature parameter	0.9334	+0.0017
Learning rate optimisation	LR search on Ce5	0.9404	+0.0070
Total gain	**Ce1 to final Ce5**	**0.9404**	**+0.7371**

The results reveal a clear hierarchy of component contributions. The dominant gain arises from the task-specific projection head (+0.6374), confirming that the pretrained BiomedCLIP representations already encode sufficient structure for effective downstream classification without backbone modification. Multi-scale CLS fusion provides the second largest contribution (+0.0664), demonstrating that intermediate transformer representations capture discriminative morphological information not present in the final pooled output alone. Partial backbone fine-tuning contributes a further +0.0246 across Ce4 and Ce4A, with deeper unfreezing providing diminishing but consistent returns. Logit scale calibration (+0.0017) and learning rate optimisation (+0.0070) provide smaller but consistent refinements, confirming that the architecture is largely saturated by Ce4A and that subsequent gains reflect fine-grained recalibration rather than structural improvements. Architectural extensions Ce6 and Ce7 did not improve upon Ce5 on the primary CELLo

dataset, suggesting that text-side adaptation and conditional expert routing provide no additional benefit once visual adaptation has been fully optimised.

4.5 Chapter Summary

This chapter presented the experimental results obtained from evaluating the proposed progressive fine-tuning pipeline across three datasets. On the primary CELLo dataset, the pipeline demonstrated consistent improvement from zero-shot baseline to Ce5, achieving a final balanced accuracy of 0.9404 ± 0.0002 under stability evaluation. Attention map analysis confirmed that the selected BiomedCLIP backbone maintains consistent spatial focus on morphologically relevant cellular structures throughout the pipeline. Cross-dataset evaluation on ISIC 2019 and Raabin-WBC demonstrated that the pipeline generalises effectively across imaging domains, with Raabin-WBC showing strong and consistent improvement through all stages and ISIC 2019 reaching peak performance at Ce4A. The domain-dependent variation in optimal adaptation depth highlights both the flexibility and the context-sensitivity of the progressive fine-tuning approach.

Chapter 5

Conclusions

This chapter presents the conclusions drawn from the experimental findings and outlines limitations and directions for future research.

5.1 Summary of Experimental Findings

This thesis presented a comprehensive evaluation of a progressive fine-tuning framework applied to vision-language foundation models across three distinct medical imaging domains, adapting a BiomedCLIP backbone using cellular morphology data from CELLo, dermoscopic images from ISIC 2019, and haematological samples from Raabin-WBC. The experimental results demonstrate that while a unified, step-wise adaptation pipeline provides substantial performance gains over zero-shot baselines across all evaluated tasks, the optimal adaptation depth and the benefit of downstream architectural extensions are highly domain-dependent.

On the primary CELLo dataset for tumour staging, the pipeline achieved stable convergence at the refined Ce5 configuration, with a final balanced accuracy of 0.9404 ± 0.0002 across three independent random seeds. This low seed-to-seed variance, particularly when contrasted against run-to-run variances of up to 0.005 observed during intermediate pipeline trials, highlights robust convergence behaviour under the selected configuration. Per-class analysis confirmed strong class separability, with Ta, T2, and T4 stages achieving AUC of 1.00, while the largest residual misclassifications occurred between T3 and NA, which also showed the lowest per-class recall at 0.906.

Evaluation on the ISIC 2019 dataset revealed an early performance plateau driven by severe class imbalance and high intra-class visual diversity. Peak performance was

achieved at stage Ce4A with a balanced accuracy of 0.7546 and macro AUC of 0.9589, with the most substantial gains contributed by multi-scale feature fusion at Ce3 and partial backbone adaptation at Ce4A. Subsequent stages, including learnable logit scale recalibration and architectural extensions, did not improve performance on this dataset, suggesting that the domain's performance ceiling had been reached under the fixed configuration.

Conversely, the framework achieved its highest overall performance on the Raabin-WBC dataset, reaching Ce7 with a balanced accuracy of 0.9745 and macro F1 of 0.9383. Unlike the other two domains, performance improved consistently through every pipeline stage, including sequential block unfreezing where Ce4B outperformed Ce4A by an absolute gain of 0.0099, soft prompt tuning at Ce6, and MoE routing at Ce7, ultimately achieving near-perfect discrimination across morphologically distinct cell categories, including 100% recall for Basophils.

5.2 Key Contributions and Theoretical Takeaways

The results of this work challenge the assumption that a single fixed adaptation strategy is sufficient across diverse biomedical vision-language tasks. The distinct variations in peak performance across evaluation domains indicate that the optimal unfreezing depth and choice of parameter-efficient extensions depend on the structural properties of the target domain. Raabin-WBC benefited consistently from additional adaptation capacity, including soft prompt tuning and MoE routing, whereas ISIC 2019, characterised by severe class imbalance and high intra-class visual diversity, reached its performance ceiling earlier during backbone adaptation and showed no benefit from subsequent extensions. This divergence may relate to differences in class morphology between the domains.

A key design decision in this work was selecting BiomedCLIP over OpenCLIP ViT-B/16 as the primary backbone, despite the latter achieving marginally higher performance both in the zero-shot setting and under a preliminary version of the pipeline. This decision was grounded in qualitative analysis of attention maps from preliminary fine-tuning evaluations, which revealed that OpenCLIP exhibited systematic activation in background and non-cellular regions across tumour stages, raising concerns about the clinical reliability of its predictions. BiomedCLIP, by contrast, demonstrated consistent

spatial focus on morphologically relevant cellular structures. This finding motivated prioritising interpretability alongside classification performance, reflecting the clinical requirements of a diagnostic support tool.

A further contribution is the qualitative analysis of attention maps across correct and incorrect classifications to assess the spatial grounding of model predictions. Attention maps across all five tumour stages in the CELLo dataset confirmed that successful classifications correlate with focused spatial activations concentrated on morphologically relevant regions, specifically the cell body, nuclear structures, and cytoplasmic regions. Misclassifications exhibited more diffuse and peripherally distributed attention patterns, suggesting that classification errors reflect genuine morphological ambiguity rather than systematic reliance on background or non-cellular cues. This finding supports the clinical interpretability of the selected configuration.

The introduction of a learnable logit scale at Ce5 not only improved classification performance modestly but also increased training stability. The variance across three independent seeds at Ce5 (± 0.0002) was substantially lower than the run-to-run variance observed at earlier stages, suggesting that temperature recalibration following backbone adaptation reduces sensitivity to random initialisation, providing a stabilising effect beyond its direct contribution to performance.

The ablation studies conducted during pipeline development further demonstrated the importance of layer-wise learning rate decay, which with a decay factor of 0.80 produced more stable training compared to uniform backbone learning rates, consistent with established practice in partial fine-tuning of vision transformers.

5.3 Limitations

Several limitations of the present work should be acknowledged. Intermediate pipeline stage selection was based on single training runs due to computational constraints, meaning that decisions such as the choice of Ce4A over Ce4B as the base for subsequent stages may be affected by run-to-run variance even though in most cases differences were higher than expected variance. Stability evaluation was conducted only on the final selected Ce5 configuration on the CELLo dataset, and was not extended to earlier stages or to the cross-dataset experiments.

The current evaluation is further limited to five tumour stages (NA, Ta, T2, T3, and T4). At the time this work was carried out, the T1 cell line data were still under-sampled

and undergoing modification and validation, and were therefore excluded to avoid inconsistencies across model experiments. Future work will include the T1 class without major difficulties, thereby enabling complete TNM staging and increasing the clinical applicability of the proposed pipeline.

No supervised baseline comparison was conducted on the CELLo dataset, as no published supervised methods evaluated on the same dataset and classification task are available in the literature at the time of writing. Performance figures, therefore, represent absolute rather than relative improvements, and direct comparison with alternative approaches is not possible.

Cross-dataset evaluation was performed without any domain-specific hyperparameter optimisation or fine-tuning, providing a direct assessment of generalisation under distribution shift but not a measure of what the pipeline could achieve with domain-specific adaptation. Results on ISIC 2019 and Raabin-WBC should therefore be interpreted as generalisation benchmarks rather than as competitive performance figures for those domains.

5.4 Future Research Directions

While the progressive fine-tuning pipeline demonstrated flexibility and strong performance across diverse imaging domains, several directions remain open for future investigation. To address the domain-dependent performance ceilings observed across tasks, future work could explore adaptive unfreezing strategies that dynamically determine the appropriate adaptation depth based on validation performance trajectories, rather than applying a fixed pipeline uniformly across datasets.

To reduce confusion between class pairs that are frequently misclassified, such as T3 and NA in cellular staging or melanoma and melanocytic nevus in dermoscopy, contrastive objectives with explicit inter-class separation penalties could be incorporated alongside the standard classification loss. Additionally, incorporating spatial boundary information during multi-scale feature fusion may improve discrimination of subtle cell wall and lesion border variations.

Finally, extending this framework toward real-world clinical deployment requires evaluation beyond offline test sets, assessing how model confidence scores and attention map visualisations influence diagnostic speed, user confidence, and error rates among practising pathologists and clinicians when integrated into digital workflow interfaces.

5.5 Chapter Summary

This chapter concluded the thesis by synthesising the experimental findings, contributions, limitations, and future directions of the progressive fine-tuning pipeline. The framework demonstrated strong performance and stable convergence, achieving balanced accuracy of 0.9404 on CELLo, 0.7546 on ISIC 2019, and 0.9745 on Raabin-WBC. Attention map analysis confirmed that the model maintains a clinically grounded focus on morphologically relevant structures, while the domain-dependent variation in optimal adaptation depth highlights both the flexibility and the context-sensitivity of the progressive fine-tuning approach. The explicit acknowledgement of limitations regarding single-run stage selection, the absence of supervised baselines, and the scope of stability evaluation provides a transparent characterisation of the boundaries of the current work.

Bibliography

Faruk Ahmed, Andrew Sellergren, Lin Yang, Shawn Xu, Boris Babenko, Abbi Ward, Niels Olson, Arash Mohtashamian, Yossi Matias, Greg S. Corrado, Quang Duong, Dale R. Webster, Shravya Shetty, Daniel Golden, Yun Liu, David F. Steiner, and Ellery Wulczyn. PathAlign: A vision-language model for whole slide images in histopathology. *arXiv preprint arXiv:2406.19578*, 2024. doi: 10.48550/arXiv.2406.19578. URL https://arxiv.org/abs/2406.19578. [Cited on pages 27, 29, and 33.]

Hussien Al-Asi, Ibrahim Yilmaz, Jordan Reynolds, Shweta Agarwal, Aziza Nassar, Abba Zubair, Craig Horbinski, Bryan Dangott, and Zeynettin Akkus. Pathology foundation models: Evolution, current landscape, challenges and opportunities from a technical and clinical perspective. *Bioengineering*, 13(5):577, 2026. doi: 10.3390/bioengineering13050577. [Cited on pages 24, 26, 31, and 33.]

Hasan Al-Sattar, Hao Ding, Oluchi Okoli, Somto Okoli, Aruni Ghose, Giuseppe Luigi Banna, Simon Wan, Athar Haroon, Jonathan Wong, Jeremy Teoh, Nikhil Vasdev, Eleni Efstathiou, Stergios Boussios, and Sola Adeleke. A multi-modal approach for decision making in bladder cancer. *Nature Reviews Urology*, 2026. doi: 10.1038/s41585-025-01122-7. [Cited on pages 7 and 32.]

Saghir Alfasly, Peyman Nejat, Sobhan Hemati, Jibran Khan, Isaiah Lahr, Areej Alsaafin, Abubakr Shafique, Nneka Comfere, Dennis Murphree, Chady Meroueh, Saba Yasir, Aaron Mangold, Lisa Boardman, Vijay H. Shah, Joaquin J. Garcia, and H. R. Tizhoosh. Foundation models for histopathology—fanfare or flair. *Mayo Clinic Proceedings: Digital Health*, 2(1):165–174, 2024. doi: 10.1016/j.mcpdig.2024.02.003. [Cited on pages 24, 25, 30, and 33.]

Omar Alqahtani, Mohamed Ghouse, Asfia Sabahath, Omer Bin Hussain, and Arshiya Begum. Multi-scale vision transformer with dynamic multi-loss function for medical

image retrieval and classification. *Computers, Materials & Continua*, 83(2):2221–2244, 2025. doi: 10.32604/cmc.2025.061977. [Cited on page 44.]

Ruqayya Awan, Ksenija Benes, Ayesha Azam, Tzu-Hsi Song, Muhammad Shaban, Clare Verrill, Yee Wah Tsang, David Snead, Fayyaz Minhas, and Nasir Rajpoot. Deep learning based digital cell profiles for risk stratification of urine cytology images. *Cytometry Part A*, 99(7):732–742, 2021. doi: 10.1002/cyto.a.24313. [Cited on pages 7, 8, and 32.]

Marko Babjuk, Maximilian Burger, Otakar Capoun, Daniel Cohen, Eva M. Compérat, José L. Dominguez Escrig, Paolo Gontero, Fredrik Liedberg, Alexandra Masson-Lecomte, A. Hugh Mostafid, Joan Palou, Bas W. G. van Rhijn, Morgan Rouprêt, Shahrokh F. Shariat, Thomas Seisen, Viktor Soukup, and Richard J. Sylvester. European association of urology guidelines on non-muscle-invasive bladder cancer (Ta, T1, and carcinoma in situ). *European Urology*, 81(1):75–94, 2022. doi: 10.1016/j.eururo.2021.08.010. [Cited on pages 7, 30, and 32.]

Mohsin Bilal, Aadam, Manahil Raza, Youssef Altherwy, Anas Alsuhaibani, Abdulrahman Abduljabbar, Fahdah Almarshad, Paul Golding, and Nasir Rajpoot. Foundation models in computational pathology: A review of challenges, opportunities, and impact. arXiv preprint arXiv:2502.08333, 2025. [Cited on page 4.]

Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Isabelle Soerjomataram, and Ahmedin Jemal. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74(3):229–263, 2024. doi: 10.3322/caac.21834. [Cited on pages 1, 6, and 32.]

Cancer Research UK. Types and grades of bladder cancer. https://www.cancerresearchuk.org/about-cancer/bladder-cancer/types-stages-grades/types, 2026. Accessed: June 2026. [Cited on pages 6 and 32.]

Kang-Yu Chang, Chi-Shun Yang, Jing-Yi Lai, Shu-Jiuan Lin, Jian-Ri Li, Tien-Jen Liu, Wei-Lei Yang, Ming-Yu Lin, Cheng-Hung Yeh, Shih-Wen Hsu, et al. An artificial intelligence-assisted diagnostic system improves upper urine tract cytology diagnosis. *in vivo*, 38 (6):3016–3021, 2024. [Cited on pages 20, 21, 31, and 33.]

Cheng-Che Chen, Tsung-Han Yen, Jian-Ri Li, Chih-Jung Chen, Chi-Shun Yang, Jing-Yi Lai, Shu-Jiuan Lin, Cheng-Hung Yeh, Shih-Wen Hsu, Ming-Yu Lin, et al. Artificial intelligence algorithms enhance urine cytology reporting confidence in postoperative follow-up for upper urinary tract urothelial carcinoma. *International urology and nephrology*, 57 (3):801–808, 2025. [Cited on pages 20, 21, and 33.]

Zhihong Chen, Shizhe Diao, Benyou Wang, Guanbin Li, and Xiang Wan. Towards unifying medical vision-and-language pre-training via soft prompts. arXiv preprint arXiv:2302.08958, 2023. [Cited on page 47.]

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. doi: 10.1109/CVPR52729.2023.00276. [Cited on pages 41 and 42.]

Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022. doi: 10.1016/j.mlwa.2021.100198. [Cited on pages 4, 23, 25, 31, and 33.]

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *arXiv preprint arXiv:1902.03368*, 2019. doi: 10.48550/arXiv.1902.03368. [Cited on pages 35 and 37.]

Marc Combalia, Noel Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C. Halpern, Susana Puig, and Josep Malvehy. BCN20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*, 2019. doi: 10.48550/arXiv.1908.02288. [Cited on page 37.]

Dawei Dai, Yuanhui Zhang, Qianlan Yang, Long Xu, Xiaojing Shen, Shuyin Xia, and Guoyin Wang. Pathologyvlm: a large vision-language model for pathology image understanding. *Artificial Intelligence Review*, 58:186, 2025. doi: 10.1007/s10462-025-11190-1. [Cited on pages 28, 29, and 33.]

Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. PubMedCLIP: How much does CLIP benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1181–1193, 2023. doi: 10.18653/v1/2023.findings-eacl.88. [Cited on page 42.]

Matteo Ferro, Ugo Giovanni Falagario, Biagio Barone, Martina Maggi, Felice Crocetto, Gian Maria Busetto, Francesco Del Giudice, Daniela Terracciano, Giuseppe Lucarelli, Francesco Lasorsa, Michele Catellani, Antonio Brescia, Francesco Alessandro Mistretta, Stefano Luzzago, Mattia Luca Piccinelli, Mihai Dorin Vartolomei, Barbara Alicja Jereczek-Fossa, Gennaro Musi, Emanuele Montanari, Ottavio de Cobelli, and Octavian Sabin Tataru. Artificial intelligence in the advanced diagnosis of bladder cancer—comprehensive literature review and future advancement. *Diagnostics*, 13(13):2308, 2023. doi: 10.3390/diagnostics13132308. [Cited on pages 7 and 32.]

Nils Gessert, Maximilian Nielsen, Mohsin Shaikh, René Werner, and Alexander Schlaefer. Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data. *MethodsX*, 7:100864, 2020. doi: 10.1016/j.mex.2020.100864. [Cited on pages 8 and 32.]

Daniele Giansanti. The augmented cytopathologist: A conceptual exploratory narrative review on immersive and vision–language models tools in digital pathology. *Journal of Imaging*, 12(3):100, 2026. doi: 10.3390/jimaging12030100. [Cited on pages 28, 29, 31, and 33.]

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 2021. doi: 10.1145/3458754. [Cited on page 42.]

Chinami Hoshino, Sayaka Kobayashi, Yoshimi Nishijima, Seiji Arai, Kazuhiro Suzuki, and Masanao Saio. Computer-assisted scatter plot analysis of cell and nuclear areas distinguishes urothelial carcinoma in urine cytology specimens. *CytoJournal*, 22:12, 2025. doi: 10.25259/Cytojournal_213_2024. [Cited on pages 20, 21, and 33.]

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J. Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical Twitter. *Nature Medicine*, 29:2307–2316, 2023. doi: 10.1038/s41591-023-02504-3. [Cited on pages 27, 28, 31, 33, and 42.]

Fabian Hörst, Moritz Rempe, Lukas Heine, Constantin Seibold, Eric Hasselberg, Jakob Dexl, Fin Bayer, Markus Kargl, Bernhard Kainz, Matthias M. Gaida, Philipp Gerber, Stephan Ferber, Moritz Schwartz, Yoni Schirris, Marco Meneghetti, Bastian Leibe, Christoph Baur, and Anirban Mukhopadhyay. CellViT: Vision transformers for precise cell segmentation and classification with a SAM-inspired encoder. *Medical Image Analysis*, 94:103143, 2024. doi: 10.1016/j.media.2024.103143. [Cited on pages 24, 25, 31, and 33.]

Wisdom Oluchi Ikezogwo, Mehmet Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Stefan Chan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda G. Shapiro. Quilt-1M: One million image-text pairs for histopathology. In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., 2023. doi: 10.48550/arXiv.2306.11207. [Cited on pages 27, 28, and 33.]

Vedrana Ivezić, Ashwath Radhachandran, Ekaterina Redekop, Shreeram Athreya, Dongwoo Lee, Vivek Sant, Corey Arnold, and William Speier. CytoFM: The first cytology foundation model. *arXiv preprint arXiv:2504.13402*, 2025. doi: 10.48550/arXiv.2504.13402. URL https://arxiv.org/abs/2504.13402. [Cited on pages 24, 25, and 33.]

Zexuan Ji, Zheng Chen, and Xiao Ma. Grouped multi-scale vision transformer for medical image segmentation. *Scientific Reports*, 15:11122, 2025. doi: 10.1038/s41598-025-95361-8. [Cited on page 44.]

Hao Jiang, Yanning Zhou, Yi Lin, Ronald CK Chan, Jiang Liu, and Hao Chen. Deep learning for computational cytology: A survey. *Medical Image Analysis*, 84:102691, 2023. doi: 10.1016/j.media.2022.102691. [Cited on pages 11 and 12.]

Yufeng Jiang and Yiqing Shen. M4oE: A foundation model for medical multimodal image segmentation with mixture of experts. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, 2024. doi: 10.1007/978-3-031-72390-2_58. [Cited on page 48.]

Ashish M. Kamat, Noah M. Hahn, Jason A. Efstathiou, Seth P. Lerner, Per-Uno Malmström, Woonyoung Choi, Charles C. Guo, Yair Lotan, and Wassim Kassouf. Bladder cancer. *The Lancet*, 388(10061):2796–2810, 2016. doi: 10.1016/S0140-6736(16)30512-8. [Cited on pages 6 and 32.]

Masatomo Kaneko, Keisuke Tsuji, Keiichi Masuda, Kengo Ueno, Kohei Henmi, Shota Nakagawa, Ryo Fujita, Kensho Suzuki, Yuichi Inoue, Satoshi Teramukai, et al. Urine cell image recognition using a deep-learning model for an automated slide evaluation system. *BJU international*, 130(2):235–243, 2022. doi: 10.1111/bju.15518. [Cited on pages 3, 15, 30, and 32.]

Farbod Khoraminia, Saul Fuster, Neel Kanwal, Mitchell Olislagers, Kjersti Engan, Geert JLH van Leenders, Andrew P Stubbs, Farhan Akram, and Tahlita CM Zuiverloon. Artificial intelligence in digital pathology for bladder cancer: hype or hope? a systematic review. *Cancers*, 15(18):4518, 2023. doi: 10.3390/cancers15184518. [Cited on pages 11 and 12.]

Zahra Mousavi Kouzehkanan, Sepehr Saghari, Sajad Tavakoli, Peyman Rostami, Mohammadjavad Abaszadeh, Farzaneh Mirzadeh, Esmaeil Shahabi Satlsar, Maryam Gheidishahran, Fatemeh Gorgi, Saeed Mohammadi, and Reshad Hosseini. A large dataset of white blood cells containing cell locations and types, along with segmented nuclei and cytoplasm. *Scientific Reports*, 12:1123, 2022. doi: 10.1038/s41598-021-04426-x. [Cited on pages 35 and 39.]

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. [Cited on pages 9, 43, and 45.]

Joshua J Levy, Natt Chan, Jonathan D Marotti, Darcy A Kerr, Edward J Gutmann, Ryan E Glass, Caroline P Dodge, Arief A Suriawinata, Brock C Christensen, Xiaoying Liu, et al. Large-scale validation study of an improved semiautonomous urine cytology assessment tool: Autoparis-x. *Cancer cytopathology*, 131(10):637–654, 2023a. [Cited on pages 19, 21, 31, and 33.]

Joshua J Levy, Natt Chan, Jonathan D Marotti, Nathalie J Rodrigues, A Aziz O Ismail, Darcy A Kerr, Edward J Gutmann, Ryan E Glass, Caroline P Dodge, Arief A Suriawinata, et al. Examining longitudinal markers of bladder cancer recurrence through a

semiautonomous machine learning system for quantifying specimen atypia from urine cytology. *Cancer cytopathology*, 131(9):561–573, 2023b. [Cited on pages 20, 21, 31, and 33.]

Joshua J Levy, Xiaoying Liu, Jonathan D Marotti, Darcy A Kerr, Edward J Gutmann, Ryan E Glass, Caroline P Dodge, Arief A Suriawinata, and Louis J Vaickus. Uncovering additional predictors of urothelial carcinoma from voided urothelial cell clusters through a deep learning–based image preprocessing technique. *Cancer cytopathology*, 131(1):19–29, 2023c. [Cited on pages 20, 21, and 33.]

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017. doi: 10.1016/j.media.2017.07.005. [Cited on pages 8 and 32.]

Tien-Jen Liu, Wen-Chi Yang, Shin-Min Huang, Wei-Lei Yang, Hsing-Ju Wu, Hui Wen Ho, Shih-Wen Hsu, Cheng-Hung Yeh, Ming-Yu Lin, Yi-Ting Hwang, et al. Evaluating artificial intelligence–enhanced digital urine cytology for bladder cancer diagnosis. *Cancer cytopathology*, 132(11):686–695, 2024. [Cited on pages 20, 21, 31, and 33.]

Yixiao Liu, Shen Jin, Qi Shen, Lufan Chang, Shancheng Fang, Yu Fan, Hao Peng, and Wei Yu. A deep learning system to predict the histopathological results from urine cytopathological images. *Frontiers in Oncology*, 12:901586, 2022. doi: 10.3389/fonc.2022.901586. [Cited on pages 3, 15, 16, 30, and 32.]

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30:863–874, 2024. doi: 10.1038/s41591-024-02856-4. [Cited on pages 27, 29, and 33.]

Xiaoyu Ma, Qiuchen Zhang, Lvqi He, Xinyang Liu, Yang Xiao, Jingwen Hu, Shengjie Cai, Hongzhou Cai, and Bin Yu. Artificial intelligence application in the diagnosis and treatment of bladder cancer: advance, challenges, and opportunities. *Frontiers in Oncology*, 14:1487676, 2024. doi: 10.3389/fonc.2024.1487676. [Cited on pages 7 and 32.]

Ewen McAlpine, Pamela Michelow, Eric Liebenberg, and Turgay Celik. Is it real or not? toward artificial intelligence-based realistic synthetic cytology image generation to augment teaching and quality assurance in pathology. *Journal of the American Society of Cytopathology*, 11(3):123–132, 2022. doi: 10.1016/j.jasc.2022.02.001. [Cited on pages 23, 25, and 33.]

Ewen McAlpine, Pamela Michelow, Eric Liebenberg, and Turgay Celik. Are synthetic cytology images ready for prime time? a comparative assessment of real and synthetic urine cytology images. *Journal of the American Society of Cytopathology*, 12(2):126–135, 2023. doi: 10.1016/j.jasc.2022.10.001. [Cited on pages 23, 25, and 33.]

Graham Mowatt, James N'Dow, Luke Vale, Ghulam Nabi, Charles Boachie, Jonathan A. Cook, Cynthia Fraser, and T. R. Leyshon Griffiths. Photodynamic diagnosis of bladder cancer compared with white light cystoscopy: systematic review and meta-analysis. *International Journal of Technology Assessment in Health Care*, 27(1):3–10, 2011. doi: 10.1017/S0266462310001364. [Cited on pages 7, 30, and 32.]

Fatima Nabiyouni and Paul Z Chiou. Enhancing urine cytopathology with artificial intelligence: a systematic review. *Am J Clin Pathol*, 165(2):aqaf135, 2026. doi: 10.1093/ajcp/aqaf135. [Cited on pages 11 and 12.]

Nima Narimani, Mohammad Mehdi Atarod, Ehsan Zolfi, Reza Shahrokhi-Damavand, Hadi Maleki, Ali Saberi, Amirhosein Naseri, Amirhossein Rahmani, Amirhossein Shahbazi, Kazem Aghili, et al. Artificial intelligence in the diagnosis and pathological assessment of bladder cancer: current evidence and clinical applications. *African Journal of Urology*, 32(1):35, 2026. doi: 10.1186/s12301-026-00578-2. [Cited on pages 11 and 13.]

Nao Nitta, Yuko Sugiyama, Takeaki Sugimura, Takahiko Ito, Koichi Ikebata, Hitoshi Abe, Shuhei Ishii, Nagisa Hosoya, Raihan Ull Islam, Aditya Jain, Meisam Hasani, Joseph Zonghi, Peter Koh, Yukihito Mase, Miki Kanematsu, Noureldin M. Z. Ali, Yoshihiko Murata, Ayumi Shikama, Yusuke Kobayashi, Daisuke Matsubara, Yukari Himeji, Hiroshi Nakamura, Akane Hashizume, Miyaka Umemori, Hiroyuki Ohsaki, Yingdong Luo, Tianben Ding, Fernando C. Schmitt, Robert Y. Osamura, Keisuke Goda, and Tomohiro Chiba. Clinical-grade autonomous cytopathology through whole-slide edge tomography. *Nature*, 651:472–481, 2026. doi: 10.1038/s41586-025-10094-y. [Cited on pages 3 and 32.]

Yanglan Ou, Ye Yuan, Xiaolei Huang, Stephen T.C. Wong, John Volpi, James Z. Wang, and Kelvin Wong. Patcher: Patch transformers with mixture of experts for precise medical image segmentation. arXiv preprint arXiv:2206.01741, 2022. [Cited on page 47.]

Alper Özcan and Ceylan Öztürk. Comprehensive data analysis of white blood cells with classification and segmentation by using deep learning approaches. *Cytometry Part A*, 2024. doi: 10.1002/cyto.a.24839. [Cited on pages 8 and 32.]

Liangrui Pan, Zhichao Feng, and Shaoliang Peng. A review of machine learning approaches, challenges and prospects for computational tumor pathology. *arXiv preprint arXiv:2206.01728*, 2022. doi: 10.48550/arXiv.2206.01728. [Cited on page 11.]

Jiuming Qin, Che Liu, Sibo Cheng, Yike Guo, and Rossella Arcucci. Freeze the backbones: A parameter-efficient contrastive approach to robust medical vision-language pre-training. In *ICASSP 2024: IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1686–1690, 2024. doi: 10.1109/ICASSP48485.2024.10447326. [Cited on page 44.]

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. doi: 10.48550/arXiv.2103.00020. URL https://arxiv.org/abs/2103.00020. [Cited on pages x, 4, 40, 41, and 45.]

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems*, volume 32, 2019. [Cited on pages 8 and 32.]

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. In *Advances in Neural Information Processing Systems*, volume 34, pages 8583–8595, 2021. doi: 10.48550/arXiv.2106.05974. [Cited on page 47.]

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. doi: 10.48550/arXiv.1701.06538. [Cited on page 52.]

Runnan Shen, Fan Jiang, Xiaowei Huang, Guibin Hong, Yun Luo, Huan Wan, Ye Xie, Mengyi Zhu, Yun Wang, Bohao Liu, et al. Artificial intelligence-assisted urine cytology for noninvasive detection of muscle-invasive urothelial carcinoma: A multi-center diagnostic study with prospective validation. *Advanced Science*, 12(39):e08977, 2025. doi: 10.1002/advs.202508977. [Cited on pages 17, 18, and 33.]

Michael Superdock, Sara E Wobker, Iain Carmicheal, and William Y Kim. Computational pathology in bladder cancer: A scoping review. *Bladder Cancer*, 12(1): 23523735251413333, 2026. doi: 10.1177/23523735251413333. [Cited on pages 11 and 12.]

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P. Langlotz, Andrew Y. Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1406, 2022. doi: 10.1038/s41551-022-00936-9. [Cited on pages 9 and 32.]

I. Tosoni, U. Wagner, G. Sauter, M. Egloff, H. Knönagel, G. Alund, F. Bannwart, M. J. Mihatsch, T. C. Gasser, and R. Maurer. Clinical significance of interobserver differences in the staging and grading of superficial bladder cancer. *BJU International*, 85(1):48–53, 2000. doi: 10.1046/j.1464-410x.2000.00356.x. [Cited on pages 7, 30, and 32.]

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5:180161, 2018. doi: 10.1038/sdata.2018.161. [Cited on page 37.]

Masayuki Tsuneki, Makoto Abe, and Fahdi Kanavati. Deep learning-based screening of urothelial carcinoma in whole slide images of liquid-based cytology urine specimens. *Cancers*, 15(1):226, 2022. doi: 10.3390/cancers15010226. [Cited on pages 3, 17, 18, and 32.]

Shuoyuan Wang, Jindong Wang, Guoqing Wang, Bob Zhang, Kaiyang Zhou, and Hongxin Wei. Open-vocabulary calibration for fine-tuned CLIP. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 51734–51754, 2024. doi: 10.48550/arXiv.2402.04655. [Cited on page 45.]

Shaoxu Wu, Runnan Shen, Guibin Hong, Yun Luo, Huan Wan, Jiaming Feng, Zhaojun Chen, Fan Jiang, Yun Wang, Caixia Liao, Xiaoxue Li, Bohao Liu, Xiaowei Huang,

Ke Liu, Peiqi Qin, Yun Wang, Ye Xie, Nuo Ouyang, Jian Huang, and Tianxin Lin. Development and validation of an artificial intelligence-based model for detecting urothelial carcinoma using urine cytology images: a multicentre, diagnostic study with prospective validation. *EClinicalMedicine*, 71:102566, 2024. doi: 10.1016/j.eclinm.2024.102566. [Cited on pages 3, 17, 18, 31, and 32.]

Lei Xiong, Xinyi Cao, Yu Shang, Zongyue Lu, Hao Jiang, Chang Shi, Chengzhi Zhang, Zhongjing Ma, Lili Tian, Xiaojie Wang, et al. Interpretable machine learning for urothelial cells classification and risk scoring in urine cytology. *iScience*, 29(1), 2026. doi: 10.1016/j.isci.2025.114259. [Cited on pages 20, 22, and 33.]

Wei-Lei Yang, Barbara A Crothers, Tien-Jen Liu, Shih-Wen Hsu, Cheng-Hung Yeh, Yi-Siou Liu, Guowei Shao, Ming-Yu Lin, Tang-Yi Tsao, Min-Che Tung, et al. Does artificial intelligence redefine nuclear-to-cytoplasmic ratio threshold for diagnosing high-grade urothelial carcinoma? *Cancer cytopathology*, 133(5):e70017, 2025. doi: 10.1002/cncy.70017. [Cited on pages 15, 16, 30, and 32.]

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Matthew P. Lungren, Tristan Naumann, and Hoifung Poon. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2023. doi: 10.48550/arXiv.2303.00915. URL https://arxiv.org/abs/2303.00915. [Cited on pages 4, 28, 29, 31, 33, and 42.]

Beidi Zhao, Wenlong Deng, Zi Han Henry Li, Chen Zhou, Zuhua Gao, Gang Wang, and Xiaoxiao Li. Less: Label-efficient multi-scale learning for cytological whole slide image screening. *Medical Image Analysis*, 94:103109, 2024. doi: 10.1016/j.media.2024.103109. [Cited on pages 17, 18, 31, and 33.]

Kaipeng Zheng, Weiran Huang, and Lichao Sun. TransMed: Large language models enhance vision transformer for biomedical image classification. arXiv preprint arXiv:2312.07125, 2023. [Cited on page 44.]

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. doi: 10.1007/s11263-022-01653-1. [Cited on pages 46 and 51.]