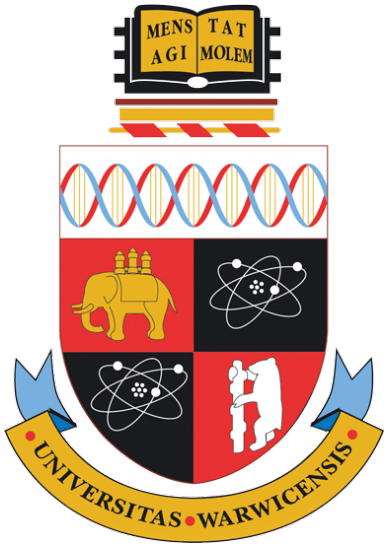
WARWICK
THE UNIVERSITY OF WARWICK



Learning with Minimal Annotations in Computational Pathology

by

Raja Muhammad Saad Bashir

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy in Computer Science

Department of Computer Science

June 2023

Contents

List of Tables

List of Figures

Acknowledgments

In the name of Allah, the Most Gracious and the Most Merciful. All gratitude to the Almighty, who gave me strength and patience to conclude this work.

I express my deep gratitude to Prof Nasir Rajpoot, my supervisor, for his guidance and support in shaping me into a proficient researcher in the field of computer science and computational pathology. His mentorship and the opportunities he provided during my PhD journey are invaluable. I am grateful to Dr Shan E Ahmed Raza, my co-supervisor, for his valuable feedback on my academic writing and helpful suggestions on my projects.

I want to express my sincere gratitude to all the current and previous Tissue Image Analytics (TIA) Centre members at the University of Warwick. Your contributions have been invaluable to my academic journey, and I am deeply thankful for your support and collaboration. Special thanks to Muhammad Shaban for his help in the early days of my PhD, Muhammad Shaban, Navid Alemi Koohbanani and Hammam Alghamdi for all the dinners and lunches we had. Muhammad Shaban, Ruqayya Awan, Manahil Raza, Dang Vu and Muhammad Dawood for their moral support. My external collaborators have played an integral part in completing my thesis. I thank my clinical collaborators who have helped develop my understanding of histopathology. Specifically, I thank Prof Ali Khurram, Dr Hanya Mahmood and Dr Ayesha Azam.

I would also like to thank Dr Talha Qaiser and Dr Mohsin Bilal. They have been incredibly helpful and available mentors, providing invaluable insights and criticisms to improve my work. Their genuine care for my well-being, professionally and personally, has been truly motivating, and I am deeply grateful for their mentorship.

I want to extend my heartfelt thanks to Ahmer Mumtaz, who has been a

guiding figure and treated me like family with his emotional and social support. His caring gestures, such as cooking meals for me and sharing his life lessons, have been invaluable. I am deeply grateful for his presence in my life and the unwavering support he has provided, akin to that of an elder brother.

I would like to express my heartfelt thanks to Gandharv Bali, Manahil Raza, Ruoyu Wang, and Kavana Kela for being true friends during the challenging times, especially throughout the COVID-19 period. Your friendship has been a source of joy and support, and I am grateful for all the fun moments and activities we had, including birthdays, trips, and occasional gatherings.

Finally, I would like to express my deepest gratitude and appreciation to my family for their unconditional and never ending moral support during my entire educational pursuit. I am grateful beyond words for my father and mother's constant prayers and belief in my abilities, my brother's/sister's unconditional love, support and confidence in me, and my grandparents' special prayers and boundless love. I would also like to extend a special thanks to my uncle Anjum and cousin Kashif, who have been a constant support during my PhD.

I dedicate this thesis to my grandfather, Bashir, who passed away in 2014.

Sponsorship and Grants

I would like to acknowledge the University of Warwick for giving me the opportunity and providing me with financial support in the form of the Chancellor's International Scholarship.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. I declare that, except where acknowledged, the material presented in this thesis is my own work, and has not been previously submitted for obtaining an academic degree.

Raja Muhammad Saad Bashir
June 2023

Publications

First-Authored Publications

Journal Articles

- **Bashir, R.M.S**., Qaiser, T., Raza, S.E.A. and Rajpoot, N.M., 2023. Consistency Regularisation in Varying Contexts and Feature Perturbations for Semi-Supervised Semantic Segmentation of Histology Images. arXiv preprint arXiv:2301.13141. [Accepted in MedIA]

- **Bashir, R. M. S.**, Shephard, A. J., Mahmood, H., Azarmehr, N., Raza, S. E. A., Khurram, S. A., and Rajpoot, N. M. (2023). A digital score of peri-epithelial lymphocytic activity predicts malignant transformation in oral epithelial dysplasia. The Journal of Pathology.

Conference and Workshops Papers

- **Bashir, R. M. S.**, Shaban, M., Raza, S. E. A., Khurram, S. A., Rajpoot, N. M. (2022, July). A Novel Framework for Coarse-Grained Semantic Segmentation of Whole-Slide Images. In Medical Image Understanding and Analysis: 26th Annual Conference, MIUA 2022, Cambridge, UK, July 27–29, 2022, Proceedings (pp. 425-439).

- **Bashir, R. M. S.**, Qaiser, T., Raza, S. E. A., Rajpoot, N. M. (2020, October). HydraMix-Net: A deep multi-task semi-supervised learning approach for cell detection and classification. In Interpretable and Annotation-Efficient Learning for Medical Image Computing: Third International Workshop, iMIMIC 2020, Second International Workshop, MIL3ID 2020, and 5th International Workshop, LABELS 2020, Held

in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 3 (pp. 164-171).

- **Bashir, R. M. S.**, Mahmood, H., Shaban, M., Raza, S. E. A., Fraz, M. M., Khurram, S. A., Rajpoot, N. M. (2020, March). Automated grade classification of oral epithelial dysplasia using morphometric analysis of histology images. In Medical Imaging 2020, Houston, Texas, United States: Digital Pathology (pp. 245-250).

Co-Authored Publications

Journal Articles

- Raza, M., Awan, R., **Bashir, R.M.S.**, Qaiser, T. and Rajpoot, N.M., 2023. Mimicking a Pathologist: Dual Attention Model for Scoring of Gigapixel Histology Images. arXiv preprint arXiv:2302.09682.

- Jiao, Y., van der Laak, J., Albarqouni, S., Li, Z., Tan, T., Bhalerao, A., **Bashir, R.M.S**., Ma, J., Sun, J., Pocock, J., Pluim, J.P. and Koohbanani, N.A., 2023. LYSTO: The Lymphocyte Assessment Hackathon and Benchmark Dataset. arXiv preprint arXiv:2301.06304.

- Pocock, J., Graham, S., Vu, Q.D., Jahanifar, M., Deshpande, S., Hadjigeorghiou, G., Shephard, A., **Bashir, R.M.S.**, Bilal, M., Lu, W. and Epstein, D., 2022. TIAToolbox as an end-to-end library for advanced tissue image analytics. Communications medicine, 2(1), p.120.

- Da, Q., Huang, X., Li, Z., Zuo, Y., Zhang, C., Liu, J., Chen, W., Li, J., Xu, D., Hu, Z., **Bashir, R.M.S.** and Yi, H., 2022. DigestPath: A benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system. Medical Image Analysis, 80, p.102485.

Conference and Workshops Papers

- Shephard, A. J., Graham, S., **Bashir, R. M. S.**, Jahanifar, M., Mahmood, H., Khurram, A., Rajpoot, N. M. (2021, October). Simultaneous nuclear instance and layer segmentation in oral epithelial dysplasia. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual (pp. 552-561).

- Jahanifar, M., Shepard, A., Zamanitajeddin, N., **Bashir, R. M. S.**, Bilal, M., Khurram, S. A., Minhas, F., Rajpoot, N. (2021). Stain-robust mitotic figure detection for the mitosis domain generalization challenge. In Biomedical Image Registration, Domain Generalisation and Out-of-Distribution Analysis: MICCAI 2021 Challenges: MIDOG 2021, MOOD 2021, and Learn2Reg 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27– October 1, 2021, Proceedings (pp. 48-52).

- Shephard, A., Azarmehr, N., **Bashir, R. M. S.**, Raza, S. E. A., Mahmood, H., Khurram, S. A., Rajpoot, N. (2022, July). A Fully Automated Multi-Scale Pipeline for Oral Epithelial Dysplasia Grading and Outcome Prediction. In Medical Imaging with Deep Learning, Zurich.

- Shephard, A. J., Graham, S., **Bashir, R. M. S.**, Jahanifar, M., Mahmood, H., Khurram, A., Rajpoot, N. M. (2021, October). Simultaneous nuclear instance and layer segmentation in oral epithelial dysplasia. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual (pp. 552-561).

Abstract

Deep learning has pushed the boundaries of Computational Pathology (CPath) models for the diagnosis and prognosis of cancer. Many methods have been proposed that are fast, reliable and reproducible, but the performance largely depends on large scale labelled data. In most cases, a large amount of data remains unlabelled and needs to be used. Therefore, this thesis focuses on developing semi-supervised and weakly-supervised approaches for automated analysis of whole slide images (WSIs) leveraging unlabelled data.

To this effect, I present a semi-supervised method for simultaneously classifying and detecting tumour cells in Diffuse Large B-Cell Lymphoma (DLBCL). I first label the unlabelled data using pseudo labels and then train the framework using MixUp augmentation, which enhances the generalisation capability of the network. Next, I segment nuclei and tissue regions in WSIs using semi-supervised and self-supervised learning. Limited labelled data challenges the model's robustness due to limited exposure and learning experience. Therefore, I propose a consistency regularisation and cross-consistency training based semi-supervised learning framework. In addition, I also incorporate entropy minimisation to improve the confidence of pseudo labels predicted during training.

Finally, I use multiple instance learning (MIL) frameworks for the diagnosis (i.e., grading) and prognosis (i.e., malignant transformation) of oral epithelial dysplasia (OED). I propose a novel digital biomarker, based on a count of peri-epithelium lymphocytes, and demonstrate its association with poor progression-free survival (PFS) in OED. Then, I propose a method based on graph neural networks (GNN) in a larger cohort. Initially, I perform coarse segmentation to delineate the epithelium into sub-layers and then train GNN models with ranking loss. The findings reveal that nuclei from the epithelium and basal layers are significant diagnostic digital biomarkers for grading. In contrast, nuclei from the basal layer and peri-epithelium tissue area are found to be significant for OED malignant transformation.

Abbreviations

AI : Artificial Intelligence

ANN : Artificial Neural Network

BCSS : Breast Cancer Semantic Segmentation

CE : Cross-Entropy

CNN : Convolutional Neural Network

CPath : Computational Pathology

CPH : Cox proportional hazard

CRCFP : Consistency Regularisation in varying Contexts and Feature Perturbations for semi-supervised semantic segmentation of Histology Images

DL : Deep Learning

DLBCL : Diffuse Large B-cell Lymphoma

EMA : Exponential Moving Average

FCN : Fully Convolutional Networks

FPR : False Positive Rate

GNN : Graph Neural Networks

H&E : Haematoxylin & Eosin

IHC : Immunohistochemistry

KM : Kaplan–Meier

mIoU : Mean Intersection Over Union

MIL : Multiple Instance Learning

MLP : Multi-Layer Perceptron

ML : Machine Learning

MLL : Multi-Label Learning

MTL : Multi-Task Learning

MoNuSeg : Multi-organ Nucleus Segmentation Challenge

MSE : Mean Square Error

OED : Oral Epithelial Dysplasia

OSCC : Oral Squamous Cell Carcinoma

OLS : Ordinary Least Square

OPMDs : Oral Potentially Malignant Disorders

ORB : Oriented FAST and Rotated BRIEF

PFS : Progression-Free Survival

PELs : Peri-epithelial Lymphocytes

ROI : Regions of Interest

SD : Standard Deviation **SCE** : Symmetric Cross-Entropy

SGD : Stochastic Gradient Descent

SOTA : State-of-The-Art

SSL : Semi-Supervised Learning

SIFT : Scale-Invariant Feature Transform

SURF: Speeded-Up Robust Features

TILs : Tumour Infiltrating Lymphocytes

TME : Tissue Microenvironment

TPR : True Positive Rate

UMAP : Uniform Manifold Approximation and Projection

WHO : World Health Organization

WSI: Whole-Slide Image

Chapter 1

Introduction

1.1 Cancer

A cell is a fundamental building block of the human body. Cells divide through a process called mitosis, which allows for growth and repair in the body. Cancer is a state where cells of an organ exhibit abnormal and uncontrolled behaviour/reproduction in such an invasive pattern that destroys surrounding healthy tissues and eventually the organ itself [3]. In 2020, 18.1 million cancer cases were recorded worldwide, with 10 million cancer related deaths as reported by the International Agency for Research on Cancer (IARC) [4]. The most common types of cancer are breast and lung cancers worldwide, with 12.5% and 12.2% of the total number of cancers diagnosed respectively. In the case of cancer, due to multiple environmental factors and a multitude of genetic mutations in these cells, this division becomes uncontrolled, resulting in abnormal/uncontrolled behaviour that can transform into a tumour. Changes in a tissue cell might not be due to cancer, but it can develop into cancer if untreated while passing through the hyperplasia and dysplasia stages before becoming cancer.

Tumours are broadly categorised into malignant (invasive) and benign (non-invasive) sub-types. A tumour that can invade nearby tissue and spread from the primary site is malignant. The benign tumour shows abnormal growth but is harmless and cannot invade or spread around [5]. Cancer can spread from the primary site (origin) to neighbouring organs or tissues via the bloodstream or the lymphatic system. This process is known as metastasising, where the extent of metastases is termed as the cancer stage. Malignant tumours can be cured to improve long-term and disease-free survival if diagnosed and treated at the right time [6]. Cancer grading, provides the extent of damage caused by cancer in the appearance and behaviours of the cells.

1.1.1 Cancer Types

Cancer types can be categorised using the primary location of the tumour and the histological type. The primary location is the site of tumour origin (i.e., first appearance), e.g., breast, lung, colon and lung. Cancers are usually named after the organs, and hundreds of different cancer types exist based on only histological type [5]. Cancer can be broadly categorised into carcinoma, sarcoma, leukaemia, lymphoma, and myeloma. Tumours originating from the epithelium (i.e., from skin or tissue lining) are known as carcinomas and account for 80-90% of cancers. Sarcomas are tumours arising from connective and supportive tissues (i.e., bone, muscles and fat) and are the second biggest after carcinomas. Leukaemia, lymphoma, and myeloma based tumours are not solid and arise from malignant growth in blood, bone marrow and lymph nodes.

1.1.2 Oral Dysplasia

The term dysplasia consists of two Greek letters *dys* meaning bad and *plasia* meaning growth. Oral Dysplasia is defined as a potentially precancerous lesion of stratified squamous epithelium diagnosed histologically based on cellular atypia and architectural abnormalities [7]. Dysplasia is a series of subtle changes in the oral cavity that are theoretically reversible if diagnosed and treated at early stages with the right treatment. Not all changes are reversible, reversible changes are characterised by accelerated cell divisions, which can lead to the death of cells undergoing neoplastic transformations. These dysplastic lesions in the oral cavity have more importance as they are much more likely to transform into oral squamous cell carcinoma (OSCC) than the non-dysplastic lesions [8] as seen in Figure 1.1. Accumulation of genetic and epigenetic alterations have been seen during the development of malignant lesions in oral mucosa in a series of clinical and histological experiments for changes depicting dysplasia [9]. Genetic alteration involves changing the DNA sequences, while epigenetic alterations do not involve changes in the DNA sequences themselves, but rather in their interactions with other genes.

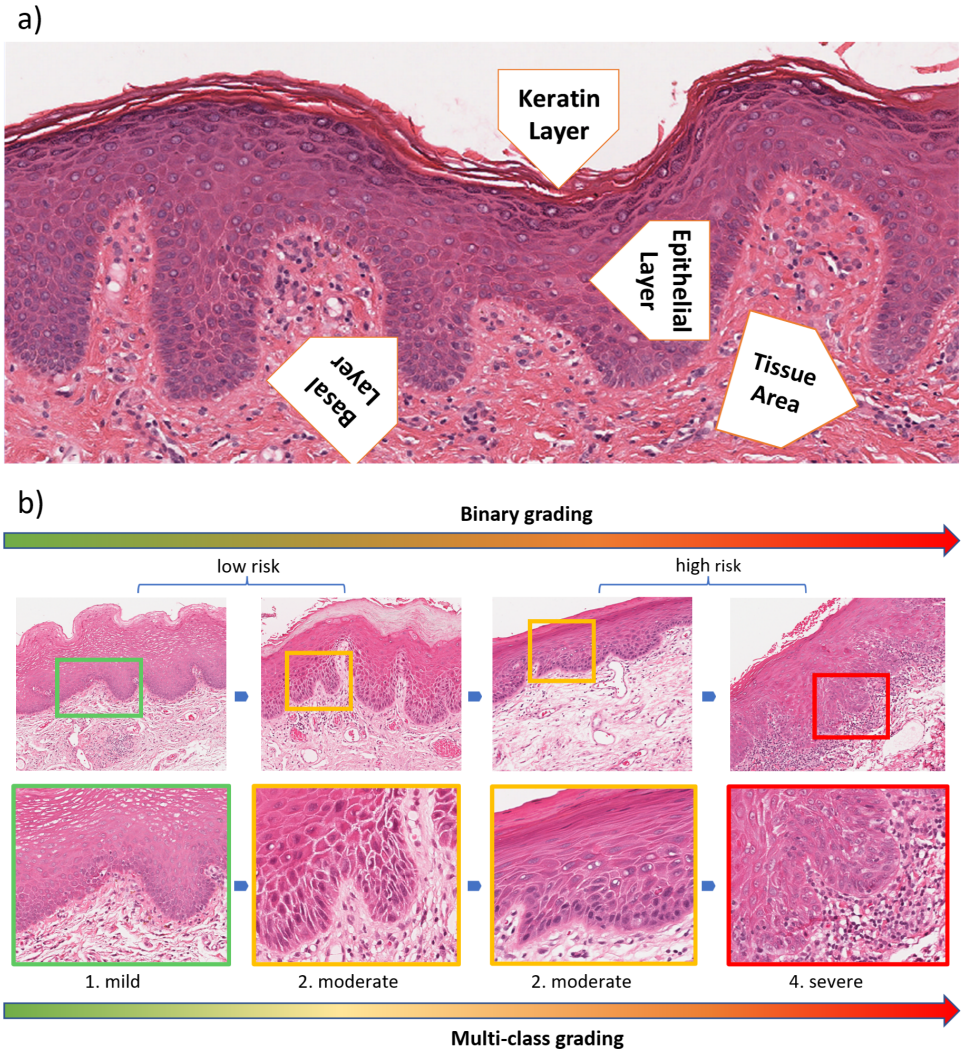


Figure 1.1: a) WSI showing layers of the oral epithelium; top/most superficial keratin layer, middle epithelial layer and bottom basal layer with underlying connective tissue. b) Different grades of oral epithelial dysplasia where first row depicts regions of interest (ROI), and the second row presents a zoomed-in version of the highlighted patches. 1) Mild dysplasia - some cytological and morphological changes restricted to the lower third of the epithelium. 2) Moderate dysplasia - significant cytological and morphological changes extending into the middle third of the epithelium 3) Severe dysplasia - significant cytological and architectural changes extending beyond the middle third and into the superficial epithelium. (Images from own work)

OED grading relies on the extent of involvement of the intra-epithelial layers (i.e., keratin, epithelium and basal layer), which are currently difficult to objectively delineate as seen in Figure 1.1 a). The diagnosis of OED involves a tissue biopsy and histological assessment using light microscopy, and the current gold standard grading system for OED is subjective and relies on evaluating at least 15 different cytological and architectural features. OED

changes usually start in the basal layer of the epithelium and progress upwards through the epithelial layers, with increasing severity. The cytological changes include nuclear and cellular pleomorphism, anisonucleosis/anisocytosis, hyperchromasia, atypical mitotic figures, and increased cellularity. Architectural changes include irregular epithelial stratification, loss of basal cell polarity, drop-shaped rete pegs, increased mitotic figures, and loss of epithelial cohesion [10]. The WHO 2005 system [11] grades cases as hyperplasia, mild, moderate, severe and carcinoma in situ. However, the presence of multiple categories was reported to add ambiguity to the treatment given to cases, and there has been an emphasis on the need for two-tier classification systems to improve reproducibility and clinical adoption. Kujan et al. [12] introduced a binary grading system, categorising cases as either low or high risk depending on the number of architectural and cytological features seen. In 2017, the WHO also released the new three-tier grading system instead, grading cases as mild, moderate or severe [13]. Despite the proprosed grading systems, the OED grading suffers from significant inter- and intra-observer variation [14] due to its subjective nature and interpretation can be hugely dependent upon the observer's experience and training. However, even these newer systems pose problems, with difficulty in the treatment given to moderate cases. Furthermore, the classification is complicated because some mild cases may progress to malignancy while some severe cases may not. Limited guidance or tools are currently available, which is critical for correct grading and aiding treatment decisions. This highlights the need for novel and objective approaches that can provide prognostic abilities along with the objective diagnosis as OED grading is vital to inform hospitals for patient management.

1.1.3 Diffuse Large B-cell Lymphoma

Malignancies derived from white blood cells (i.e., lymphocytes) are known as lymphomas and can be categorised into B-cell and T-cell lymphoma with respect to their origin in cells. Further, these B-cell lymphomas can be categorised into high- and low-grade lymphoma. Diffuse large B-cell lymphoma (DLBCL) is a high grade aggressive and fast-growing malignancy that affects the growth of anti-bodies known as B-type lymphocytes. It is more prevalent in western countries [15] affecting people with median age of 70 at diagnosis time [16]. The most common test for DLBCL is to remove part or all of the enlarged lymph node which is then checked by the hematopathologist under a microscope. Visual examination is assessed on the basis of grading system proposed by Ann Arbor [1] where stages I and II are considered as low or early stage DLBCL and stages III and IV are considered as high or advanced stage as shown in Table 1.1. Early stages can be treated using simple chemotherapy while for advanced

stages chemotherapy is combined with a drug named Rituximab. Although, advancements in treatment have improved the overall survival [17] of DLBCL patients with modern chemotherapy and Rituximab. However, approximately 40% of patients doesn't show the lasting response to therapy and inevitably die with DLBCL [18].

Table 1.1: DLBCL classification system proposed by Ann Arbor [1]

Stage	Staging Description
I	A single organ or site contains tumour (around a single lymph node)
II	At least two organs or sites in lymphatic regions of same side of diaphragm contain tumour
III	Lymphatic regions (including organs and lymph nodes) contain tumour on both side of the diaphragm
IV	Diffuse or disseminated association of one of multiple extra-lymphatic organs (like liver, lung nodules, bone marrow)

1.2 Cancer Diagnosis

Pathology is the gold standard for investigating the cancer, its cause and its effects using biopsies/resections. Cancer diagnosis usually involves physical examination, various laboratory tests, and samples of body tissues (i.e., biopsies and resections). The diagnostic process starts with extracting tissue samples from different organs through biopsies/resections carefully examined by surgeons and histopathologists, as detailed in the next section. Pathologists look for different patterns and anomalies at the cellular and architectural levels under the microscope, which might explain the underlying disease [19]. As a part of diagnosis, pathologists also determine the cancer grade/aggressiveness, which is then used in treatment options and predictive analysis along with other factors, e.g., site of origin, type and grade. Accurate diagnosis is critical to patient management, as a patient may be under- or over- treated based on the diagnosis. Treatment decisions include removing the cancerous region or using drug therapy and radiotherapy to reduce the chances of cancer relapses and help improve the patient's quality of life.

1.2.1 Slide Preparation

The journey of a tissue specimen from a biopsy/resection to analysis involves various steps, namely fixation, embedding, sectioning, and staining, explained below and shown in Figure 1.2.

Fixation: As a result of removal from an organ, cells start to die in the tissue, known as autolysis, and the original structure of the tissue is lost. To prevent

autolysis and preserve the original structures, the tissue specimen is first fixed using a fixative solution, such as formalin.

Embedding: To further preserve the tissue specimen and prepare it for the next procedures, it is hardened using an embedding medium, e.g., paraffin wax. Before embedding, the tissue specimen is dehydrated, where water is replaced with an organic solvent, such as ethanol, which is miscible with the embedding medium. This dehydration not only prepares the tissue for embedding but also facilitates the staining process as most dyes do not penetrate cells effectively if they are not properly dehydrated.

Sectioning: To prepare glass slides from a fixed, embedded tissue specimen, sectioning is performed using microtomy where slices of thickness $3 - 5\mu m$ are prepared and transferred to glass slides. It is an important step that ensures a proper tissue specimen analysis later on.

Staining: Finally, before putting slides in microscopes for analysis, these are stained with different dyes, which increases their contrast because most cells are virtually transparent, and it is very difficult to differentiate between them. The most commonly used stains for diagnosis is hematoxylin (blue) and eosin (red), abbreviated as H&E. However, there are many others, e.g. Giemsa stain, Masson's trichrome, Periodic acid Schiff (pas) and Congo red, etc. Another staining protocol known as Immunohistochemistry (IHC) employs specialised antibodies to detect over-expressed proteins, e.g., estrogen (ER), progesterone (PR), and human epidermal growth factor receptor (EGFR) etc.

Figure 1.2: The journey of a tissue slide from resection to becoming a glass slide before a histopathologist analyse it under the microscope for malignancy. 1) The tissue resection/biopsy is carried out of an organ and is fixed with formalin. 2) Tissue is divided into small sections for analysis, and 3) these small sections are then pre-processed, which involves eliminating water etc. 4) These are then embedded in paraffin wax to make tissue blocks which are used for 5) slicing the whole tissue blocks into multiple tissue slides for staining. 6) Different staining dyes are used to increase the contrast of the tissue cells and other areas. 7) To safeguard the tissue when viewed under a microscope, a slim layer of plastic or glass is placed over the stained portion on the slide. (Images from own work.)

1.2.2 Challenges in Routine Examination

Examining H&E stained histology slides for grading tissues is a meticulous and potentially time-consuming task for pathologists. They must carefully examine each case to ensure an accurate diagnosis and prognosis. It becomes especially challenging in the case of biopsy screening, where thousands of cases must be diagnosed in multiple hospitals each year. This challenge is exacerbated by staff shortages in most histopathology departments worldwide [20]. In addition, due to inherent inter- and intra- observer variability, different pathologists may give different diagnoses, leading to significant variation in diagnosis [14, 21]. This is because certain cancer grading guidelines, such as oral epithelial dysplasia (OED) grading, involve various nuclear architectural and cytological patterns and rely on the pathologist's subjective interpretation of the nuclei's appearance. Trainee pathologists tend to exhibit more variability in diagnosis compared to the experienced ones [22]. Additionally, there is often a low agreement between pathologists when presented with a rare cancer type [23]. Given these challenges, there is a need for a more objective measure of histopathology slides that can help to reduce the pathologist's workload.

1.3 Digital and Computational Pathology

Digital Pathology enables routine pathological practices to shift from a manual to a digital environment with the involvement of acquisition, management, and sharing of information digitally. With the advent of high-quality digital scanners, glass slides can now be scanned at very high resolution to capture an entire glass slide with remarkable precision. These slides can be stored, shared over the network, and can be used to apply image analytical tools to finding new digital biomarkers within the tissue section [24]. It has also enabled pathologists around the globe to share their work to engage and collaborate remotely to improve patient care with better diagnosis and prognosis of diseases [25]. It requires one additional step of scanning the slide using a digital scanner from the previous pipeline.

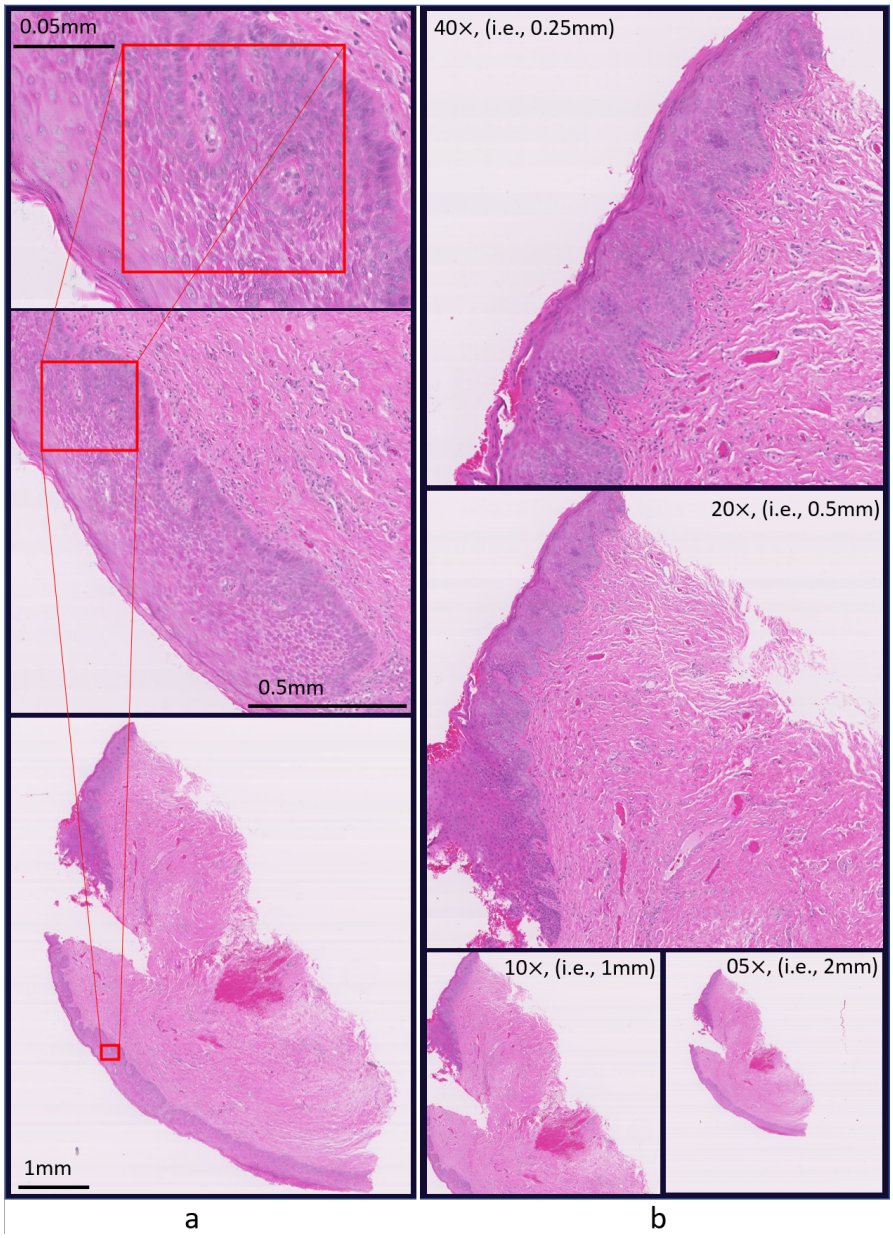


Figure 1.3: Illustration of WSI stored in pyramid structure where a) shows the sample ROIs of WSI at different micro per pixels (mpp) and (b) shows the the sample ROIs from different levels of the WSI with highest magnification 40×. (Images from own work.)

1.3.1 Whole Slide Image

Whole Slide Images (WSIs), the output of a digital scanner that converts a glass slide into a digital slide, are large images consisting of billions of pixels (e.g., 100,000 x 100,000). These images are stored digitally in compressed multi-resolution pyramid structures, which are down-sampled versions of the original image, as shown in Figure 1.3. In these pyramids, the images are

stored in a tiled fashion to improve reading capabilities by fast retrieval from sub-regions of an image. Because these images are compressed, opening them in standard image viewers is quite challenging as these often cannot fit into memory. Different compression and storing protocols are used for WSI formats, e.g., svs (Aperio Technologies, USA), ndpi (Hamamatsu Photonics, Japan), mexs (3DHISTECH, Hungary), isyntax (Philips, Netherlands), etc., as there is no standard file format globally accepted for WSIs. Openslide [26] is a C based library developed to provide an interface for reading WSI content from various formats, which paved the way for the development of algorithms for typical tasks. Similarly, KaKadu (C++) [27] provided the same interface for loading jp2 images. However, apart from programming interfaces, many WSI viewers are equipped with capabilities of performing different tasks from basic annotation to nuclei segmentation, e.g., QuPath [28], ASAP [29], HistomicTK [30] etc.

1.3.2 Computational Pathology

Computational Pathology (CPath) involves the analysis of raw WSI pixels with associated metadata for diagnosis using different mathematical and statistical models, where these models extract biological and clinically significant information for inference and prediction [31]. It helps in better patient management and precision medicine by assisting pathologists in their critical decision-making by reducing inter- and intra- observer variability using quantitative measures. The revolutionary digital developments in pathology have paved the way towards the possibility of computer-assisted diagnosis, where computers can assist a pathologist in their routine workflow by interpreting underlying hidden patterns from WSIs.

Pre-Processing: The appearance of tissue slides in WSIs is influenced by numerous factors, including the type of organ, staining conditions, optics, and image acquisition devices [32, 33]. These variations in appearance can pose challenges for computational algorithms that are used to analyse and diagnose the tissue. While pathologists can still diagnose slides with varying stains, the performance of CPath algorithms can be negatively affected by these variations. To address this issue, researchers are developing algorithms to standardise stain appearance across multiple images before analysis [34–37]. By removing inconsistencies in staining, these algorithms can ensure that the same tissue type and stain intensity are accurately represented across all images. In addition, to stain variation, WSIs may contain various artefacts such as tissue folds, ink markings, and out-of-focus regions [38]. These artefacts can complicate the tissue analysis, potentially leading to inaccurate diagnoses.

To address this issue, pre-processing algorithms can detect and account for artefacts, enabling pathologists to focus their analysis on artefact-free regions of the slide. Sometimes, these algorithms may even determine if a glass slide needs to be re-scanned [39, 40]. Standardising the appearance of WSIs and accounting for artefacts through pre-processing algorithms can significantly improve the accuracy and reliability of computational algorithms, thus enabling pathologists to make more informed and precise diagnoses.

Detection and Segmentation: WSIs are complex digital images that contain a vast amount of information that needs to be analysed by pathologists to reach a diagnosis. However, manual analysis of WSIs is time-consuming and may not always provide consistent and accurate results due to the complexity and variability of tissue structures and the staining process. Computational algorithms have emerged as a promising solution for analysing WSIs efficiently and accurately. These algorithms can assist pathologists in detecting, quantifying, and localising tissue components, improving diagnostic accuracy and reducing the time required for diagnosis. For example, CPath algorithms can identify nuclei in WSIs, which can be challenging to achieve through visual examination due to the large number of cells present on each slide [41–43]. Automated detection can also help identify difficult-to-spot objects such as mitotic cells [44], which can aid in diagnosing cancer and other diseases. In addition to identifying specific tissue components, segmentation algorithms can separate different tissue structures within the WSI. This segmentation can be particularly useful in examining morphological features associated with cancer grade and patient prognosis, such as breast tissue or oral cancer [45, 46]. By examining the structures and relationships between different components of the tissue, computational algorithms can help pathologists gain a more comprehensive understanding of the sample and it's potential disease state [47].

Cancer Type and Grade Prediction: In routine practice, the pathologists diagnose the type and grade of cancer using the glass slide or WSI, as the grade and type of cancer significantly impacts patient treatment and management [48, 49]. However, this task is subject to variability in diagnosis, which can lead to differences in treatment recommendations [23]. To address this issue, CPath is a computational tool that provides objective and reproducible measures, thus reducing diagnostic variability in cancer grading. Specifically, CPath can automatically perform OED grading, a grading system for oral dysplasia, subject to significant variation in pathologist diagnosis [50, 51]. Moreover, CPath algorithms can diagnose the cancer type automatically based on a tissue sample extracted from a specific organ, which is crucial since

different types of cancer may require different treatment regimens. CPath can provide a more objective measure of histopathology slides, which is a crucial factor for accurate diagnosis and reducing the workload of pathologists. Ultimately, these benefits can lead to improved patient outcomes.

Prediction of prognosis: Diagnosing cancer involves a standardised set of guidelines to determine the type and grade of cancer [10, 52]. However, predicting factors such as disease specific/free/overall survival, recurrence, and precision medicine can be more complicated. In these cases, the latest CPath algorithms can automatically extract a representative set of features related to the task from the raw WSI input. These features can also be used to train new pathologists on the most relevant diagnostic factors or patterns for a specific task. CPath algorithms can also discover new digital biomarkers that may enable superior diagnostic performance compared to the already established ones [47, 53–55]. By analysing the vast amount of information in WSIs, these algorithms can identify features that are not visible to the human eye and may provide additional insights into the disease state.

The revolutionary digital developments in pathology have led towards the possibility of computer based diagnosis, where computers can intelligently assist a pathologist in their routine workflow by interpreting underlying hidden patterns from WSIs. The main force behind the success of CPath is the use of artificial intelligence (AI), especially machine learning and deep learning, which enables CPath algorithms to uncover the underlying patterns and hidden relations between different types of tasks.

1.4 Artificial Intelligence for CPath

1.4.1 Machine Learning

Machine Learning (ML) is a sub-domain of larger domain artificial intelligence which started to gain importance in the 1990s with the manifesto of making computers learn (i.e., from Experience **E**) without being explicitly programmed for a task (i.e.,**T**) with some performance metrics (i.e., **P**). It is said to increase in **P** for **T** with an increase in **E** (Tom Mitchell, 1997). Earlier, there was not much data or computational power available for the machine to learn. For several years it has been based on rules and heuristics being programmed into the logic of a system, e.g., checkers-playing program [56]. Recently, machine learning is becoming indispensable across all aspects of life, from a smartphone, autonomous vehicles, speech recognition, visual surveillance, healthcare, etc. Machine learning is further divided into sub-types depending on the types of learning involved: supervised, semi-supervised, weakly supervised,

unsupervised, and reinforcement learning.

- **Supervised Learning** involves training with labels given some data in the form of $X = \{x_1, x_2, \ldots, x_N\}$, and labels $Y = \{y_1, y_2, \ldots, y_N\}$, where x_i denotes the i^{th} data point, y_i its corresponding target label, and N the number of data points or the length of the dataset.

- **Semi-supervised Learning** involves training with partially labelled data: $X = \{x_1, x_2, \ldots, x_n\} \subset \Omega$ with labels $Y = \{y_1, y_2, \ldots, y_n\} \subset \Psi$ and unlabelled data $U = \{u_1, u_2, \ldots, u_m\} \subset \Omega$. Here $N = n + m$ is the total number of data points, while n and m represent the number of labelled and unlabelled data points. Ω represents the space of all possible input data points and Ψ represents the space of all possible labels. The goal is to learn from the labelled data (X, Y) and predict pseudo labels $Z = \{z_1, z_2, \ldots, z_m\}$ for the unlabelled data U. Finally, we want to learn a model $g : \Omega \to \Psi$, where $X' = X \cup U$ and $Y' = Y \cup Z$, X' represents the full set of data, and Y' represents the complete set of labels including both original and predicted ones.

- **Weakly Supervised Learning** involves training with lower quality labelled data, where the labels are not entirely informative or accurate and are noisy. Multiple Instance Learning (MIL) is a type of weakly supervised learning where $X = \{B_1, B_2, \ldots, B_n\} \subset \Omega$ consists of a "bag" $B_i = \{x_1^i, x_2^i, \ldots, x_{m_i}^i\}$ containing m_i instances for each bag B_i, where n is the total number of bags. Each bag B_i has the label $y_i \in \Psi$ where Ω represents the space of all possible input data points and Ψ represents the space of all possible labels. The objective of MIL is to learn a function $f : \Omega \to \Psi$ that can accurately predict the label y of a new bag B based on its instances.

- **Unsupervised Learning** involves training without labels or desired output. Given some data $X = \{x_1, x_2, \ldots, x_N\}$, where N is the number of data points or the length of X, it aims to discover hidden relationships within the set X and group similar data points into clusters $C = \{c_1, c_2, \ldots, c_M\}$, where M represents the number of desired clusters.

- **Reinforcement Learning** involves learning from the environment with interactions and their associated rewards/penalties without supervision. An agent, denoted as a, learns from its state s_t in an environment with the rewards r_t at each time step t, such that it maximises the cumulative reward for an objective.

Traditional machine learning involves domain-specific hand-crafted features for models to learn from, e.g., from images extracting features like edges, Scale-

Invariant Feature Transform (SIFT), Speeded-Up Robust Features (SURF), and Oriented FAST and Rotated BRIEF (ORB) features, eliminating the possibility of learning from the raw input. Deep Learning (DL), on the other hand, has emerged recently and taken all these tasks by storm in setting new benchmarks with its remarkable ability to learn from raw input. DL is a further sub-type of Machine Learning (ML) where it learns the hidden representation in data by directly mapping input to output.

1.4.2 Deep Neural Networks

A neural network or artificial neural network (ANN) consists of multiple layers, i.e. (input layer, hidden layer, and output layer) of highly interconnected neurons stacked together for a task, each activating incoming input in non-linear space. The input layer receives input of different types, e.g., features vectors, images and text. The next layer forwardes these input values through a weighted connection with neurons. The hidden layer activates incoming weighted input with non-linear activation functions, e.g., Sigmoid, Tanh, ReLU. The output layer outputs the input signal according to the learned distribution function into the required output, i.e., continuous or discrete value. This whole process of taking input, processing it, and predicting a value is known as Forward Pass, as shown in Equation 1.1 where $\hat{\mathbf{y}}$ is the predicted value, $\mathbf{W}$ is the learned weights matrix for all neurons, $\mathbf{X}$ is the input to the network and $\mathbf{b}$ is the bias.

$$\hat{\mathbf{y}} = \mathbf{W}\mathbf{X} + \mathbf{b} \tag{1.1}$$

Finally, the loss function calculates the error between the output prediction and the target value, which estimates the difference between the two distributions. An Optimiser, e.g., Adam, Stochastic Gradient Descent (SGD), AdaGrad etc., adjusts the network parameters according to the loss to improve the prediction accuracy. This whole process is known as Backward Pass, as shown in Equation 1.2 where l is the loss between actual and predicted value summed over all items m and is optimised on $\mathbf{W}$. Neural networks, in short, are fully interconnected neurons, i.e., all neurons are interconnected to all other neurons of their previous and subsequent layers.

$$\arg\min_{\mathbf{W}} \frac{1}{m} \sum_{i=1}^{m} l(\hat{y}_i - y_i) \tag{1.2}$$

Real world problems are different in nature, and they require different architectures. Relatively small problems require small or shallow networks, while complex problems require deeper and larger networks [57, 58]. These simple ANNs do not work for images because they do not consider the spatial

context, as they need flattened input for processing. On the other hand, convolutional neural networks (CNN) perform best when it comes to images due to their intrinsic property of taking the whole image as input.

1.4.3 Convolutional Neural Networks

Convolutional neural networks (CNNs) [59] use a special operator for processing inputs and multiplying neurons known as convolution, where convolutions are linear operations spatially. CNNs have three essential features because of convolution i) they have sparse interactions, unlike ANNs where every neuron is interconnected to every other neuron; ii) they share the same parameters across all the input iii) they are equivariant to translation [60, 61]. CNNs perform best where spatial context is vital in data or grid-like structure, e.g., images, sound and are not suitable for data where spatial structure does not matter, like a relational database. CNNs consist of three components: convolution operation, pooling operation and dense layers. CNNs initially perform convolution operations over the input, and several convolutional operators are stacked together to extract different features. Convoluted inputs are activated using some non-linear activation function, e.g., Sigmoid, Tanh and ReLU. Pooling operations reduce the input size and introduce tolerance to spatial translations in the model. There are different pooling operations, e.g., max pooling, min pooling, and average pooling. Lastly, there are dense fully connected layers for the final computation of the underlying tasks.

CNNs in CPath

Convolutional Neural Networks (CNNs) have revolutionised the field of computer vision and medical imaging, especially computational pathology, by enabling automated analysis of large-scale and complex WSIs with high accuracy and efficiency [62, 63]. The power of CNNs lies in their ability to extract and hierarchically learn image features, starting with simple low-level features such as edges and textures and gradually learning more complex and high-level features such as cell morphology, tissue structure, and other image-level characteristics [64]. This hierarchical learning approach allows CNN based models to effectively capture the underlying patterns and structures in histopathology images and use this information for tasks such as tissue segmentation, nuclei detection, cell classification, and cancer diagnosis.

Despite their remarkable success, developing effective CNN based models for computational pathology is still challenging. The main challenges are the large size of WSI and stain variability. Large size of WSI leads to higher memory demands and intensive computational requirements, which in result complicates the model training and deployment. Stain variability on the other hand can

hinder the generalisation power of a model which can lead to poor performance when deployed in real-time on external cohorts (i.e., data that is sourced from outside the primary study or experiment) [34, 36, 44]. Additionally, the limited availability of high-quality labelled data for training and validation can make it challenging to develop robust and accurate models [65]. Moreover, the high computational requirements of deep learning methods can make training and inference times prohibitively long [46]. To address the challenges, ongoing research efforts focus on developing more advanced CNN based models and addressing the limitations of current models. For example, some studies have proposed using multi-scale and multi-task learning techniques to improve model accuracy and efficiency [66, 67]. Others have proposed using weakly supervised and semi-supervised learning methods to reduce the reliance on high-quality labelled data [54, 68]. In addition, various studies have explored transfer learning, data augmentation, and other techniques to improve model performance and generalisation [69, 70]. Although CNN based models show promise in computational pathology, further research is required to improve their accuracy, robustness, efficiency, interpretability and generalisation for clinical use.

1.4.4 Graph Neural Networks

Graph Neural Networks (GNNs) [71] are specialised neural network architectures developed to handle graph-structured data, which primarily consists of nodes and edges. The fundamental spine of such architectures is the 'message passing framework', a mechanism that collects information from adjacent nodes and calculates a unique representation for each node. Messages typically contain node features or embeddings. This message passing enables GNNs to learn complex representations of graph data by embedding both individual node features and their relational context. In this framework, h_l^v denotes the state of a node v at the l^{th} layer, u represents a neighbour of v, and $\mathcal{N}(v)$ specifies the set of neighbours of node v. Mathematically, the framework first aggregates the states h_l^u of the neighbouring nodes $u \in \mathcal{N}(v)$ using the AGGREGATE function:

$$m_l^v = \text{AGGREGATE}^l(h_l^u : u \in \mathcal{N}(v)) \tag{1.3}$$

Next, it computes the state h_{l+1}^v of node v at the next layer $(l+1)$ by applying the UPDATE function on the node's current state h_l^v and the aggregated message m_l^v:

$$h_{l+1}^v = \text{UPDATE}^l(h_l^v, m_l^v) \tag{1.4}$$

Following these operations, a READOUT function orchestrates the final node representations into a graph-level output. The readout function combines the information from all node representations into a unified graph-level representation. Common choice is a simple operation like summation or averaging, but more complex operations can also be used.

This classic approach of GNNs has inspired several variations and extensions such as Graph Attention Networks (GATs) [72] and Convolutional Graph Networks (GCNs) [73]. Numerous applications have leveraged these models with significant impact, from social network analysis [74] and drug discovery [75] to recommendation systems [76]. Despite their successes, GNNs present ongoing research challenges related to the management of large graphs, the integration of domain knowledge, and the necessity for interpretability and fairness.

GNN in CPath

Graph Neural Networks (GNNs) have emerged as a promising tool for computational pathology analysis due to their ability to model complex relationships between features in histopathology images. GNNs can learn from the graph structure of the tissue samples, which contains information about the spatial arrangement and connectivity of the different components within the tissue microenvironment. By leveraging this information, GNNs can better capture the spatial dependencies between different image regions and produce more accurate segmentation and classification results. Recent studies have demonstrated the potential of GNNs in various computational pathology applications from segmenting tumour regions [77] to classification [78, 79], detection of nuclei [80] and survival analysis [81]. Despite their potential, GNNs are still relatively new in computational pathology, and there is much room for further exploration and optimisation. Continued research is needed to improve GNNs' robustness, scalability, and interpretability for clinical use.

1.4.5 Evaluation Metrics

In this thesis, we utilise a range of evaluation metrics to thoroughly assess the performance of our proposed models. These include F1-Score (also known as Dice Score), Precision, Recall, Accuracy, Area Under the Receiver Operating Characteristic Curve (AUROC), Mean Intersection over Union (mIoU), and Mean Absolute Error (MAE). Each of these metrics provides a unique perspective on the effectiveness of our models in different aspects of the classification or segmentation task. The evaluation metrics used in this thesis for both classification and segmentation tasks are defined as follows:

- **True Positives (TP):** These are the correctly predicted positive values, meaning the model correctly predicted the positive class.
- **True Negatives (TN):** These are the correctly predicted negative values, meaning the model correctly predicted the negative class.
- **False Positives (FP):** These occur when the negative class is incorrectly predicted as positive.
- **False Negatives (FN):** These occur when the positive class is incorrectly predicted as negative.
- **F1-Score (Dice Score):** This is the harmonic mean of precision and recall. It is calculated as:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{1.5}$$

- **Precision:** This is the ratio of correctly predicted positive observations to the total predicted positive observations. It is calculated as:

$$Precision = \frac{TP}{TP + FP} \tag{1.6}$$

- **Recall (Sensitivity):** This is the ratio of correctly predicted positive observations to all observations in the actual class. It is calculated as:

$$Recall = \frac{TP}{TP + FN} \tag{1.7}$$

- **Accuracy:** This is the ratio of correctly predicted observations to the total observations. It is calculated as:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{1.8}$$

- **Area Under Receiver Operating Characteristic Curve (AUROC):** This is a performance measurement for classification problem at various thresholds settings. It tells how much the model is capable of distinguishing between classes.
- **Mean Intersection over Union (mIoU):** This measures the overlap between the predicted segmentation and the ground truth. It is defined as:

$$IoU = \frac{TP}{TP + FP + FN} \tag{1.9}$$

The mIoU is then the average IoU across all classes.

- **Mean Squared Error (MSE):** This is a measure of the average of the squares of the errors or deviations. It is calculated as:

$$MSE = \text{Average}((\text{TrueValue} - \text{PredictedValue})^2) \tag{1.10}$$

In segmentation tasks, each pixel is treated as a separate prediction, hence the area of pixels is used in the above definitions.

1.4.6 Challenges in Computational Pathology

Despite advancements, the availability of labelled data remains a significant challenge for deep learning algorithms. Labelled data is essential for training machine learning models to accurately identify various pathologies, particularly for deep learning models, as they require large amounts of labelled data to achieve peak performance. Techniques like semi-supervised learning, weakly supervised learning, unsupervised learning, and reinforcement learning can perform better than fully supervised models in the absence of labelled data. However, even with these techniques, fully supervised with largely annotated data outperforms them, underscoring the need for more labelled data. Apart from the labelled data, the gigapixel nature of the pathology images is another critical challenge to training deep learning models. This can make it difficult to train machine learning models as it is impossible to fit these images into the memory with current hardware. One solution is the tessellation of WSI, i.e., dividing them into smaller patches that can be processed with batch processing. However, this approach risks losing important contextual information, which can be crucial for accurate diagnosis. Recent developments have focused on compressing the WSIs into latent space, which the deep learning models can further process without losing critical information. Computational costs for processing and storing WSIs and intermediate results are also a significant challenge in CPath, particularly for smaller healthcare facilities that may need more financial resources to invest in the necessary hardware and software. Interpretability is another major challenge in CPath, as machine learning models can often produce accurate results but can be challenging to understand. This is particularly important in the medical field, where experts need to be able to justify the underlying reasons for specific diagnoses.

1.5 Aims and Objectives

This thesis seeks to develop robust deep learning based methods to overcome the significant challenges associated with developing reliable AI models for the diagnosis and prognosis from WSIs with limited or noisy labelled data.

The proposed research aims to leverage limited labels to classify and segment different tissue regions in a complex tissue microenvironment (TME). The TME is a microscopic niche consisting of multiple cell types with unique biological roles and spatial relationships with respect to each other. Another major challenge addressed in the thesis is the prediction of malignant transformation in oral epithelial dysplasia (OED) using a top-down approach via weak labels associated with the patient outcomes.

1.5.1 Main Contributions

- I develop a semi-supervised learning based method called HydraMix-Net [82] for the simultaneous detection and classification of tumour cells in Diffuse Large B-Cell Lymphoma (DLBCL). The framework utilises the concepts of MixUp and symmetric cross-entropy (SCE) for improving performance in the absence of large annotated data.

- I develop a semi-supervised learning based semantic segmentation framework for tissue regions and nuclei [70]. The framework uses contrastive learning with cross-consistency training to make the model robust against varying contexts and perturbation. I also incorporate entropy minimisation further to improve the accuracy and reliability of the segmentation results.

- I investigate the top-down approach for the prognosis of oral epithelial dysplasia using weak labels [83]. I train a MIL model for predicting malignant transformation. Moreover, I also delve into analysing hotspots highlighted by the model for significant prognostic digital biomarkers in epithelium and peri-epithelium. I identify significant digital biomarkers that are associated with malignant transformation, providing valuable insights for prognosis in oral epithelial dysplasia.

- I further enhance the top-down approach for the diagnosis and prognosis of oral epithelial dysplasia (OED) by incorporating the segmentation of epithelium [46] into sub-layers and training graph neural networks (GNNs) for analysis using sub-layer features. GNNs provide a powerful tool for modelling large whole slide images (WSIs) as interconnected graphs, offering explainability and interpretability at the node level. This led to statistically significant digital biomarkers strongly correlated with OED grading and prediction of malignant transformation. These findings contribute to the advancement of accurate and reliable diagnostic and prognostic approaches for OED.

1.6 Thesis Organisation

Chapter 2: Deep Multi-Task Semi-Supervised Learning Approach for Cell Detection and Classification. Semi-supervised techniques have removed the barriers of large scale labelled sets by exploiting unlabelled data to improve the performance of a model. In this chapter, I propose a semi-supervised deep multi-task classification and localisation approach HydraMix-Net in the field of computational pathology where labelling is time consuming and costly. Firstly, pseudo labels are generated by using the model's prediction on the augmented set of unlabelled images with averaging. The high entropy predictions are further sharpened to reduce the entropy and mixed with the labelled set for training. The model is trained in a multi-task learning manner with noise tolerant joint loss for classification and localisation, where it achieves better performance when given limited data compared to a convolutional neural network (CNN) trained in a supervised manner. On DLBCL data, it achieves 80% accuracy in contrast to the CNN, achieving 70% accuracy when given only 100 labelled examples.

Chapter 3: Semi-Supervised Learning for Segmenting Tissue Regions and Nuclei Histology Images. Semantic segmentation of various tissue and nuclei types in histology images is fundamental to many downstream tasks in the area of computational pathology (CPath). In recent years, Deep Learning (DL) methods have been shown to perform well on segmentation tasks, but DL methods generally require a large amount of pixel-wise annotated data. Pixel-wise annotation sometimes requires an expert's knowledge and time, which is laborious and costly to obtain. In this chapter, I present a consistency based semi-supervised learning (SSL) approach that can help mitigate this challenge by exploiting a large amount of unlabelled data for model training, thus alleviating the need for a large annotated dataset. However, SSL models might also be susceptible to changing context and feature perturbations exhibiting poor generalisation due to the limited training data. I propose an SSL method that learns robust features from both labelled and unlabelled images by enforcing consistency against varying contexts and feature perturbations. The proposed method incorporates context-aware consistency by contrasting pairs of overlapping images in a pixel-wise manner from changing contexts resulting in robust and context invariant features. I show that cross-consistency training makes the encoder features invariant to different perturbations and improves the prediction confidence. Finally, entropy minimisation is employed to further boost the confidence of the final prediction maps from unlabelled data. I conduct an extensive set of experiments on two publicly available large datasets (BCSS and MoNuSeg) and show superior performance compared to

the state-of-the-art methods.

Chapter 4: Weakly Supervised Learning for Predicting Malignancy in Oral Epithelial Dysplasia (OED). Oral squamous cell carcinoma (OSCC) is amongst the most common cancers worldwide, with more than 377,000 new cases worldwide each year. OSCC prognosis remains poor, related to cancer presentation at a late stage, indicating the need for early detection to improve patient prognosis. OSCC is often preceded by a premalignant state known as oral epithelial dysplasia (OED), which is diagnosed and graded using subjective histological criteria leading to variability and prognostic unreliability. In this work, I propose a deep learning approach for the development of prognostic models for malignant transformation and their association with clinical outcomes in histology whole slide images (WSIs) of OED tissue sections. I train a weakly supervised method on OED (n= 137) cases with transformation (n= 50) status and a mean malignant transformation time of 6.51 years ($\pm$5.35 SD). Performing stratified 5-fold cross-validation achieves an average AUROC of 0.78 for predicting malignant transformations in OED. Hotspot analysis reveals various features of nuclei in the epithelium and peri-epithelial tissue to be significant prognostic factors for malignant transformation, including the count of peri-epithelial lymphocytes (PELs) ($p < 0.05$), Epithelial layer nuclei count (NC) ($p < 0.05$) and Basal layer NC ($p < 0.05$). Progression free survival (PFS) using the Epithelial layer NC ($p < 0.05$, C-index = 0.73), Basal layer NC ($p < 0.05$, C-index = 0.70) and PEL count ($p < 0.05$, C-index = 0.73) showed association of these features with a high risk of malignant transformation in our univariate analysis. Our work shows the application of deep learning for prognostication and prediction of PFS of OED for the first time and has significant potential to aid patient management. Further evaluation and testing on multi-centric data is required for validation and translation to clinical practice.

Chapter 5: Graph Based Learning for Predicting Grade and Malignancy in Oral Epithelial Dysplasia (OED). In this chapter, I investigate the use of graph neural networks (GNNs) for diagnostic and prognostic purposes in OED (n=241) cases with transformation (n=50) status and mean transformation time of 6.51 years ($\pm$5.35 SD). The diagnostic task is predicting the OED binary grading of low-risk vs high-risk, while the prognostic task involves predicting the OED malignant transformation status. I employ a GNN with EdgeConvs and multi-layer perceptrons (MLP) in the final aggregation and show that it is able to predict OED grades with an AUROC of 0.81 and malignant transformation with an AUROC of 0.76, as determined by a stratified 5-fold cross-validation bootstrapped using three different random

seeds. Hotspot and cellular composition analysis within the epithelial layer and peri-epithelial tissue regions for both tasks reveal significant diagnostic and prognostic nuclear features, e.g., nuclei count, crowdedness, solidity etc. In a univariate analysis, higher proportions of basal and epithelial layer nuclei count showed a correlation with poor progression-free survival (PFS) with the significance of ($p < 0.05$, C-index = 0.81) and ($p < 0.05$, C-index = 0.70). Similarly, the higher proportions of nuclei count in the peri-epithelium showed the most correlation with poor PFS in both univariate and multivariate analysis with ($p < 0.05$, C-index = 0.83). Our study demonstrates the use of DGNN to predict OED grades and malignant transformation for PFS. Our work has significant potential towards clinical adoption to aid patient management and care. However, additional testing on data from multiple centres is required to validate and implement our findings in clinical practice.

Chapter 6: Conclusions and Future Directions. This chapter summarises the main contributions of this thesis and lists potential future research directions for extending this work.

Chapter 2

Deep Multi-Task Semi-Supervised Learning Approach for Cell Detection and Classification

2.1 Introduction

Detecting and classifying nuclei in histology images is crucial in various downstream analyses, including cell counting, cell segmentation, and studying inter-cellular connections. However, the task of detection and classification is difficult due to the intricate texture of histology images due to stain variability or type, the variability in the shape of nuclei, and the presence of touching cells. Numerous techniques have been proposed to address these challenges, with deep learning (DL) methods being the most effective in terms of performance. In recent years, deep learning has revolutionised computer vision and achieved state-of-the-art (SOTA) performance in various vision-related tasks. The inevitable fact is that most of the DL success is attributed to the availability of large scale datasets and compute power available these days. To achieve SOTA performance, it is essential to train models using single-task learning approaches on large-scale datasets with their associated labels. The costs associated with labelling the datasets are often very high, especially for medical imaging data involving expert knowledge to collect the ground truth. In contrast, semi-supervised learning (SSL) approaches [84] take advantage of the limited labelled data and leverage readily available unlabelled data to improve the model performance.

SSL techniques have been successfully applied in computer vision, especially in the medical domain with adaption from natural images, with popular

methods such as Mean Teacher [85] and Virtual Adversarial Training [86]. In recent years, they have also been used in computational pathology for tasks such as clustering, segmentation, and image retrieval. This also alleviates the need for the time-consuming and laborious task of manual annotations and assists in training more complex models for better performance. Generally, SSL techniques follow a two-step approach a) predict pseudo labels for unlabelled data from the model trained on limited labelled data and b) retrain the model on pseudo labels and limited labelled data to improve the performance. To improve the learning ability of SSL by introducing regularisation [87, 88], and entropy minimisation [89] to avoid high-density predictions and train models into an end-to-end manner. However, due to the unique challenges posed by the giga-pixel WSIs, multi-scale resolutions, contextual information, and stain invariability of patches extracted from WSIs, directly applying popular semi-supervised algorithms in pathology classification may not be straightforward.

The purpose of a semi-supervised task is to learn from unlabelled data during learning such that it improves the model's performance. To achieve this goal, these approaches take advantage of different techniques to mitigate the issues faced during learning, e.g., consistency regularisation, entropy minimisation and pseudo label noise reduction etc. Decision boundary passing through high-density regions can be minimised using entropy minimisation techniques like [89], which minimise entropy with the help of a loss function for the unlabelled data. Consistency regularisation can be achieved using standard augmentation such that the network knows if the input was being altered in some ways, e.g., rotation, etc. [87, 88]. Semi-supervised approaches also suffer from noisy pseudo labels, an issue where pseudo labels used in semi-supervised learning, can be inaccurate or false due to the false prediction of the trained model. Labels for unlabelled data are generated using a model trained on a small amount of labelled data. These generated labels, called pseudo-labels, are likely to contain errors, especially if the initial amount of labelled data is small or unrepresentative of the whole population. These wrong pseudo labels can introduce noise in the training batches, but this can be handled using noise reduction methods such as those proposed by [90]. Using these standard approaches, there have been semi-supervised methods for the classification of natural images, e.g., Berthelot et al., [91] used simple data augmentation and MixUP technique [88] for consistency regularisation and used sharping [92] for entropy minimisation for semi-supervised training. Tarvainen et al., [85] improved the temporal ensembling over labels to use the moving average of the weights of the student model in the teacher model after comparing the student's prediction with its teacher's prediction, which in turn improves learning of the teacher model.

Regarding cell classification and detection, Cirean et al., [87] proposed a

simple deep learning based classification model to differentiate between the mitotic and non-mitotic cells in the breast WSI's. Sirinukunwattana et al., [41] used the locality sensitive information to localise the cell nuclei while using the Neighbouring Ensemble Predictor (NEP) for classification purposes. Qaiser et al., [93] proposed the joint multi-task framework to explore the spatial arrangements of the tumour cells and their localisation with the collagen VI in DLBCL by proposing the novel digital proximity signature (DPS) marker in the tumour rich collagen regions. Inspired by all these methods and techniques, we present our novel deep multi-task join training framework for end-to-end classification and detection.

In this chapter, we present a multi-task SSL method to alleviate the need for the time-consuming and laborious task(s) of manual labelling for histology whole-slide images (WSI). In this regard, we opted to use diffuse large B-cell lymphoma (DLBCL) data because manual annotation of cell type and nuclei localisation is very hard due to the large number of cells present in WSIs. The cell detection may also help in understanding the spatial arrangement of malignant cells within the tumour micro-environment with collagen VI. DLBCL malignancy originates from B-cell lymphocytes, and it is the most common high-grade lymphoma among the western population with relatively poor disease prognosis among lymphomas [94, 95]. The use of modern chemotherapeutic treatments has improved the survival rates of patients with DLBCL [95]. However, despite these advancements, approximately 40% of patients do not respond well to treatment and eventually succumb to the disease. This variability in treatment response is partly due to the tumour heterogeneity [96]. Recently, significant progress has been made in understanding this diversity, with studies exploring the role of the tumour microenvironment (TME) in DLBCL [97].

We present a novel deep multi-task learning framework, HydraMix-Net, for simultaneous detection and classification of cells, enabling end-to-end learning in a semi-supervised manner. To the best of our knowledge, we are the first to enhance a semi-supervised approach by improving a single loss term with noisy pseudo labels. This advancement enables the joint training of multi-task problems, thereby improving performance. Our main contributions are a) a novel multi-task SSL framework (HydraMix-Net) for cell detection and classification and b) combating noisy pseudo labels using a symmetric cross-entropy loss function.

2.2 Materials and Methods

2.2.1 Data

A total of 32 WSIs were collected for this study and stained with immunohistochemistry and Hematoxylin counter-stain to simultaneously detect collagen VI and nuclear morphology. The cohort for the Diffuse Large B-cell Lymphoma (DLBCL) study consisted of twelve participants, ranging in age from 24 to 90 years. Of these, ten were female and two were male. An expert pathologist in the VSM tool annotated the ground truth for cell detection and classification for 10 cases, resulting in a total of 2617 annotated cells, of which 2039 were tumour cells, 462 lymphocytes, and 116 macrophages. Patches of size 41 $\times$ 41 pixels, with the cell centroid kept at centre, were extracted, yielding 12553 patches. Given the inherent class imbalance in patches, offline augmentations including flipping, rotation, and crop were applied to balance the dataset. This process resulted in 24000 patches, with each class being equally represented by 8000 patches. For training and testing purposes, 7 and 3 WSIs were employed, respectively. Based on a 70-30 split, this resulted in 18000 training patches and 6000 test patches with three classes tumours, lymphocytes and background.

2.2.2 Methods

The semi-supervised method HydraMix-Net is a novel approach that integrates a variety of multi-task and semi-supervised techniques to resolve different learning challenges. The model applies techniques such as consistency regularisation using standard augmentations and the MixUP method [88], entropy minimisation assisted by label sharpening [92], and addresses noisy pseudo labels through modified loss terms like symmetric cross entropy (SCE) loss [90]. The HydraMix-Net jointly optimises the combined loss function to classify and localise centroids for the cell patches. Our multi-task learning framework consists of a CNN backbone model with three heads responsible for the classification and regression (i.e., localisation of cell nuclei). The schematic diagram of the model can be seen in Figure 2.1. The following sections delineate the data augmentation, pseudo label generation, noise handling and training in the semi-supervised HydraMix-Net model. To facilitate the reader's comprehension, frequently used mathematical notations are listed and defined in Table 2.1

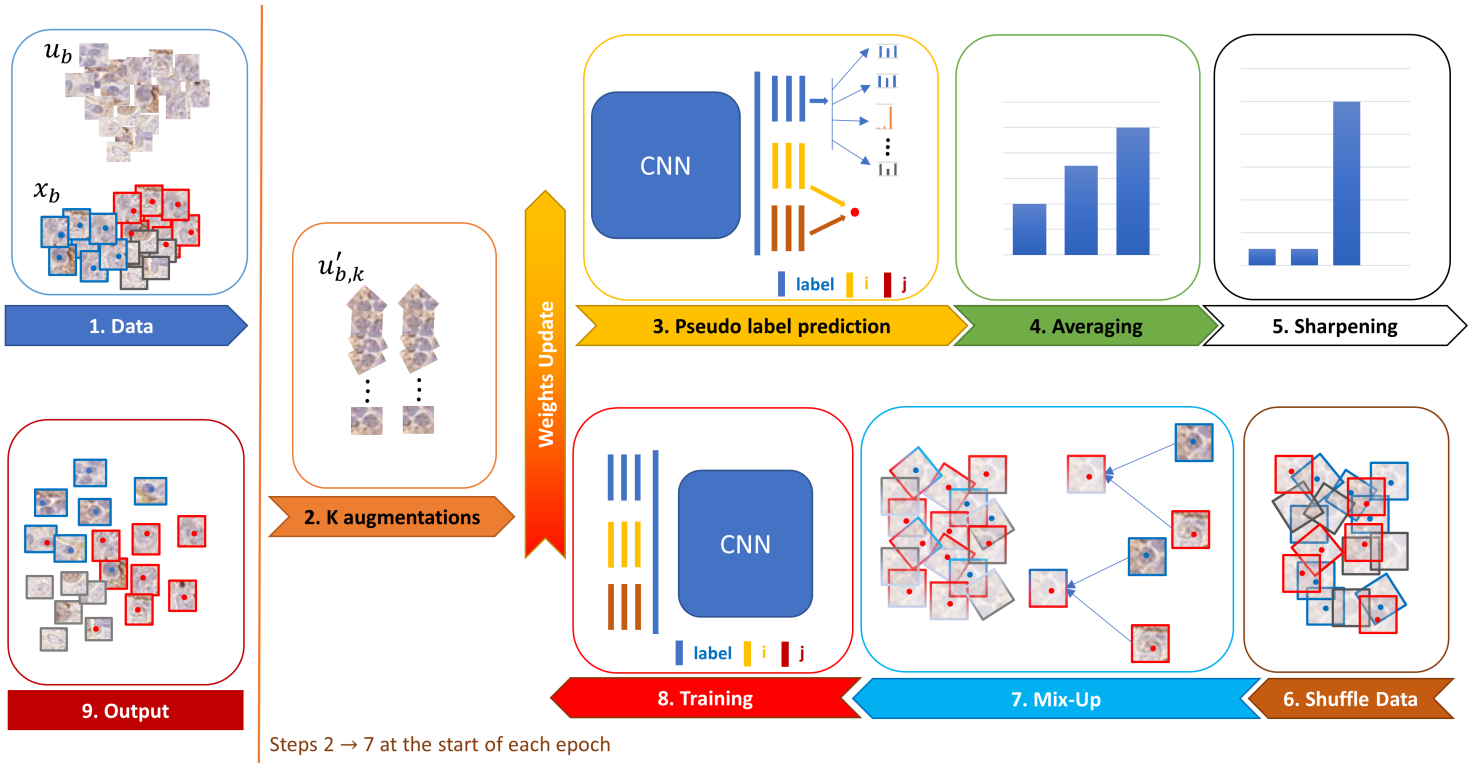


Figure 2.1: The schematic diagram of the HydraMix-Net. The unlabelled data u_b is first subjected to k augmentations (see Section 2.2.3) to generate $u'_{b,k}$ and then process them from the model to generate pseudo labels, after which the predicted labels are averaged and sharpened to minimise entropy in the prediction distribution (see Section 2.2.4). Once pseudo labels are assigned, unlabelled set u_b is mixed-up (see Section 2.2.5) with labelled data x_b to help the model iteratively learn more generalised distributions with pseudo label noise suppression (see Section 2.2.6).

Table 2.1: Frequently used mathematical notations in Chapter 2

Notation	Explanation
X	Set of labelled images
U	Set of unlabelled images
B	Total number of batches
X_b	Input b^{th} batch of labelled images
U_b	Input b^{th} batch of unlabelled images
l^c	One-hot encoded class labels
l^i	Nuclei centroid x-coordinate
l^j	Nuclei centroid y-coordinate
$U'_{b,k}$	b^{th} batch and k^{th} augmentation of unlabelled images
K	Number of augmentations
X'_b	Augmented b^{th} batch of labelled images
l^{uc}	Pseudo labels for unlabelled batch
l^{ui}	Pseudo centroid x-coordinate for image u_i
l^{uj}	Pseudo centroid y-coordinate for image u_i
φ	Model
θ	Weights of the model
$\hat{y}$	Model prediction
T	Temperature hyperparameter
x_i	Instance of image for MixUp
γ	Mixing ratio between pair of images x_i, x_j
W	Concatenated set of augmented images
N	Total number of samples in W
α, η	Parameters of Beta distribution for MixUp
x^m	Mixed image from MixUp
l^{cm}	Mixed label from MixUp
l^{im}, l^{jm}	Centroids from x^m used after MixUp
μ	Weight for combining classification and regression loss
δ, ρ	Weight parameters for SCE loss
$\mathcal{L}^{total}$	Total loss function
$\mathcal{L}^{c-sce}$	Symmetric cross-entropy loss for labelled part
$\mathcal{L}^{uc-sce}$	Symmetric cross-entropy loss for unlabelled part
$\mathcal{L}^{ri}$, $\mathcal{L}^{rj}$	Mean squared error loss terms for nuclei centroid (ri, rj) of the labelled data
$\mathcal{L}^{rui}$, $\mathcal{L}^{ruj}$	Mean squared error loss terms for nuclei centroid (uri, urj) of the unlabelled data
augment(k, U_b)	Augmentation function
sharpening$(l^{uc}, T)_i$	Label Sharpening function

2.2.3 Data Augmentation

During training, the model took an input batch of labelled images X_b from $X = \{X_b\}_{b=1}^{B}$ and unlabelled images U_b from $U = \{U_b\}_{b=1}^{B}$. The number B, representative of the total number of batches, could vary depending on the available labelled and unlabelled images according to the annotation budget. One-hot encoded labels l^c, and point coordinates l^i, l^j represent the type

and nuclei centroid. To generate the pseudo labels l^{uc} and their centroids l^{ui}, l^{uj}, the model applied k augmentations such as horizontal flip, vertical flip, and random rotation to U_b. This yielded an augmented batch U'_b, defined as $U'_{b,k} = \text{augment}(k, U_b)$, where $k \in \{1, ..., K\}$. X_b was also subjected to a single augmentation per image, generating X'_b according to $X'_b = \text{augment}(k, X_b)$, where $k = 1$.

2.2.4 Pseudo Label Generation

To generate pseudo labels l^{uc} for the batch U_b, predictions from the model φ for k augmented images U'_b were averaged out on class distributions. While for pseudo centroids, prediction on only the original image from the model was used. This is due to the fact that after various augmentations, the centroids are not in the same place because of transformations, and hence averaging the centroids of augmentations will lead to incorrect centroids as in Equation (2.1).

$$l^{uc}, l^{ui}, l^{uj} = \begin{cases} \frac{1}{k} \sum_{k=1}^{K} \varphi(\hat{y} \mid U_{b,k}; \theta), & \text{for } l^{uc} \text{ (classification)} \\ \varphi(\hat{y} \mid U_b; \theta), & \text{for } l^{ui}, l^{uj} \text{ (centroids)} \end{cases} \tag{2.1}$$

where φ is the model and θ are the corresponding weights yielding the prediction $\hat{y}$ which was split into patch label l^{uc} and centroids l^{ui} and l^{uj}.

Pseudo Label Sharpening The generated pseudo labels l^{uc} tend to have large entropy in the prediction as a result of averaging different distributions. Therefore, label sharpening [92] was used to reduce or minimise the prediction's entropy by adjusting the temperature like parameter as in Equation (2.2).

$$\text{sharpening}(l^{uc}, T)_i = \frac{l_i^{1/T}}{\sum_{c=1}^{C} l_c^{1/T}} \tag{2.2}$$

where l^{uc} is the categorical distribution of label predictions averaged over k augmentations, T is the temperature hyperparameter falling within the range $(0, \infty)$, which influences the output distribution, where lower values (near 0) enhance the label sharpening effect and higher values lead to a more uniform distribution. i denotes the i-th element in the input distribution, and $c \in \{1, ..., C\}$, which is the full count of the classes. As T approaches 0, a one-hot encoded output is produced, signifying that a lower temperature will result in low entropy output distributions.

2.2.5 MixUP

To bridge the gap between unseen examples, avoid over-fitting, and achieve generalisation in semi-supervised approaches, the MixUP [88] technique was used. Given a pair of images and their labels as (x_1, l_1^c) and (x_2, l_2^c), the images were mixed with their one-hot encoded labels in an appropriate proportion γ. However, the centroids were not mixed due to their numeric nature and transformations. Therefore, centroids from x_1 were used after MixUp as shown in Equation (2.3). Our method used the modified MixUp [91] technique where γ was extracted from a beta distribution. Then max between γ and $1-\gamma$ was taken as γ. This ensures that the maximum of the original image was preserved and output was closer to x_1.

$$\begin{aligned} \gamma &= \max(\text{Beta}(\alpha, \eta), 1 - \text{Beta}(\alpha, \eta)) \\ x^m &= \gamma x_1 + (1-\gamma) x_2 \\ l^{cm} &= \gamma l_1^c + (1-\gamma) l_2^c \\ l^{im}, l^{jm} &= l_1^i, l_1^j \end{aligned} \tag{2.3}$$

where γ is the mixing ratio derived from a Beta distribution which determines the weight on each pair of images and labels in the MixUp. α and η are the parameters of the Beta distribution guiding the shape of the distribution from which the mixing ratio γ is drawn. x^m is the mixed image resulting from the MixUp operation, a weighted combination of the images x_1 and x_2, with the weights being γ and $1-\gamma$. l^{cm} is the mixed label, a weighted combination of the labels l_1^c and l_2^c with weights in line with the image mixing. l^{im} and l^{jm} are the centroids of the mixed image x^m, directly taken from the centroids of the first image x_1, denoted as l_1^i and l_1^j, and not subjected to the MixUp.

To implement the MixUp technique, the sets X_b' and U_b' are initially concatenated to form the set W. After a shuffle operation, the set W is used in the MixUp process. The mixed-up set X_b' is formed with segments of W ranging from 0 to $|X_b'|$, and similarly, U_b' is mixed-up with the segments of W that extend from $|X_b'|$ to N. Here, $|X_b'|$ denotes the size of the augmented mixed-up set X_b', and N represents the total number of samples contained in W.

2.2.6 Noise Reduction

To handle pseudo label noise, symmetric cross entropy (SCE) loss [90] was used for both labelled and unlabelled loss instead of relying on categorical cross-entropy for labelled loss and mean squared loss for the guessed labels. SCE handles the noisy pseudo labels by incorporating cross-entropy terms for labelled loss and reverse cross-entropy for prediction loss. This also provides a way to learn from model predictions instead of relying on given labels, as in Equation (2.4). As with iterative progressive learning, the model gets more

confident in its learning and predictions, which is why more weight is assigned to predictions for unlabelled loss, and in labelled loss, more weight is assigned to original labels.

$$\mathcal{L}^{sce} = \delta(-\sum_{c=1}^{C} q(c|x_m) \log p(c|x_m)) + \rho(-\sum_{c=1}^{C} p(c|x_m) \log q(c|x_m)) \tag{2.4}$$

where δ and ρ are weight parameters that control the significance of input labels and model predictions, respectively. These parameters can be tuned according to the specific problem to balance the influence of class labels and model's predictions on the result. $p(c|x^m)$ and $q(c|x^m)$ represent the true and predicted probability distributions respectively, with each data sample x^m conditioned to belong to a class c. In this context, c represents a class among the C classes of the given classification problem, and m refers to a specific sample within the dataset.

2.2.7 Training

The learning mechanism of the HydraMix-Net jointly optimises the combined loss function for classification and regression to predict label and location tuples for labelled and unlabelled batches as in Equation(2.5).

$$\mathcal{L}^{total} = \mu(\mathcal{L}^{c-sce} + \mathcal{L}^{uc-sce}) + (1-\mu)(\mathcal{L}^{ri} + \mathcal{L}^{rj} + \mathcal{L}^{rui} + \mathcal{L}^{ruj}) \tag{2.5}$$

where $\mathcal{L}^{c-sce}$ represents the symmetric cross-entropy loss for the labelled part, where $\mathcal{L}^{uc-sce}$ represents the symmetric cross-entropy loss for the unlabelled part, both coupled together in weight μ which weights the classification head more to provide more accurate labels. While the $\mathcal{L}^{ri}$ and $\mathcal{L}^{rj}$ are the mean squared error loss terms for the labelled data, whereas the $\mathcal{L}^{rui}$ and $\mathcal{L}^{ruj}$ are the mean squared error loss terms for the unlabelled data for the regression head being weighted by the $(1-\mu)$. While calculating loss for regression heads, the predictions of the classification head were multiplied by regression heads to avoid the loss incorporated by background patches, which is why the classification head was given more weight in the loss term.

2.2.8 Implementation Details

The whole framework was implemented in TensorFlow 2.0, where the base CNN was selected as WidesResNet [98] with an additional three heads i) classification head, ii) two-regression heads. In the classification head, the final output of the WideResNet was the global average pooled and passed through three dense layers of sizes 128, 64 and 32 before the classification layer. In contrast, the

regression heads take the flattened layer results of the output layer, which is then passed through 2 dense layers of sizes 128 and 32 before going to the regression output. The dense layers were activated using the ReLU activation with $l2$ regularisation. The model was optimised with the Adam optimiser with the adaptive learning rate from 0.001 to 0.00001 trained for 100 epochs and a batch size of 32 and $K = 2$. μ, δ and ρ were selected using the empirical evaluation as was set to 0.7, 1.0 and 0.1 for this study.

2.2.9 Experimental Settings

The experimental settings used to test the effectiveness of the HydraMix-Net were i) fully supervised, ii) partial data iii) semi-supervised. In the first one, all the available data was used to train a simple CNN, i.e., WideRes-Net [98], while in a partial setting, WideRes-Net was trained on partially labelled data. Lastly, HydraMix-Net used a semi-supervised approach for training where both labelled and unlabelled data were used in a way discussed earlier in section 3. Further, for labelled and unlabelled data, we tested different configurations from 50 labelled images to 100, 200, 300, 500, 700 and so on.

2.2.10 Evaluation metrics

Patch-wise F1-score was used for evaluation purposes of the experiments as it is a measure of a test's accuracy that considers both the precision and the recall to compute the score.

Table 2.2: Test F1-score of the HydraMix-Net and partial data approaches with various amounts of labelled data provided.

labelled examples	**50**	**100**	**300**	**500**	**700**	**1000**	**3000**
Simple CNN	0.62	0.70	0.76	0.83	0.85	0.84	**0.90**
HydraMix-Net w/o SCE	0.66	0.70	0.70	0.35	0.35	0.35	0.35
HydraMix-Net	**0.66**	**0.80**	**0.81**	**0.85**	**0.85**	**0.86**	0.89

2.3 Results and Discussion

Table 2.2 shows the accuracy achieved by the HydraMix-Net in contrast to the simple CNN on partially labelled data, e.g., when provided with the random 50 labelled examples, the simple CNN model underperformed by achieving 62% accuracy where the HydraMix-Net leveraged the unlabelled data and achieved superior performance with 66% accuracy. Similarly, when increasing the data from 50 labelled examples to 100 and 300, the HydraMix-Net achieved higher performance and reached up to 81% accuracy, while the simple CNN model trained on only these labelled examples only gave the best performance of

76% accuracy, which shows higher efficiency of our approach in scarcity of the labelled examples. When trained with all the data, the highest accuracy achieved is 90%, where this threshold is reached by approximately 3000 labelled data by both techniques. However, it is interesting to see that performance plateaus after a certain number of labelled examples. Figure 2.2 shows a) the confusion matrix showing the performance of HydraMix-Net when trained on 100 labelled examples and b) the cell centroid detection in for the test set shown in red. Figure 2.3 shows the actual predictions for the HydraMix-Net for the 100 labelled training set. We can notice that the model is performing good in terms of identifying the type of cell but is failing to accurately find the centre of the cell. This is due to the inherent bias included in the dataset as most of the nuclei were extracted around the centre of the cell.

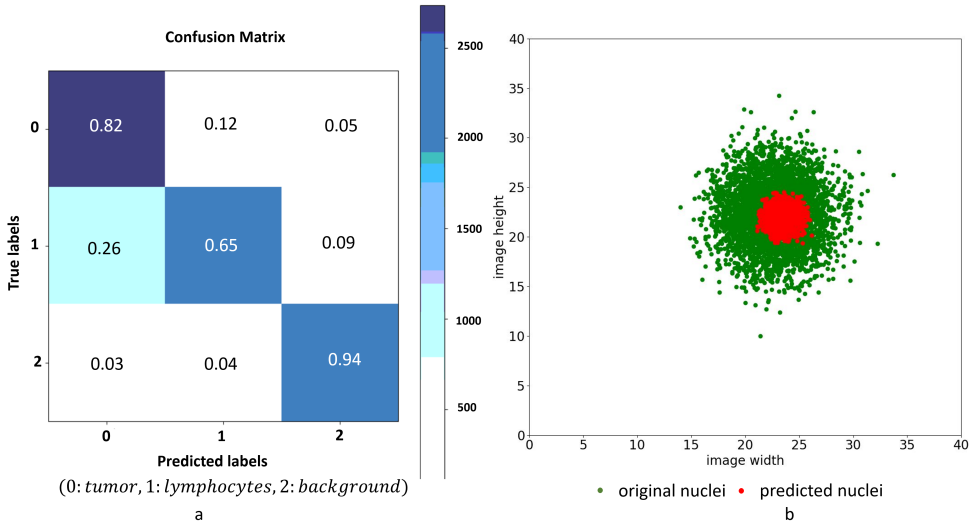


Figure 2.2: (a) Represents the confusion matrix for the HydraMix-Net while (b) Represents the prediction and distribution of the centroid in the HydraMix-Net trained on 100 labelled instances.

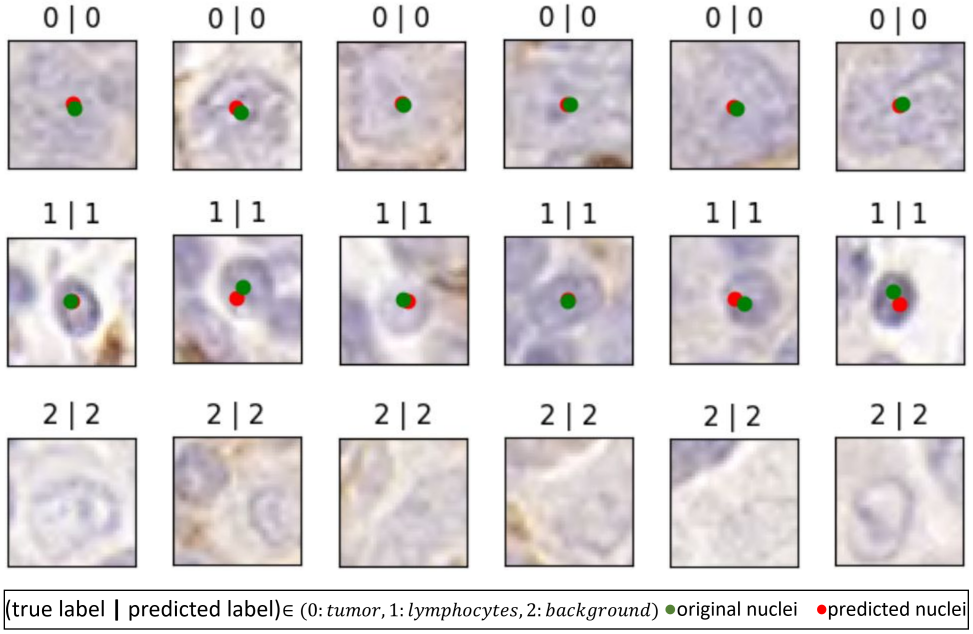


Figure 2.3: The prediction of labels and distribution of the centroid on an example set where the HydraMix-Net was trained on 100 labelled examples.

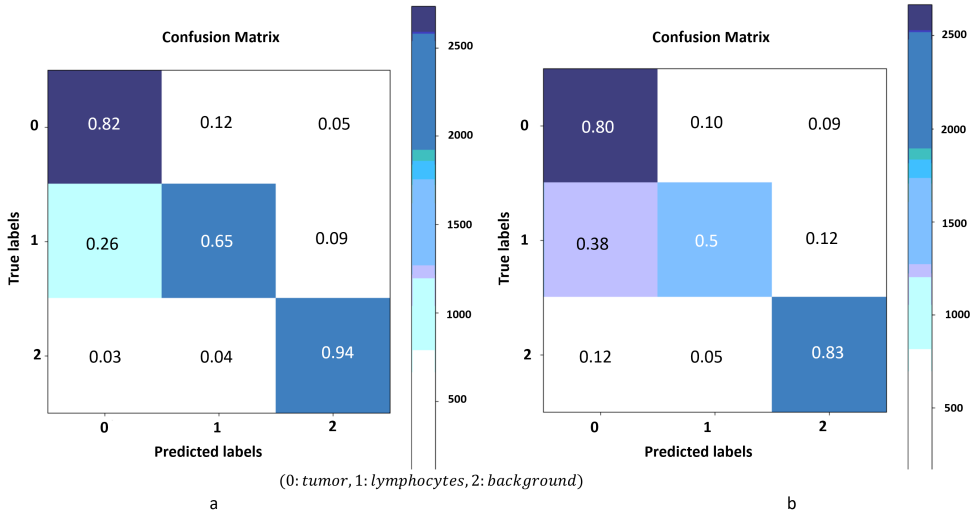


Figure 2.4: (a) Represents confusion matrix for the HydraMix-Net while (b) represents confusion matrix for simple CNN model trained on partial data of size 100. It can be seen from the matrix that false positives in the HydraMix-Net are less than false positives in partial data.

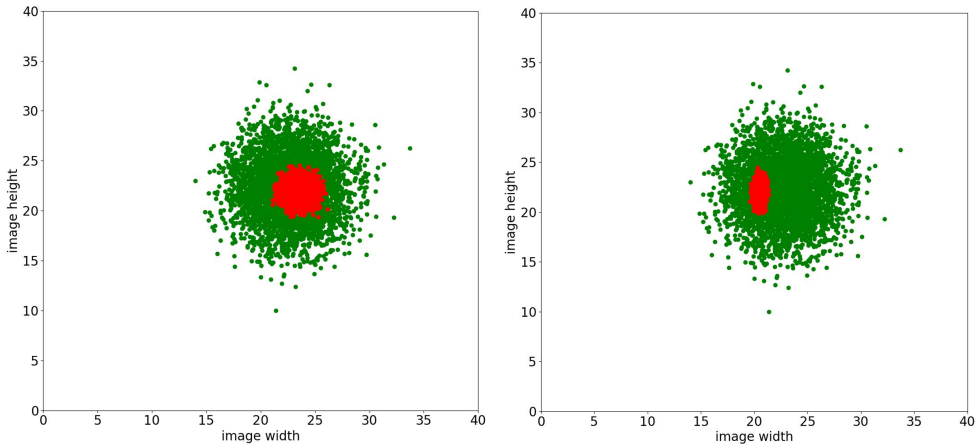


Figure 2.5: (a) Represents prediction and distribution of centroid in the HydraMix-Net trained while (b) shows the distribution of centroid learned by simple model on partial data of size 100.

In order to compare the performance gains of HydraMix-Net over the vanilla CNN we can look at the confusion matrix show in Figure 2.4 where it shows comparative results models trained using only 100 labelled set. Similarly, Figure 2.6 shows the confusion matrix for both models train on 300 labelled images set. It is evident from both the figures that the our approach performs superior to the vanilla CNN in differentiating between the tumour and lymphocyte cells. Furthermore, apart from correctly classifying the cell types we can see from the Figure 2.5 that in the scarcity of enough labels the simple CNN performance in cell detection is inferior to the proposed approach. The simple CNN fails to diversify the x and y coordinates and predicts within a small vicinity. Further, it can be seen that nuclei locations are biased towards the patch's centre because of the training data inherent biases. Figure 2.7 shows predictions for the HydraMix-Net for the 100 labelled set in (a) while 300 labelled set in (b), it can be seen that the model is learning to classify the patch accurately along with nuclei prediction among tumour, lymphocytes and background patches.

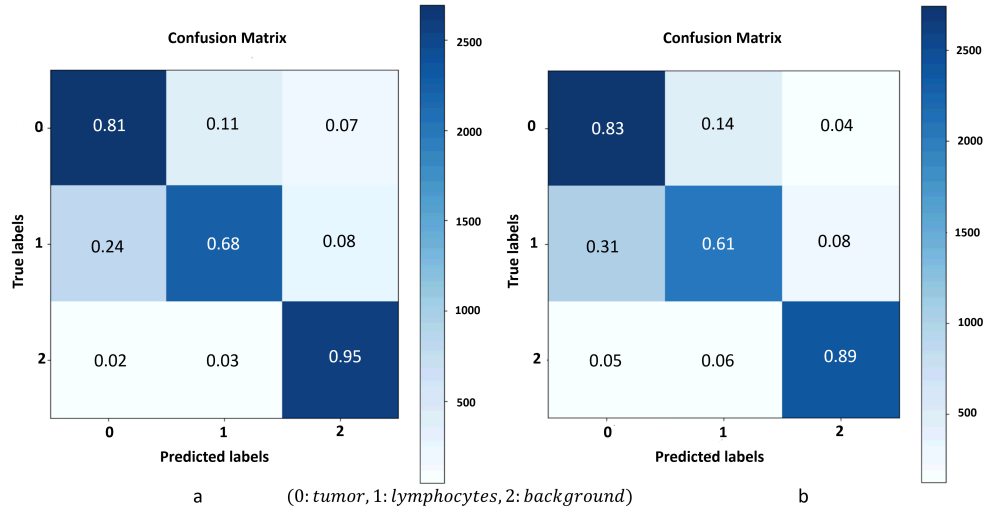


Figure 2.6: (a) Represents confusion matrix for the HydraMix-Net while (b) represents confusion matrix for simple CNN model trained on partial data of size 300. It can be seen from the matrix that false positives in the HydraMix-Net are more in case of the tumour, while for background and lymphocytes, false positives in partial data training are in abundance.

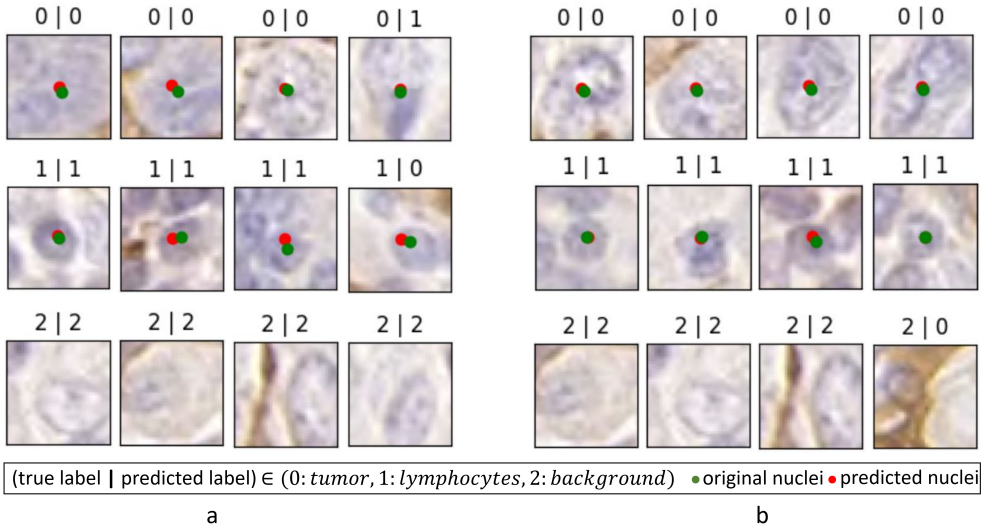


Figure 2.7: (a) Shows prediction and distribution of the centroid in the HydraMix-Net trained on 100 labelled examples (b) shows prediction and distribution of the centroid in the HydraMix-Net trained on 300 labelled examples.

2.3.1 Noise Reduction

In this chapter, we have included a symmetric cross-entropy loss to reduce the effect of the noisy pseudo label and ease our learning capabilities. Labelled

data was given more weightage while computing the SCE loss because there is less noise in the original labelled set and labels are not much noisy (i.e., MixUp doesn't add much noise in the labels), while in the case of unlabelled data loss, the new predictions were given more importance as it was believed that the newly predicted values were more accurate as the model has learned and improved the previous mistakes. Hence, we experimented with a few configurations to see the effectiveness of SCE loss, and it turned out that the addition of SCE made the model learn more than simple cross-entropy and l^2 loss, as it can be seen in the Figure 2.8. Interestingly, when more data is provided, the chances of model overfitting increase as training is sensitive towards the pseudo label noise and starts to overfit the dominant class distribution, as seen in Table 2.2. Hence, adding SCE improves the overall learning of the approach by making this technique less susceptible to pseudo label noise.

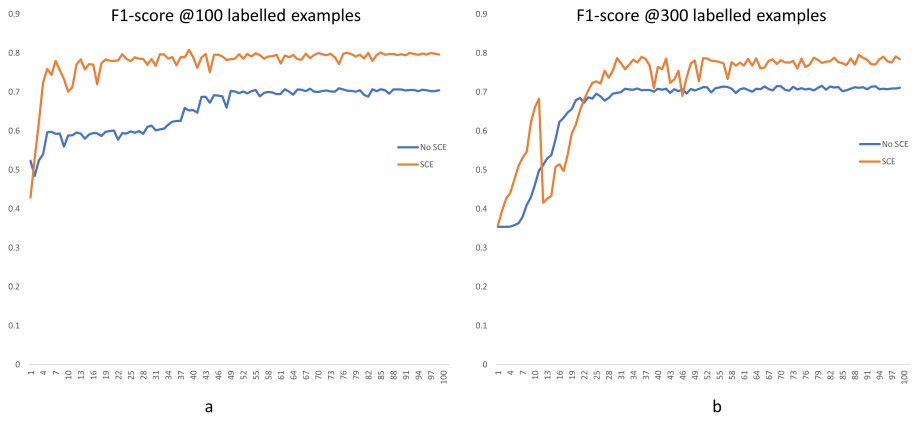


Figure 2.8: (a) Represents the F1-score curves of the models trained with 100 labelled examples with the orange line showing the model with SCE and the blue line showing the model without SCE, and it can be seen that the model without SCE under-performs the model with SCE with a margin of 10% in F1-score. Similarly, (b) Represents the F1-score curves of the models trained with 300 labelled examples, with the orange line showing the model with SCE and the blue line showing the model without SCE, and it can be seen that the model without SCE under-performs the model with SCE with a margin of 5% in F1-score.

2.3.2 Knowledge vs F1-score

In this chapter, we have also examined the behaviour of increasing the knowledge, i.e., increasing labelled samples while training corresponding to the

model's F1-score. It has been shown through experiments that increasing knowledge does increase F1-score. As with a more accurate labelled data training model, it gets a chance to learn it more accurately and performs better on validation and test sets, as seen in Table 2.2 and in Figure 2.9.

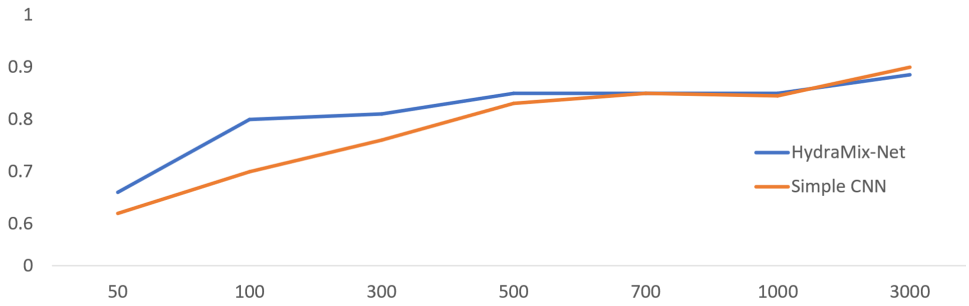


Figure 2.9: Represents the increase in knowledge vs increase in F1-score where the knowledge is the number of labelled samples which can help the model to learn more accurately on the true labels, and it can be seen that the HydraMix-Net leverages semi-supervised approach and outperformed the simple CNN trained on partial data.

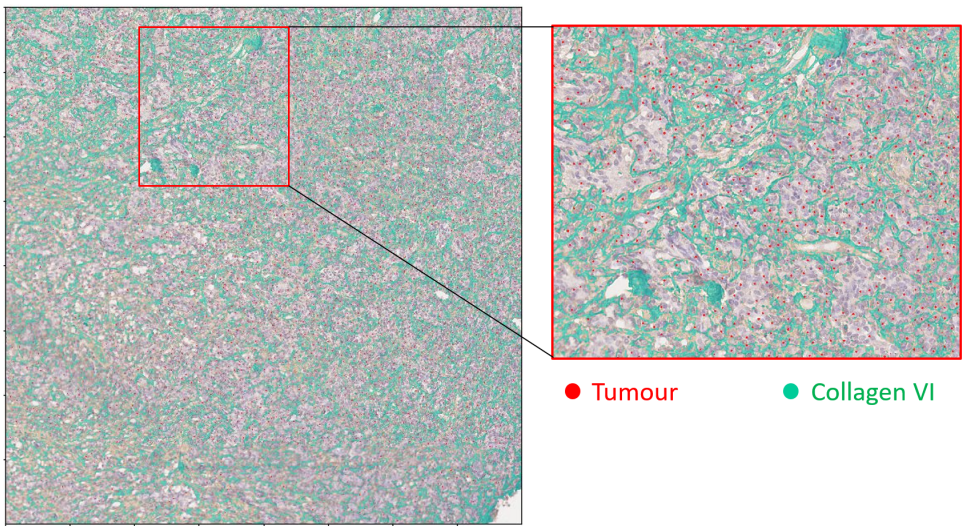


Figure 2.10: HydraMix-Net prediction of tumour cells in a large ROI show in red dots while the cyan colour shows the collagen VI.

2.4 Chapter Summary

In this chapter, we presented a novel end-to-end novel multi-task SSL approach for simultaneous classification and localisation of nuclei in DLBCL. Further, we plan to extend this work by improving the technique with the help of strong

augmentations and validating the performance of our HydraMix-Net on larger cohorts from multiple tumour indications. The cell detection and classification may also help in performing follow-up analysis like survival prediction and understanding the spatial arrangement of malignant cells within the tumour micro-environment with collagen VI as seen in Figure 2.10.

Chapter 3

Semi-Supervised Learning for Segmenting Tissue Regions and Nuclei Histology Images

3.1 Introduction

Segmentation of fundamental objects and regions in histology images is key to several downstream analysis tasks in computational pathology (CPath) [99, 100] e.g., cancer type classification [101–104], tumour and glandular segmentation [105], and other tasks like mutation prediction [54, 106]. Their utility is not limited to diagnosis but has also been employed for prognostic purposes, e.g., tumour infiltrating lymphocytes (TILs) have been found to be a significant prognostic bio-marker in various types of tumours [107]. Similarly, tumour progression has been linked with the interaction between the tumour epithelial cells and tumour associated stroma [108]. Hence, it is important to segment different types of histological objects precisely as their quantification is vital to downstream analysis.

Machine learning based traditional methods accomplished this task using different hand-crafted features, e.g., colour [109], texture [110, 111] and morphological features [112]. Deep learning (DL) algorithms have recently gained increasing attention in semantic segmentation due to their superior performance on natural and medical images [113–116]. However, DL methods are known to be "data hungry" and require a large amount of annotated data. Precise annotation of histology images is expensive and laborious, requiring up to ∼5-6 hours of an expert histopathologist's time to annotate one whole-slide image (WSI) [117]. To alleviate the annotation burden, other modes of training have been proposed, such as patch based segmentation [47, 55], coarse segmentation [46, 118] and interactive segmentation [99, 119]. However, these methods still require large-scale weak annotations involving a human expert.

Semantic segmentation is a pixel-level classification task of predicting labels for each pixel using pixel values. Most of the early DL methods were based on fully convolutional networks (FCN) [120] where pooling layers aggregate the information by focusing on "what" rather than "where", resulting in loss of spatial information. The subsequent studies addressed this shortcoming by using pooling layers with more advanced techniques involving skip connections, encoders and boundary information. As semantic segmentation is more than just assigning labels to pixels, it inevitably requires some contextual information along with knowledge of colour, edges and resolutions. In this regard, algorithms like UNet [113], PSPNet [121], HRNet [122], and DeepLab-v3 [114] use techniques like encoder-decoder architecture, wider receptive fields and dilated/*atrous* convolutions to improve the segmentation performance. More recently, another line of work focused on transforming the task of semantic segmentation to sequence-to-sequence prediction, where the self-attention mechanism is introduced using transformers [123] to encode the global context in each layer [115, 124] for subsequent decoding. However, a downside of using transformer based techniques is their computational complexity.

On the other hand, semi-supervised learning (SSL) can train DL models with a small set of annotated data by leveraging the unlabelled data for better representation learning, boosting performance. SSL methods consist of different techniques to incorporate unlabelled data for learning, including pseudo labelling [125–127], generative adversarial modelling [86, 128–130], consistency training [127, 131–133] and entropy minimisation [68, 89, 134]. However, SSL methods have an additional issue related to overfitting small labelled input data, which may lead to poor generalisation. Self-supervised learning enables SSL to learn more robust representation by enabling supervised tasks out of unlabelled data, thus improving the overall performance.

3.1.1 Semantic Segmentation

The transformation of pixel values of an image to class labels using high level features is known as semantic segmentation and is fundamentally a challenging task. FCN extracts meaningful visual hierarchical features for various computer vision tasks, e.g., classification [135], segmentation [120] and object detection [136]. However, due to the pooling layers, spatial information is lost in aggregation, which is vital in segmentation tasks and results in smaller output [120]. Encoder-decoder based architectures solve this issue by recovering and refining the output spatially in a step wise fashion [62, 137, 138]. Further improvements can be made possible with the help of skip connections which results in more refined boundaries and confident predictions [113]. However, the downside of the encoder-decoder architectures is a limited

receptive field resulting in missing long-range dependencies. Dilated/atrous convolutions [46, 114, 139, 140], spatial pyramid pooling [105, 121, 141, 142] and attention based algorithms [115, 143–145] can enable aggregation of context by using larger receptive fields or maintaining spatial information. More recently, attention mechanism [123] has been used to replace convolutions' limited local receptive field with global contexts using transformers. Images are transformed into a sequence of patches for transformer [146] to process as transformers capture more consistent global contexts due to their self-attention mechanisms [115, 124, 147]. Despite the advancements and improvements in semantic segmentation, the bottleneck for high accuracy still remains to be dependent on pixel-wise annotations.

3.1.2 Semi-Supervised Learning

Semi-supervised learning (SSL) exploits the unlabelled data instead of limited labelled data for improving the model performance and internal feature representation. Recently SSL based methods have been widely adopted in the computer vision domain [148]. Popular SSL techniques include pseudo labelling [125–127] where the model trained on limited data is used to predict the labels for unlabelled data known as pseudo labels. Generative adversarial based methods improve the generalisability of the trained model using various perturbations in the direction of maximum vulnerability, resulting in aligning the distributions of labelled and unlabelled input in latent space [86, 129, 130, 149]. Data interpolation based methods aim to augment input space to create perturbed linear inputs for training models [82, 91, 150], Temporal ensembling based methods aim to ensemble predictions over the epochs using momentum/moving average to enforce consistency between the predictions [85, 151]. Self-supervised learning based consistency training aims to contrast the unlabelled input using pre-text tasks for learning meaningful representations [127, 133, 152–154] and entropy minimisation based method aims to maximise label assignment to either of the labels [68, 89, 134].

3.1.3 Self-Supervised Learning

Self-Supervised learning is a hybrid technique of representation learning in the machine learning paradigm where robust and high quality data representations are learned from unlabelled data. The concept behind self-supervised learning is to build supervised tasks out of unlabelled data, known as pretext tasks. Self-supervised learning can help SSL learn high quality representations, thus making it easy to preserve and transfer valuable key insights available in data for downstream analysis. The key here is that the task-agnostic pre-trained networks can help better with downstream tasks than the task specific pre-

trained networks. Self-supervised learning started with Autoencoders [155], where neural networks learned to compress and reconstruct the input data and were often used for feature extraction. Famous examples in this line of work were Word2Vec [156] and GloVe [157], which were used to map words to word embeddings for natural language processing tasks. Moving on from autoencoders was the concept of learning similarities from the input data using self-organising neural networks, e.g., Siamese networks [158]. These networks were used extensively in computer vision, from signature verification to face recognition. Similarly, in the early days, other techniques, e.g., restricted Boltzmann machines, autoregressive modelling and metric learning, were also used to learn good representation. Self-Supervised learning methods can be broken down into two broad categories, i) self-prediction and ii) contrastive learning.

Self-prediction: Given the input data with missing parts, the self-prediction tasks involve predicting missing parts using the available parts. E.g., given a masked input image with missing areas, the task is to generate the whole image using the available clues. Supervised loss functions can be used to regularise the training as the input is being used as a label. Self-prediction can be thought of as an intra-sample prediction task. There are different methods of self-prediction, e.g., autoregressive generation, masked generation, innate relation prediction and hybrid self-prediction.

- **Autoregressive Generation**: Innate sequence prediction is the type of autoregressive model where it predicts the future based on the past. It can be applied to 1-dimensional input data, e.g., audio [159], text [160] and raster scans [161].
- **Masked Generation**: Masking a random portion of the input data and predicting the masked portion using the unmasked portion. Popular examples are masked language modelling [162] and denoising autoencoder [163], context autoencoder [164] etc.
- **Innate Relationship Prediction**: Distorting the innate nature of the input data without disturbing the semantic meaning and predicting it using the correct pairs. Innate relations can be the relative position of objects, rotation, jigsaw puzzle [165] etc.
- **Hybrid Self-Prediction Models**: Mixing the different generation models with other self-prediction techniques, e.g., VQ-VAE with autoregression [166]

Contrastive learning: Learning by contrasting pairs of similar (positive) and dissimilar (negative) images for improved representation learning is known

as contrastive learning [152, 167, 168]. Several loss functions have been proposed from maximum margin loss [169], triplet loss [170], N-pair loss [171] to contrastive predicting coding (CPC) [172] proposing mutual information based InfoNCE loss to improve contrastive learning. Contrastive learning has been used in both supervised and unsupervised learning tasks in conjunction with self-supervision [152, 153, 173]. Recently, it has been established that using more accurate positive and negative pairs along with larger batch sizes improves the quality of learned representations with heavy augmentations. Memory banks are adopted when large batches are not computationally feasible (i.e., don't fit the GPU memory) for contrastive loss using a large set of negative samples. It is the task of learning the relationships among the data points. These relationships can be of multiple types, e.g., multiple views of the same object, augmented versions of the same object and so on. Contrastive learning can be thought of as an inter-sample prediction task. The goal of contrastive learning is to learn such embedding in hyperspace where similar data points are clustered together. The main subtypes of contrastive learning are inter-sample classification, feature clustering and multiview coding.

- **Inter-Sample Classification**: Given similar (i.e., positive) or dissimilar (i.e., negative) data samples, find out the similar ones to the reference samples (i.e., anchor). Similar data points can be made using the distorted versions of the original input using transformation etc. For training, a simple supervised loss can be used, but there has been a lot of research on building contrastive loss functions, e.g., contrastive loss [169], triplet loss [170], infoNCE loss [172] etc.

- **Feature Clustering**: Feature clustering works by clustering the same data samples in feature space using clustering algorithms, e.g., K-NN, on the learned feature representations. Pseudo labels were assigned to the data samples based on their centroids, and then intra-sample contrastive learning was used for training, e.g., DeepCluster [174] and InterCLR [175].

- **Multiview Coding**: Multiview coding involves applying contrastive loss, especially infoNCE, on multiple views of the same input data [176].

3.1.4 Semi-Supervised Semantic Segmentation

SSL based semantic segmentation approaches utilise the aforementioned techniques to extract knowledge from unlabelled data. Recently, CutMix, MixUp, and CutOut based augmentation techniques were used together with the student-teacher model, where consistency was enforced between the mixed predictions [177]. Guided collaborative training (GCT) by [178] performed

network perturbations with the help of different network initialisation and enforced the dynamic consistency constraint between the predictions. Cross-consistency training (CCT) by [132] performed perturbations on the main encoder's features and enforced consistency over the multiple decoders' output, making it robust to various perturbation types. Context-aware consistency by [133] proposed directional consistency loss for contrasting different contexts by cropping two overlapping patches of the same input to improve the representation learning. Recently, in the field of computational pathology, a few methods for semi-supervised semantic segmentation have been proposed. [179] proposed a semi-supervised method for signet detection using the help of self-supervised learning for label generation. [180] proposed a two stage SIM-FixMatch approach utilising self-supervised learning in the first stage and then using FixMatch for pseudo label generation along with consistency regularisation. [181] proposed an exponential moving average (EMA) student-teacher framework where the model is trained using the noisy labels to enforce consistency over similar and dissimilar patch pairs. Cross-patch dense contrastive learning by [68] proposed a student-teacher based method to enforce EMA based consistency over predictions and to improve the internal representations. The pixel-wise contrastive loss was applied to background and foreground patches to improve the internal feature representations.

In this chapter, we present a novel consistency based SSL method for semantic segmentation which leverages unlabelled data in varying contexts and CNN feature perturbations such as dropout and noise using self-supervised learning techniques like consistency regularisation and cross-consistency training. Consistency regularisation is enforced by using context-aware contrastive learning in changing contexts, and cross-consistency training is used to handle CNN feature perturbations along with entropy minimisation for confident predictions. The primary purpose of consistency regularisation is to enforce the CNN model to output consistent predictions for unlabelled data under changing conditions. For consistency to work effectively, input space must hold the cluster assumption constraint, i.e., the same label is most likely to be shared among the nearby samples, thus forming a cluster. Therefore, high density regions correspond to clusters (i.e., samples with the same labels), whereas the low density regions are separation spaces (i.e., object boundaries). As for histology and natural images, the pixel space might not hold the constraint of cluster assumption, as seen in distance map of the RGB space in Figure 3.1. The low density regions (i.e., high average distance) do not align well with the class boundaries in most of the scenarios, e.g., in 1^{st} row, we observe low density regions throughout the image, while, in the last row there exist a cluster of high density regions for a foreground object, i.e., road. However, the cluster assumption holds in CNN encoders latent feature space [132], as we

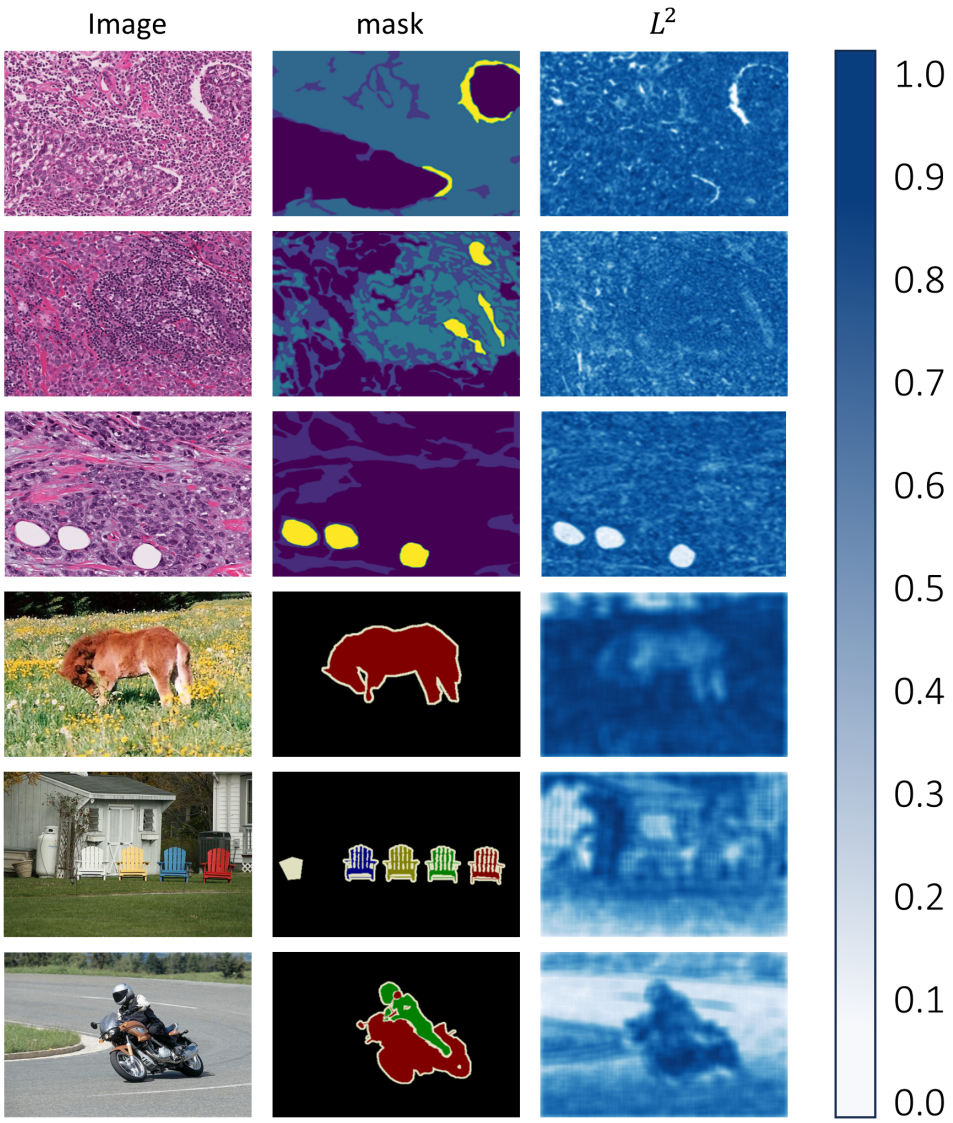


Figure 3.1: (1^{st} column) Example images from histological (BCSS) and natural (PASCAL VOC 2012) datasets; (2^{nd} column) Respective masks showing the foreground and background objects with boundaries; (3^{rd} column) Distance map (i.e., Average Euclidean distance L^2) between the central patch of size 21×21 with four overlapping patches in the immediate neighbours in RGB colour space. Note that the darker blue colour represents the low density regions corresponding to the high average distance.

show and discuss later in Figure 3.10. Therefore, we applied the CNN feature perturbations to the CNN encoder's output rather than the input images. Also, due to the limited labelled data, the CNN model may become overly dependent on just context (i.e., its surrounding) overlooking the objects themselves, losing object-awareness [133]. Changes in context refers to the changes or shifts in the conditions, environment, or circumstances surrounding an image or scene that can impact the interpretation, understanding, or analysis of

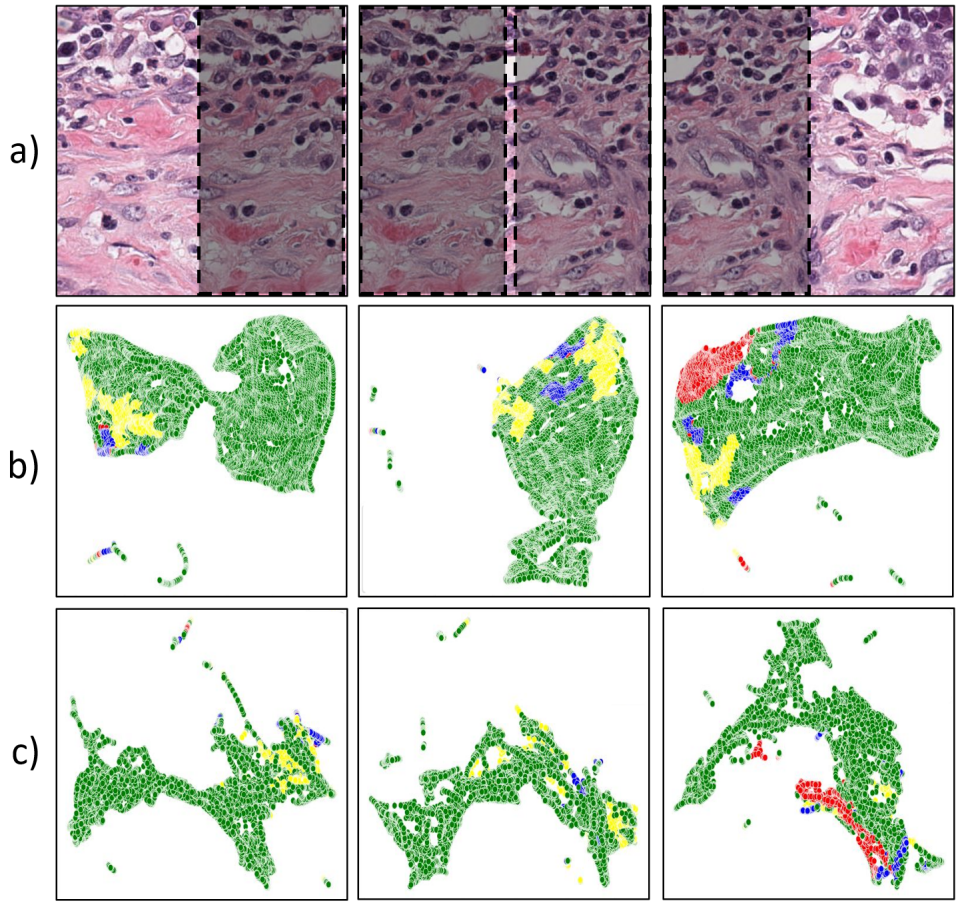


Figure 3.2: (a) Images from the BCSS dataset with overlapping regions cropped sequentially (i.e., dashed grey boxes) from the same image to mimic changing contexts; (b) UMAP visualisations of CNN features embeddings extracted from a fully supervised model; (c) UMAP visualisations of CNN feature embedding extracted from our semi-supervised model. The semi-supervised model benefits from unlabelled data, enabling it to capture the underlying data distribution more comprehensively, resulting in more consistent representations (i.e., roughly the same location) for each class as compared to the CNN feature embeddings obtained from a fully supervised model. Note that the CNN feature embeddings are represented in the same UMAP space where dots with the same colour represents CNN feature embedding from the same class.

that image. Therefore, to enforce consistency against changes in contexts, we present context-aware contrastive learning, which helps the CNN model learn high-level semantic CNN features by contrasting the positive (i.e., similar) and negative (i.e., dissimilar) pairs of images in different contexts. As shown in Figure 3.2, under varying contexts, the CNN model trained in a fully supervised manner is unable to produce consistent CNN feature distributions as compared to proposed method, **C**onsistency **R**egularisation in varying **C**ontexts and **F**eature **P**erturbations for semi-supervised semantic segmentation of Histology Images (CRCFP) with consistent feature distributions. While context-aware consistency brings robustness to changing contexts, cross-consistency training can help the CNN model learn robust invariant CNN feature representation to small perturbations. While context-aware and cross-consistency training regularisation can bring consistency in the CNN encoder's features representations, it often fails to optimise the pixel classifier leading to less confident prediction maps. Finally, entropy minimisation coupled with the aforementioned techniques helps the CNN model acquire high quality and confident predictions. We extensively evaluated our CRCFP on two publicly available histology image datasets BCSS [45] and MoNuSeg [69] for two different semantic segmentation tasks, i.e., tissue region segmentation and nuclei segmentation. In summary, our contributions are as follows:

- We present a consistency regularisation based SSL method against varying contexts and perturbations using a novel combination of context-aware consistency loss and cross-consistent training for feature generalisability.
- To improve the confidence of final predictions for pseudo labelling, entropy minimisation is employed on top of context-aware and cross-consistent regularisation.
- We demonstrate our method on two different semantic segmentation tasks, i.e., cancer region and nuclei segmentation on two publicly available large histology datasets.
- Extensive experiments showed superior performance of our method outperforming the state-of-the-art (SOTA) SSL methods with extensive ablation studies.

3.2 Materials and Methods

3.2.1 Data

We evaluated our framework on two publicly available datasets, the Breast Cancer Semantic Segmentation (BCSS) [45] and Multi-organ Nucleus Segmentation Challenge (MoNuSeg) [69] dataset for semantic segmentation. The data

was obtained from the respective challenge pages hosted on Grand Challenge for Medical Image Analysis website (https://grand-challenge.org/).

MoNuSeg. The MoNuSeg challenge was organised as a MICCAI 2018 satellite event and contains 21,623 annotated nuclei from 30 H&E stained images for training and contains 7223 annotated nuclei from 14 H&E stained images for testing purposes. Annotations were done by engineering students, and then an expert pathologist served as quality control for the annotated nuclei. Each image is of size 1000 $\times$ 1000 extracted from a WSI scanned at 40$\times$ resolution of an individual patient obtained from The Cancer Genome Atlas Program (TCGA) [25]. WSIs are sampled from 18 different centres and seven different organs, including the breast, liver, kidney, prostate, bladder, colon and stomach, with various tumour stages.

BCSS. The BCSS challenge was conducted in 2021 and contains over 20,000 annotated regions of interest (ROI) from 151 H&E stained WSIs with the same number of patients from TCGA [25]. 25 annotators, including pathologists, residents, and medical students, helped annotate this large scale data into 25 refined categories, which are later merged into five broad categories as a tumour, stroma, inflammatory, necrosis, and others. For this work, we have used the same five broad categories by relabelling the regions and then split them into training and test centres according to the [45] where there were 14 centres for training and seven centres for testing.

Data Preparation

In order to validate the CRCFP framework, we evaluated it against different labelled data proportions for each dataset. Where for BCSS different data proportions were created using different centres (hospitals) to make training more susceptible to variation in colours enabling domain shift. DL methods often fail to perform well on samples from different domains (centres), mainly due to domain shift, this also makes it a domain generalisation problem. Therefore, the training set was divided into data portions by diving the total training centres as 1/1 (full), 1/2 (half), 1/4 (quarter), and 1/8 (one-eighth) centres where 1/8 results in training images coming from only 1 centre, while the test set remains intact as it is. Similarly, for 1/4 (quarter) training images comes from 4 centres and 7 centres for 1/2 (half). For MoNuSeg, different label proportions were based on training images themselves and are then divided into 1/1, 1/8, 1/16, and 1/32 proportions to make it comparable to the work of [68]. Further, this whole process is repeated using 3 different random seeds and then the results are reported using mean aggregation with standard deviation.

To facilitate the reader's comprehension, frequently used mathematical notations are listed and defined in Table 3.1

Table 3.1: Frequently used mathematical notations in Chapter 3

Notation	Explanation
L	Set of labelled images
U	Set of unlabelled images
N	Total number of labelled images
M	Total number of unlabelled images
H	Height of image
W	Width of image
D	Depth of image
C	Number of classes
x^l	Labelled image
x^u	Unlabelled image
y^l	Labelled image mask
$g(\cdot;\theta^g)$	shared decoder
$h(\cdot;\theta^h)$	shared encoder
$f(\cdot;\theta^f)$	Feature extractor
C^f	Pixel classifier for final prediction
$\phi(\cdot;\theta^z)$	Projection head
θ^g	Parameter of the shared decoder
θ^h	Parameter of the shared encoder
θ^f	Parameter of the feature extractor
θ^p	Parameter of the pixel classifier
θ^z	Parameter of the projection head
f^l	Feature maps for x^l
f^u	Feature maps for x^u
x^{u1}, x^{u2}	Overlapping patches extracted from x^u
$\hat{y}^l$	Class map for x^l using pixel classifier C^f
$\hat{y}^u$	Class map for x^u using pixel classifier C^f
φ^u	Projected feature map for x^u
$\mathcal{L}^{cont}(\varphi^{ui}, \varphi^{uj})$	Directional contrastive loss $x^{ui} \rightarrow x^{uj}$
λ	Threshold for positive feature filtration
η	Negative samples for contrastive loss
$\mathcal{U}$	Uniform distribution
$\mathcal{L}$	Overall loss for training the framework
$\mathcal{M}$	Binary mask for positive and negative samples
$\mathcal{L}^{sup}$	Cross-entropy loss for supervised training
$\mathcal{L}^{t-cont}$	Total directional contrastive loss
$\mathcal{L}^{cross}$	Cross-consistency training loss
$\mathcal{L}^{ent}$	Entropy loss for improving prediction confidence
w^{sup}	Weight for supervised loss
w^{t-cont}	Weight for direction context-aware contrastive loss
w^{cross}	Weight for cross-consistency training loss
w^{ent}	Weight for entropy loss

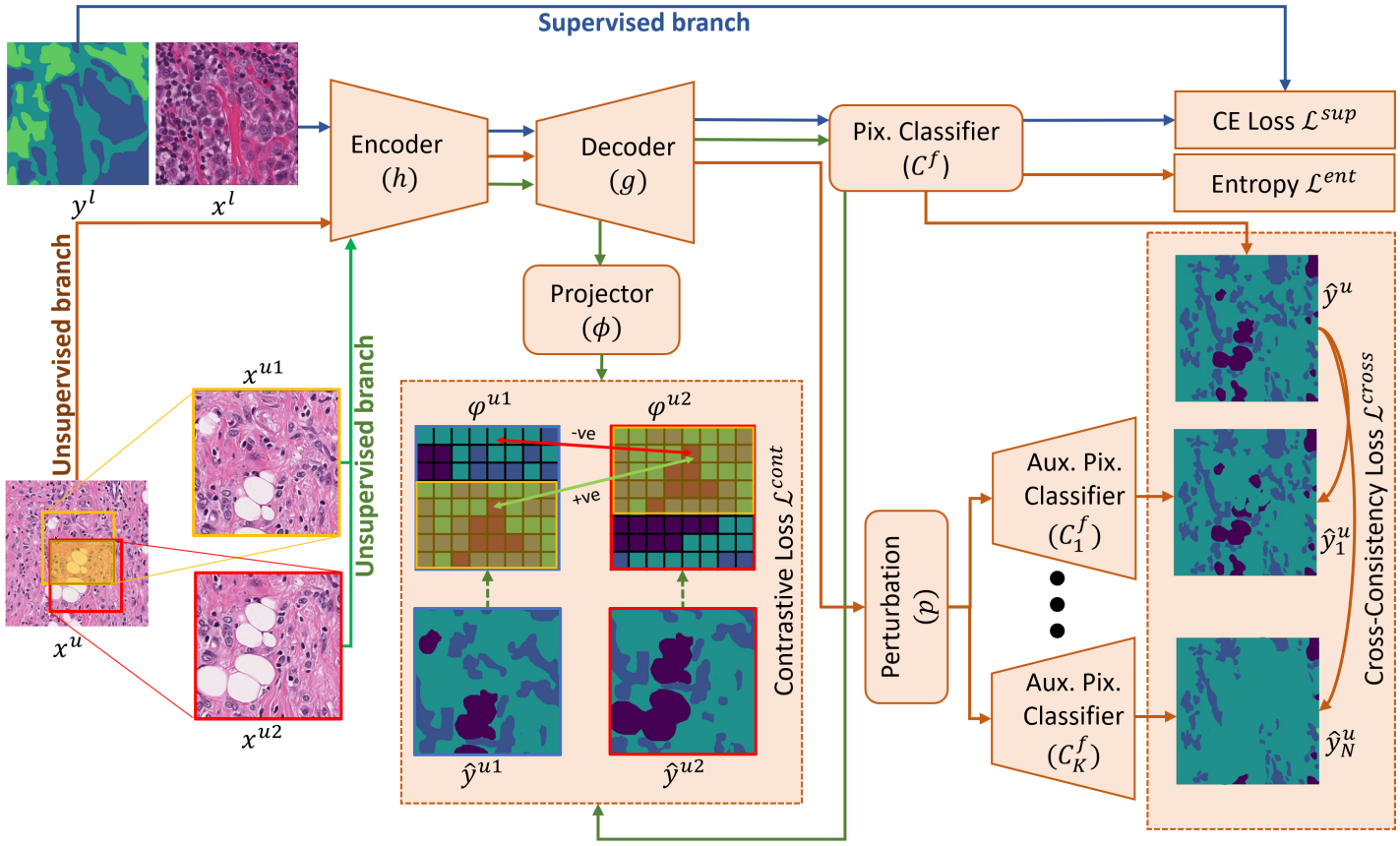


Figure 3.3: The proposed framework, called CRCFP, consists of an encoder and decoder trained in a supervised manner using cross-entropy (CE) loss for labelled instances (represented by blue arrows). For unlabelled instances, the framework employs a combination of cropped patches with partial overlap (represented by green arrows) and the input image (represented by brown arrows), which are fed through the encoder. The green arrow pathway shows the contrastive learning pathway where the encoded features are projected to a lower dimension before applying directional consistency loss. Similarly, the brown arrow pathway shows the cross-consistency training where encoded features undergo various perturbations comparisons.

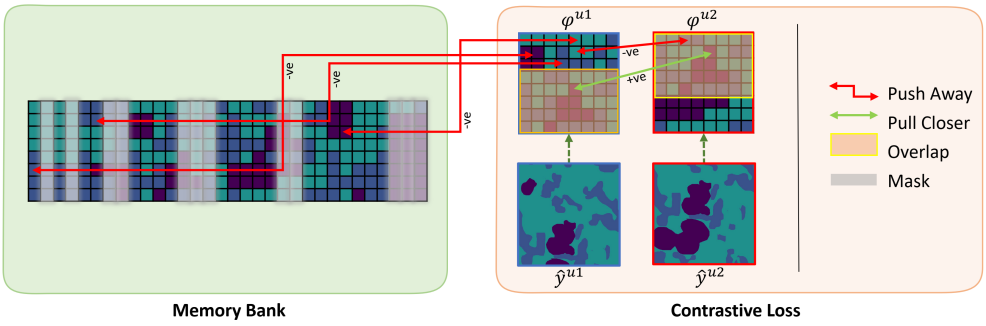


Figure 3.4: Directional contrastive loss working for context-aware consistency, where from $\varphi^{u1}, \varphi^{u2}$ overlapping area's (yellow overlay) positive pixels with higher confidence were used to pull each other closer (green arrows) while negative pixels from φ^{u2} as well as from memory bank were used to push each other apart (red arrows). To obtain these negative samples, class masks $\hat{y}^{u1}$ and $\hat{y}^{u2}$ (depicted as dashed green arrows) are applied. These masks guide the selection of negative samples from both φ^{u2} and the memory bank, which is represented by the grey overlay.

3.2.2 Methods

Figure 3.3 shows an overview of the proposed framework (CRCFP), where $L = \{(x_1^l, y_1^l), ..., (x_n^l, y_n^l) : n \in \{1, ..., N\}\}$ represents the N labelled images while $U = \{(x_1^u), ..., (x_m^u) : m \in \{1, ..., M\}\}$ represents the M unlabelled images. Labelled and unlabelled images x^l and x^u were sampled from L and U respectively in batches. Both images x^l, x^u are of $H \times W \times D$ spatial dimensions with corresponding pixel-wise mask $y^l = \mathbb{R}^{C \times H \times W}$ only for labelled image where C is the number of classes. Each labelled image x^l is passed through the supervised pathway of the CRCFP framework (blue arrows in Figure 3.3) whereas the unlabelled images x^u pass through the unsupervised pathways of the framework (brown arrows in Figure 3.3) along with two overlapping patches extracted randomly from x^u denoted as x^{u1}, x^{u2} (green arrows in Figure 3.3). Feature maps are extracted from the input image using the shared encoder $h(\cdot; \theta^h)$ and decoder $g(\cdot; \theta^g)$ as $f(\cdot; \theta^f) = h(\cdot; \theta^h) \circ g(\cdot; \theta^g)$ resulting in feature maps for each input as $f^l = f(x^l; \theta^f)$, $f^u = f(x^u; \theta^f)$, $f^{u1} = f(x^{u1}; \theta^f)$ and $f^{u2} = f(x^{u2}; \theta^f)$. Further, f^l and f^u are processed by a pixel classifier C^f for final prediction as $\hat{y}^l = C^f(f^l; \theta^p)$ and $\hat{y}^u = C^f(f^u; \theta^p)$ where $\hat{y}^l$ is optimised using the cross-entropy loss over y^l as $\mathcal{L}^{sup}$ shown in Equation 3.1.

$$\mathcal{L}^{sup} = -\frac{1}{N} \sum_{c=1}^{C} \sum_{i=1}^{N} y_{c,i}^l \log(\hat{y}_{c,i}^l) \tag{3.1}$$

where $\hat{y}_{c,i}^l$ denotes predicted label map of class c of the i^{th} instance. Similarly, $y_{c,i}^l$ denotes label map of class c of the i^{th} instance

Context-Aware Consistency

With only the supervised loss $\mathcal{L}^{sup}$, the model may start relying excessively on contexts due to limited labelled data. Context-aware consistency can alleviate this issue by aligning the two different contexts of the same patch with the help of contrastive learning. For this purpose, encoded feature maps f^{u1} and f^{u2} are projected to a low-dimensional space using a non-linear projector ϕ to preserve important contextual information. The choice of non-linear projection head as compared to linear and identity projection head is due to its superior performance [152]. The projection head $\phi(\cdot;\theta^z)$ outputs projection maps as $\varphi^{u1} = \phi(f^{u1};\theta^z)$ and $\varphi^{u2} = \phi(f^{u2};\theta^z)$. Similar to [133], context-aware consistency is maintained between the overlapping regions of φ^{u1} and φ^{u2} using the directional contrastive loss $\mathcal{L}^{cont}$ to keep the feature representation consistent under different contexts as shown in Figure 3.4. For computing directional consistency loss, first class maps $\hat{y}^{ui}$ were extracted using pixel classifier C^f and then maximum probability among all classes C is maintained using max probability as it is linked with higher confidence as shown in Equation 3.2.

$$\hat{y}^{ui} = \underset{c \in C}{\arg\max}\, C^f(f^{ui};\theta^p) \tag{3.2}$$

where $i \in \{1,2\}$ as there are two cropped patches and higher probability features are used to align less confident features towards the more confident features [68, 133, 178] which can improve learning by avoiding the exchange of unreliable knowledge from the less confident features. In order to extract positive and negative samples, class prediction maps as $\hat{y}^{u1} = C^f(f^{u1};\theta^p)$ and $\hat{y}^{u2} = C^f(f^{u2};\theta^p)$ were used as pseudo labels along with their probabilities. For computing the directional loss between $\varphi^{u1} \rightarrow \varphi^{u2}$, φ^{u1} acts as a positive (+ve) instance where a positive feature projection φ^{u1+} was computed from φ^{u1} with condition on the pseudo label $\hat{y}^{u1}$ as $\hat{y}^{u+}$ having probability greater than $\hat{y}^{u2}$ as $\hat{y}^{u-}$. Threshold λ is further applied on the φ^{u+} as this positive feature filtration enables us to avoid the exchange of less confident features while computing the loss. For negative (-ve) samples η, all other instance meeting the criteria of $(\hat{y}^{u+} \neq \hat{y}^{u-})$ were selected, where all other samples were treated as $\hat{y}^{u-}$ as shown in Equation 3.5. The $\mathcal{L}^{cont(\varphi^{u1},\varphi^{u2})}$ loss for one pair is calculated as shown below,

$$\mathcal{L}^{cont(\varphi^{u1},\varphi^{u2})} = -\frac{1}{M}\sum_{H,W} \mathcal{M}^{+}.\log \frac{\text{sim}(\varphi^{u1},\varphi^{u2})}{\text{sim}(\varphi^{u1},\varphi^{u2}) + \sum\limits_{\varphi^{u-}\in\eta} \mathcal{M}^{-}\cdot \text{sim}(\varphi^{u1},\varphi^{u-})} \tag{3.3}$$

$$\text{sim}(\varphi^{u1},\varphi^{u2}) = \exp\left(\frac{(\varphi^{u1})^{\mathrm{T}}\varphi^{u2}}{\|\varphi^{u2}\|\,\|\varphi^{u2}\|\,\tau}\right) \tag{3.4}$$

$$\mathcal{M}^{-} = \begin{cases} 1 & \text{if } \hat{y}^{u+} \neq \hat{y}^{u-}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.5}$$

$$\mathcal{M}^{+} = \begin{cases} \mathcal{M}^{c+}, & \text{if } \max C^{f}(f^{u1}; \theta^{p}) > \lambda, \\ 0, & \text{otherwise.} \end{cases} \tag{3.6}$$

$$\mathcal{M}^{c+} = \begin{cases} 1 & \max C^{f}(f^{u1}; \theta^{p}) < \max C^{f}(f^{u2}; \theta^{p}), \\ 0 & \text{otherwise.} \end{cases} \tag{3.7}$$

where sim(.) is the cosine similarity measure with temperature τ, $\mathcal{M}^{c+}$ represents the binary mask for extracting confident features corresponding to φ^{u1+}. $\mathcal{M}^{+}$ is the binary mask for positive confident samples above threshold λ. $\mathcal{M}^{-}$ is the binary mask for negative samples indicating different pseudo labels between φ^{u+} and φ^{u-}. To increase the negative samples, we have used the memory bank which stores features from recent batches to further increase the negative samples for better contrastive performance [68, 133, 152]. Similarly, for computing the directional loss between $\varphi^{u2} \rightarrow \varphi^{u1}$, φ^{u2} acts as a positive (+ve) instance where a positive feature projection φ^{u+} was computed from φ^{u2} with condition on the pseudo label $\hat{y}^{u2}$ as $\hat{y}^{u+}$ having probability greater than $\hat{y}^{u1}$ as $\hat{y}^{u-}$. Finally, the directional contrastive loss $\mathcal{L}^{t-cont}$ is calculated as below:

$$\mathcal{L}^{t-cont} = \mathcal{L}^{cont}(\varphi^{u1}, \varphi^{u2}) + \mathcal{L}^{cont}(\varphi^{u2}, \varphi^{u1}) \tag{3.8}$$

Cross-Consistency Training

As context-aware consistency improves the model's robustness towards changing contexts without losing object-awareness, the model is still susceptible to small perturbations in the input due to limited labelled data. Therefore, in order to leverage unlabelled data and make the model invariant to small perturbations, we utilise the cross-consistency training [132] where f^{u} is perturbed K times for each perturbation type and consistency is maintained between the output of pixel classifier and auxiliary classifiers. This not only improves the model's robustness but also regularises the main pixel classifier towards correct predictions. We use $\hat{y}^{u}$ to regularise the pixel classifier over the mean square error (MSE) loss by measuring the distance between the output of the main pixel classifier C^{f} and the output of auxiliary classifiers C^{f}_{k}. Formally, a perturbation function p_k with $k \in \{1, K\}$ perturbations outputs a perturbed version of the f^{u} as $f^{u'}_{k} = p_k(f^{u})$ for a perturbation type and the cross-consistency training loss $\mathcal{L}^{cross}$ can be defined as below,

$$\mathcal{L}^{cross} = \frac{1}{M}\frac{1}{K}\sum_{x^u \in U}\sum_{k=1}^{K} d(\hat{y}^u, C_k^f(f_k^{u'})) \tag{3.9}$$

where d measures the squared distance between the output probabilities of the main pixel classifier and perturbed pixel classifier output. The following perturbations are applied to enforce consistency:

Feature Noise: From uniform distribution $\mathcal{U}$, a uniformly sampled noise in the interval $[\alpha, \beta]$ is added to the features map f^u in two steps. First sampled noise is multiplied with f^u to scale the noise relative to feature activations. Second, the scaled sample noise is then added to the feature map f^u. This makes the noise proportional to each feature activation, as shown below.

$$\Omega \sim \mathcal{U}(\alpha, \beta) \tag{3.10}$$

$$f^{noise} = (f^u \odot \Omega) + f^u \tag{3.11}$$

where Ω is random noise sampled from $\mathcal{U}$ and $\odot$ represents the element wise multiplication.

Feature Dropout: From uniform distribution $\mathcal{U}$, a uniform sample threshold γ is used to prune the less confident activations to stop the model from relying on those activations. This is done by first summing the f^u over different channels and then normalising it using min-max normalisation resulting in $f^{u'}$. Anything below γ is dropped, as seen below:

$$\gamma \sim \mathcal{U}(\alpha, \beta), \tag{3.12}$$

$$\mathcal{M}^{drop} = \begin{cases} 1, & \text{if } f^{u'} < \gamma, \\ 0, & \text{otherwise.} \end{cases}, \tag{3.13}$$

$$f^{drop} = \mathcal{M}^{drop} \odot f^{u'}. \tag{3.14}$$

where $\mathcal{M}_{drop}$ is the binary mask containing threshold values for pruning the activations.

DropOut: A fraction of activations are dropped out spatially, where the fraction is decided using the Bernoulli distribution with probability δ.

$$\mathcal{M}^{dropout} \sim \text{Bernoulli}(\delta) \tag{3.15}$$

$$f^{dropout} = \mathcal{M}^{dropout} \odot f^u \tag{3.16}$$

Entropy Minimisation

Context-aware contrastive learning and cross-consistency training improves the encoder's features but it often fails to improve the final pixel classifier leading to less reliable pseudo labels corrupting the training from unlabelled data. As higher confidence means better prediction maps resulting in more refined pseudo labels which can help train both context-ware and cross-training with improved positive/negative pairs and pseudo labels. Hence, in order to improve the confidence of predictions, we employ entropy regularisation following its applications in semi-supervised learning [68, 89, 168, 182] as shown in Equation 3.17 where it penalises the uncertain prediction in the unlabelled data and improves the overall confidence of the prediction maps.

$$\mathcal{L}^{ent} = -\frac{1}{C}\sum_{c=1}^{C}\sum_{m=1}^{M}\hat{y}^{u}\log\hat{y}^{u} \tag{3.17}$$

3.2.3 Training

Finally, the entire framework is trained in an end-to-end fashion using a weighted combination of the above mentioned losses, as shown below,

$$\mathcal{L} = w^{sup}\mathcal{L}^{sup} + w^{t-cont}\mathcal{L}^{t-cont} + w^{cross}\mathcal{L}^{cross} + w^{ent}\mathcal{L}^{ent} \tag{3.18}$$

where w^{sup}, w^{t-cont}, w^{cross} and w^{ent} correspond to the weights for each loss component respectively.

3.2.4 Implementation Details

Network Architecture

We used DeepLab-v3 [114] as the base segmentation network with ResNet-50 [135] encoder pretrained on ImageNet [183] where the projector consists of two fully connected (FC) layers of size 128 with ReLU as an intermediate activation layer, FC $\rightarrow$ ReLU $\rightarrow$ FC. Pixel classifiers consist of convolutional layers with a kernel of size 1 $\times$ 1 to reduce the number of channels to total classes with non-linear ReLU activation. The final layers upsample the output using bi-linear interpolation to match the input size as $H \times W \times C$.

3.2.5 Experimental Settings

In order to address the potential impact of centre selection on performance, particularly when dealing with smaller data fractions, we randomly selected the training centres for each trial using different random seeds. However, to

ensure fair evaluation and mitigate any bias, we have conducted testing on fixed and unseen test sets derived from the two challenge contests. In each trial, we have performed a three fold cross-validation to enhance the reliability and robustness of our results. This cross-validation approach allows us to assess the performance of our method across multiple iterations, ensuring that the evaluation is not biased by the specific partitioning of the data. This approach allowed us to account for the influence of centre selection while maintaining consistent and unbiased evaluation across our experiments. The input size for the proposed framework for both labelled and unlabelled images was 320 $\times$ 320. For contrastive learning, two patches x^{u1} and x^{u2} were randomly cropped from the unlabelled image with an overlap in the range of [0.1, 1.0] and are resized to match the input dimensions. For positive filtering mask λ was set to 0.75 by empirical evaluation and $\tau = 0.1$ as temperature for cosine similarity. For cross-consistency training, number of auxiliary pixel classifiers were set to $K = 4$ for each perturbation type and for feature noise perturbation the parameters $\alpha = -0.3$, $\beta = 0.3$ were used. For feature dropout perturbation, $\alpha = 0.75$, $\beta = 0.9$ were used as they can help remove approximately 10% to 30% of active regions from the feature map. Also, for simple Dropout the probability for Bernoulli distribution was set to $\delta = 0.5$. During training, a set of standard augmentations were applied to the input images including horizontal and vertical flipping, gaussian blur, colour and grey scaling. PyTorch was used for implementing this framework where for optimisation we train the whole framework for 80 epochs. For the initial 5 epochs, only supervised loss $\mathcal{L}^{sup}$ was used to train the whole framework as this provides a stable head start for the semi-supervised learning. The batch size of 8 was used for labelled and unlabelled images with stochastic gradient descent (SGD) optimiser and a learning rate of 0.001. As a common practice, *poly* learning rate decay policy was used where the learning rate is scaled using $1 - (\frac{iter}{\max\_iter})^{power}$ at each iteration with $power = 0.9$. Weights with respect to different losses $\mathcal{L}^{sup}$, $\mathcal{L}^{t-cont}$, $\mathcal{L}^{cross}$ and $\mathcal{L}^{end}$ were set to fixed values as $w^{sup} = 1$, $w^{t-cont} = 0.1$, $w^{cross} = 0.01$ and $w^{ent} = 0.01$ respectively after empirical evaluation. All models were trained with the same configurations for both datasets where two Nvidia GeForce 1080Ti GPUs are used for training.

3.2.6 Evaluation metrics

In order to compare our method quantitatively with other state-of-the-art methods (SOTA), we have used different pixel-wise quantitative measures, including accuracy, F1-score (Dice) and mean intersection over union (mIoU) for both datasets.

Table 3.2: Comparison of the state-of-the-art methods with mIoU, dice score and accuracy aggregated for 3 different random seeds as mean (standard deviation). The first column represents the fraction of data used for training the model.

BCSS				
Fraction	Method	mIoU	Dice	Accuracy
1/8	DeepLab-v3 [114]	40.99 (7.96)	55.96 (9.1)	66.53 (6.07)
1/8	CCT [132]	22.84 (0.54)	32.01 (0.69)	56.14 (0.70)
1/8	CAC [133]	44.67 (6.32)	58.97 (7.51)	72.43 (3.40)
1/8	CRCFP	**47.09 (6.18)**	**61.84 (6.59)**	**73.20 (3.31)**
1/4	DeepLab-v3 [114]	53.03 (0.88)	68.52 (0.94)	75.70 (0.65)
1/4	CCT [132]	30.63 (2.19)	43.24 (3.33)	62.43 (0.89)
1/4	CAC [133]	58.65 (0.65)	73.33 (0.55)	78.60 (0.42)
1/4	CRCFP	**61.06 (0.98)**	**75.21 (0.74)**	**80.87 (0.89)**
1/2	DeepLab-v3 [114]	56.26 (1.19)	71.33 (0.98)	78.07 (1.031)
1/2	CCT [132]	29.78 (2.56)	41.63 (4.28)	61.94 (0.17)
1/2	CAC [133]	60.44 (1.48)	74.67 (1.16)	80.50 (0.92)
1/2	CRCFP	**61.86 (0.63)**	**75.73 (0.59)**	**81.18 (0.27)**
1/1	DeepLab-v3 [114]	61.29 (0.26)	75.49 (0.12)	81.10 (0.09)

3.3 Results and Discussion

The performance of our method CRCFP compared to recent SOTA semi-supervised semantic segmentation methods, including DeepLab [114], CCT [132], CAC [133] and CDCL [68] [1] is shown in Table 3.2 and Table 3.3. As these methods were implemented originally using different configurations and baseline segmentation models. We have implemented these methods within a unified framework with the same segmentation baseline, experimental settings and data augmentations for a fair comparisons.

Table 3.2 shows the performance of our CRCFP model compared to supervised and semi-supervised methods on all matrices for the BCSS dataset. Particularly, when 1/8 proportion of the training centres was used, it can be seen that in terms of mIoU our method performs ∼6% better than the supervised method and ∼3% better than the recent CAC [133]. Similarly, its worth noting that with 1/4 of the total centres, the CRCFP performance is almost similar to fully supervised method with all data. however, it can also be seen that training the model on a single centre (i.e., 1/8) results in a model with high standard deviation across the folds, this shows that training from a single source is highly susceptible to stain variations. On the other hand, the poor performance of CCT [132] can be attributed towards heavy perturbations applied directly to the features where it brings perturbed features from different contexts closer without pushing dissimilar apart whereas

[1] CDCL [68] cannot be applied to the multi-class problem as it divides the patches into foreground and background only for contrastive learning.

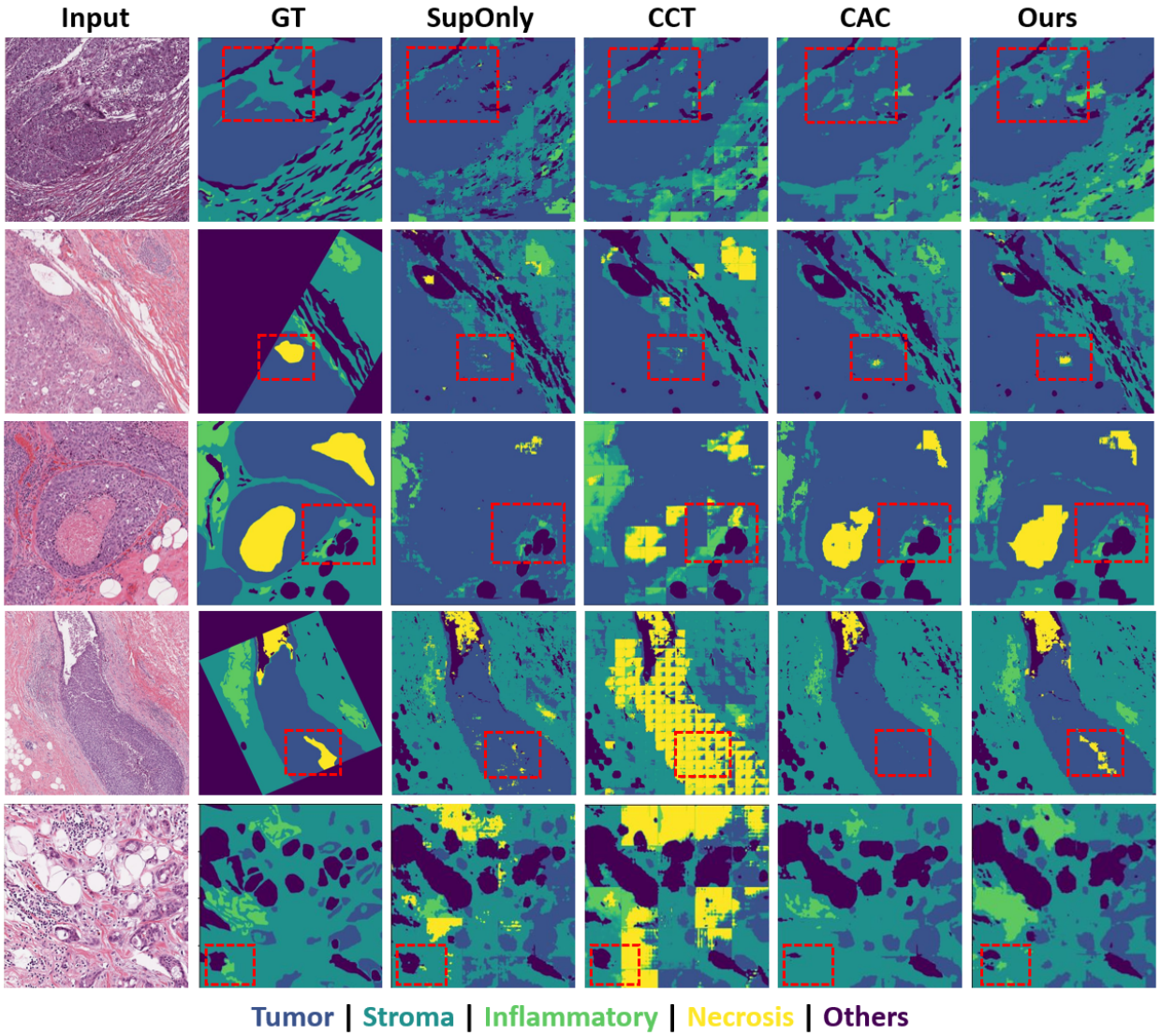


Figure 3.5: Visual comparison of the CRCFP with different state-of-the-art methods for tissue region segmentation with 1/2 training data only. The dashed red box highlights the superior performance of our method as compared to SOTA methods.

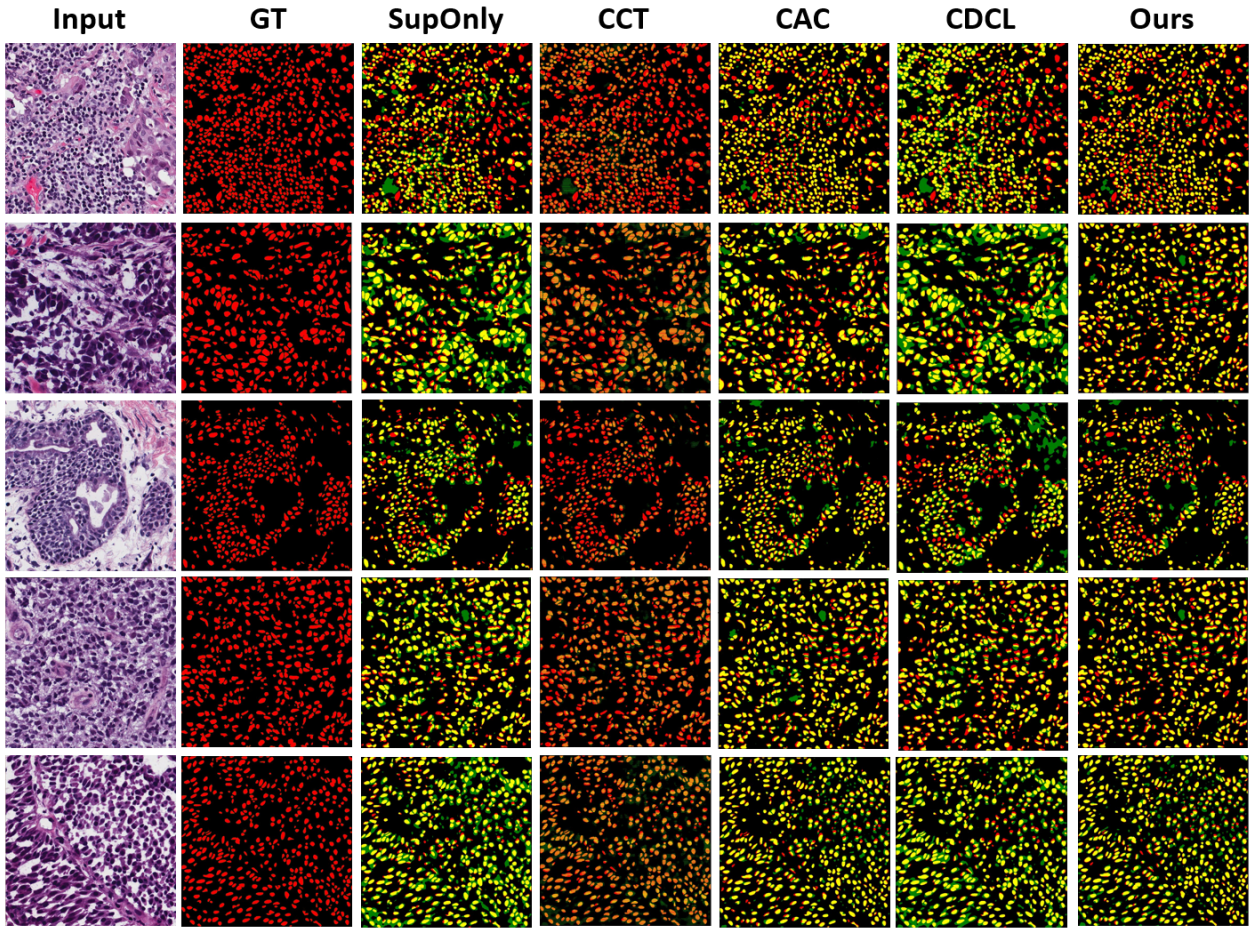


Figure 3.6: Visual comparison of the CRCFP with different state-of-the-art techniques in nuclei image segmentation with 1/8 training data only. GT represents the ground truth nuclei masks, and SupOnly shows the models trained with labelled training data only. Red pixels correspond to the ground truth, while green shows the prediction. Yellow pixels represent the overlap regions between the prediction and ground truth.

Table 3.3: Comparison of the state-of-the-art methods with mIoU, dice score and accuracy aggregated for 3 different random seeds as mean (standard deviation). The first column represents the fraction of data used for training the model.

MoNuSeg				
Fraction	Method	mIoU	Dice	Accuracy
1/32	DeepLab-v3 [114]	60.09 (2.07)	73.89 (1.77)	79.45 (1.40)
1/32	CCT [132]	41.13 (0.06)	50.31 (0.29)	74.33 (0.35)
1/32	CAC [133]	67.40 (1.12)	79.33 (0.90)	86.14 (0.62)
1/32	CDCL [68]	62.72 (1.83)	75.95 (1.35)	81.66 (1.57)
1/32	CRCFP	**71.72 (0.22)**	**82.60 (0.24)**	**88.86 (0.23)**
1/16	DeepLab-v3 [114]	56.20 (5.76)	70.80 (4.58)	75.27 (75.27)
1/16	CCT [132]	40.99 (0.08)	49.56 (0.29)	75.17 (0.36)
1/16	CAC [133]	71.44 (1.11)	82.47 (0.92)	88.27 (0.11)
1/16	CDCL [68]	60.63 (1.15)	74.40 (0.90)	79.99 (1.37)
1/16	CRCFP	**72.08 (2.07)**	**82.91 (1.52)**	**88.58 (1.16)**
1/8	DeepLab-v3 [114]	59.67 (2.99)	73.59 (2.32)	78.98 (2.89)
1/8	CCT [132]	40.9 (0.13)	50.00 (0.54)	74.63 (0.42)
1/8	CAC [133]	74.56 (0.42)	84.73 (0.30)	89.91 (0.23)
1/8	CDCL [68]	57.07 (1.45)	71.63 (1.14)	76.29 (1.48)
1/8	CRCFP	**75.57 (0.85)**	**85.19 (0.54)**	**90.28 (0.42)**
1/1	DeepLab-v3 [114]	71.29 (0.16)	82.49 (0.11)	87.52 (0.09)

CAC [133] not only bring them closes but also pushes away the features from different classes. However, it focuses more on encoder feature generalisation leaving pixel-classifier with less confident features. Figure 3.5 shows visual comparison of CRCFP with the SOTA algorithms, where, it can be observed that prediction maps of CRCFP are better as compared to the rest, specially highlighted in the dashed red boxes. There are also some noticeable stitching artefacts in some prediction maps, especially in the case of CCT. Stitching artefacts are common in patch-based segmentation methods, especially when using large ROI, i.e., the case in BCSS. These artefacts can be removed with the help of overlapping patch processing, but that takes additional computing time.

Table 3.3 shows the performance of CRCFP surpassing other SOTA methods in all data proportions and metrics, especially in 1/32 proportion of the MoNuSeg dataset. It can be seen that our CRCFP outperforms the CAC [133] by 4.32% in mIoU with a smaller standard deviation of 0.22. It can also be observed that fully supervised models are more susceptible to domain generalisation problems from the table as in 1/32 proportion of the training images, the performance of DeepLab-v3 [114] is 4% better than the 1/16 proportion of the training images whereas there is more data available in the latter. This is due to the fact that in a random sampling of training images, some training images are better indicators of the testing distribution due to

similarities in the same stain, organ and tumour stage. However, most of the SOTA semi-supervised algorithms solve this issue with the help of unlabelled data, as it can be seen that the performance increase with the increase in data for all these methods. Figure 3.6 shows a visual comparison of CRCFP with SOTA methods where it can be seen that our approach predicts fewer false positives as compared to CDCL [68].

Further, in order to validate the contribution of each component (i.e., context-aware consistency, cross-consistency training and entropy minimisation), we conducted an extensive ablation study. The ablation study is performed on the BCSS dataset due to its complexity and multi-class nature, where we studied the effect of using all data proportions for the different encoders and stripping the framework. While studying the effect of negative samples and the number of auxiliary pixel classifiers, we used 1/8 data proportion.

3.3.1 Encoder

To verify the performance boost by plugging in a bigger encoder in the base segmentation network, we replaced ResNet-50 with ResNet-101 for all data proportions. Table 3.4 shows the performance of the CRCFP framework with a bigger encoder, and it can be seen that there is a performance boost overall for most of the methods, especially for CCT [132]. However, it can be observed that CRCFP with a smaller encoder (i.e., ResNet-50) still performs comparable/better than other SOTA techniques with a bigger encoder, e.g., in 1/8 proportion CAC [133] with ResNet-101 achieves mIoU of 46.91 where CRCFP with ResNet-50 achieves mIoU of 47.09 showing the superiority of our method. Also, it is worth mentioning that with ResNet-101, the standard deviation we observed with ResNet-50 was reduced, owing to the fact that bigger encoders are more stable for semi-supervised learning frameworks. Overall the CRCFP framework provides improved and stable performance with bigger encoders as compared to the other methods.

3.3.2 Network Schemes

We validated the contribution of each component by breaking down the whole framework with respect to different losses and called them network schemes. We started with a baseline segmentation network, i.e., DeepLab-v3 with ResNet-50 as SupOnly, Scheme.1 consists of using context-aware consistency loss, Scheme.2 consists of using context-aware consistency loss with entropy minimisation, and finally, Scheme.3 is our framework with context-aware consistency loss with cross-consistency training and entropy minimisation. Table 3.5 shows the schemes with respect to their respective losses being used. It can be seen that with each component's addition, we can see an improvement in

Table 3.4: Comparison of the state-of-the-art methods on the mean (standard deviation) of the mean intersection of union (mIoU), dice score and accuracy with baseline encoder as ResNet-101. The first column represents the fraction of data used for training the model.

BCSS				
Fraction	Method	mIoU	Dice	Accuracy
1/8	DeepLab-v3 [114]	37.50 (6.61)	51.73 (7.51)	64.89 (5.92)
1/8	CCT [132]	31.71 (4.64)	45.66 (5.96)	59.42 (3.46)
1/8	CAC [133]	46.91 (6.79)	61.92 (6.74)	72.01 (3.85)
1/8	CRCFP	**47.15 (6.76)**	**61.27 (7.72)**	**72.57 (2.82)**
1/4	DeepLab-v3 [114]	55.18 (1.88)	70.30 (1.70)	77.37 (1.25)
1/4	CCT [132]	42.63 (0.98)	58.35 (1.26)	66.94 (0.73)
1/4	CAC [133]	61.48 (0.73)	75.52 (0.47)	80.78 (0.84)
1/4	CRCFP	**62.01 (0.40)**	**75.94 (0.29)**	**81.18 (0.49)**
1/2	DeepLab-v3 [114]	60.37 (1.89)	74.5 (1.58)	80.57 (0.86)
1/2	CCT [132]	44.01 (0.65)	59.64 (0.55)	67.66 (1.33)
1/2	CAC [133]	61.95 (0.72)	75.77 (0.67)	81.09 (0.27)
1/2	CRCFP	**63.01 (0.09)**	**76.57 (0.09)**	**81.67 (0.12)**
1/1	DeepLab-v3 [114]	62.33 (1.04)	76.22 (0.73)	81.68 (0.58)

overall performance. E.g., in 1/8 data proportion, the addition of context-aware consistency brings about 4% of improvement while entropy minimisation further bumps it up by 1%, and finally, cross-consistent training beings about 2% of improvement, accumulating the overall performance to ∼7% from baseline supervised model. Also, for other data proportions, the performance boost is not that much significant with the addition of these Scheme.2 and Scheme.3 as compared to Scheme.1. However, its worth mentioning that the standard deviation of Scheme.2 and Scheme.3 as compared to Scheme.1 is smaller which is due to the fact that these schemes brings confidence in prediction maps thus improving the overall performance with stability.

3.3.3 Negative Samples

As increasing the negative samples in training contrastive learning framework boosts the performance of the underlying model. This is done mostly by increasing the batch size to 2048 or 4096 where possible, as the bigger the batch size, the more samples you get for comparisons [152, 184]. However, where it is not possible, another workaround is to use a memory bank where negative samples from previous batches are stored for later use. Therefore, in order to get the upper bound of performance in our framework with respect to negative samples, we have experimented with the different numbers of negative samples as seen in Table 3.6. It can be noticed that with increasing negative samples, the performance increases for a while, and then it reaches the plateau and then increases with very little gain, as it can also be observed visually

Table 3.5: CRCFP breakdown with BCSS splits in different Schemes with respect to their loss functions. SupOnly correspond to baseline segmentation model with $\mathcal{L}^{sup}$ loss only. Scheme.1 corresponds to addition of $\mathcal{L}^{t-cont}$ loss on top of SupOnly. Scheme.2 corresponds to addition of $\mathcal{L}^{ent}$ on top of Scheme.1 and finally Scheme.3 is addition of $\mathcal{L}^{cross}$ on top of Scheme.2.

						BCSS
Method	Split	$\mathcal{L}^{sup}$	$\mathcal{L}^{t-cont}$	$\mathcal{L}^{ent}$	$\mathcal{L}^{cross}$	mIoU
SupOnly	1/8	✓	×	×	×	40.99 (7.96)
Scheme.1	1/8	✓	✓	×	×	44.67 (6.32)
Scheme.2	1/8	✓	✓	✓	×	45.76 (6.12)
Scheme.3	1/8	✓	✓	✓	✓	**47.09 (6.18)**
SupOnly	1/4	✓	×	×	×	53.03 (0.88)
Scheme.1	1/4	✓	✓	×	×	58.65 (0.65)
Scheme.2	1/4	✓	✓	✓	×	59.97 (1.47)
Scheme.3	1/4	✓	✓	✓	✓	**61.06 (0.98)**
SupOnly	1/2	✓	×	×	×	56.26 (1.19)
Scheme.1	1/2	✓	✓	×	×	60.44 (1.48)
Scheme.2	1/2	✓	✓	✓	×	60.87 (1.39)
Scheme.3	1/2	✓	✓	✓	✓	**61.86 (0.63)**

Table 3.6: Performance of CRCFP with respect different number of negatives samples used while training $\mathcal{L}^{t-cont}$ loss with BCSS data split of 1/8

	BCSS		
#	mIoU	Dice	Accuracy
100	45.62 (8.10)	60.46 (8.87)	67.38 (8.14)
500	45.81 (7.78)	59.86 (8.88)	70.05 (5.30)
1200	47.09 (6.18)	61.84 (6.59)	73.20 (3.31)
1600	47.16 (6.70)	61.81 (7.05)	72.68 (3.07)
2400	47.60 (6.09)	62.14 (6.80)	**73.58 (3.43)**
3200	**48.34 (5.25)**	**63.83 (5.01)**	73.06 (3.59)

in Figure 3.7. This can be due to the fact that there might not be many variations to cover in training set with more negative samples, thus reaching stable performance or very little performance gain. Also, due to gradient checkpoint functionality in PyTorch, adding more negative samples does not affect the training efficiency drastically but does consume more compute time and memory. Hence, based on these observations, for this study, we set the number of negative samples to 1200 for its memory vs accuracy trade-off.

3.3.4 Auxiliary Pixel Classifier

We employed separate auxiliary classifiers for each perturbation to ensure accurate classification of each variation introduced by perturbations. Sharing a single classifier or using the main classifier alone would limit the model's ability to differentiate between the different perturbations and accurately. Additionally,

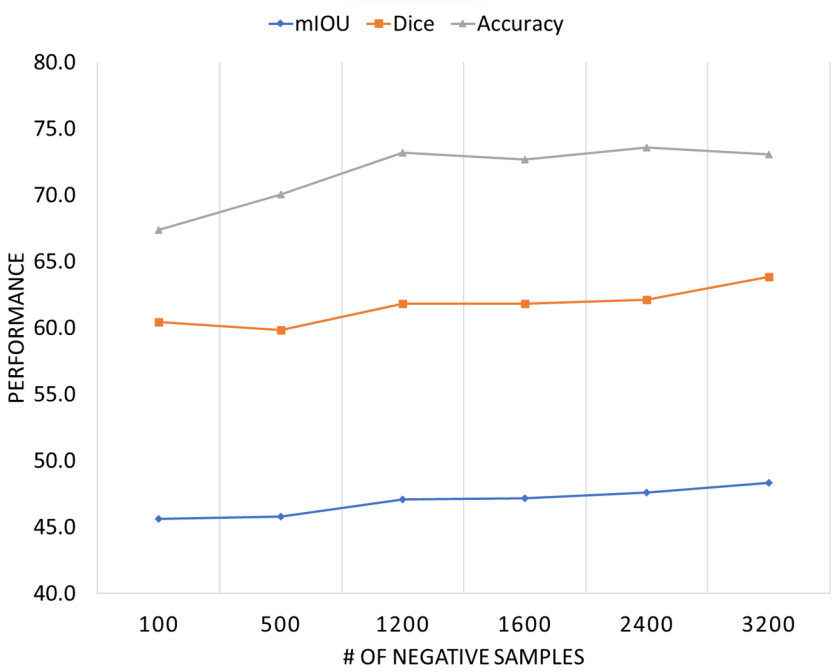


Figure 3.7: Performance graph with respect varying number of negatives samples used while training $\mathcal{L}^{t-cont}$ loss with BCSS data split of 1/8

relying solely on the main classifier would provide only one supervision signal during training, potentially insufficient for capturing the diverse variations in the input data caused by the perturbations. To see the effect of a varying number of auxiliary pixel classifiers with respect to different perturbations we conducted experiments with $K \in \{1, 2, 4, 6, 8, 10\}$ as seen in Table 3.7. It can be seen that increasing the number of pixel classifiers per perturbation increases the performance but the upper bound is achieved soon after it reaches $K = 4$, from where the performance drops slightly as can be observed in the Figure 3.8. Increasing the number of perturbations can result in more aggressive penalisation of the model overall as it accumulates to $K \times 3$ losses which can deviate the model from learning meaningful representations. Based on this observation we set the number $K = 4$ for our study for the rest of the comparisons for both datasets.

Interpretable features from histology slides can be extracted by segmenting objects/structures from ROIs, e.g., nuclei, glands, stroma, tumours etc. Interpretable features can enable the discovery of novel digital bio-markers with explanations for histology images for hard tasks like survival analysis [107, 185] and mutation prediction [54, 186, 187]. Therefore, it is vital for the downstream tasks to have good quality and precise segmentation of the region of interests. For this purpose, utilising unlabelled data for representation learning not only improves performance but also improves the internal representations for better learning. The qualitative and quantitative results, along with the ablation

Table 3.7: Performance of CRCFP with respect different number of K auxiliary classifiers used while training $\mathcal{L}^{cross}$ loss with BCSS data split of 1/8

	BCSS		
#	mIoU	Dice	Accuracy
1	43.94 (7.95)	58.9 (8.27)	69.14 (5.54)
2	45.76 (7.51)	60.44 (7.88)	71.23 (4.23)
4	**47.09 (6.18)**	**61.84 (6.59)**	**73.20 (3.31)**
6	46.48 (6.26)	61.01 (6.73)	72.60 (3.68)
8	46.72 (6.88)	61.38 (7.29)	72.25 (3.89)
10	45.68 (6.79)	60.64 (7.20)	71.84 (3.99)

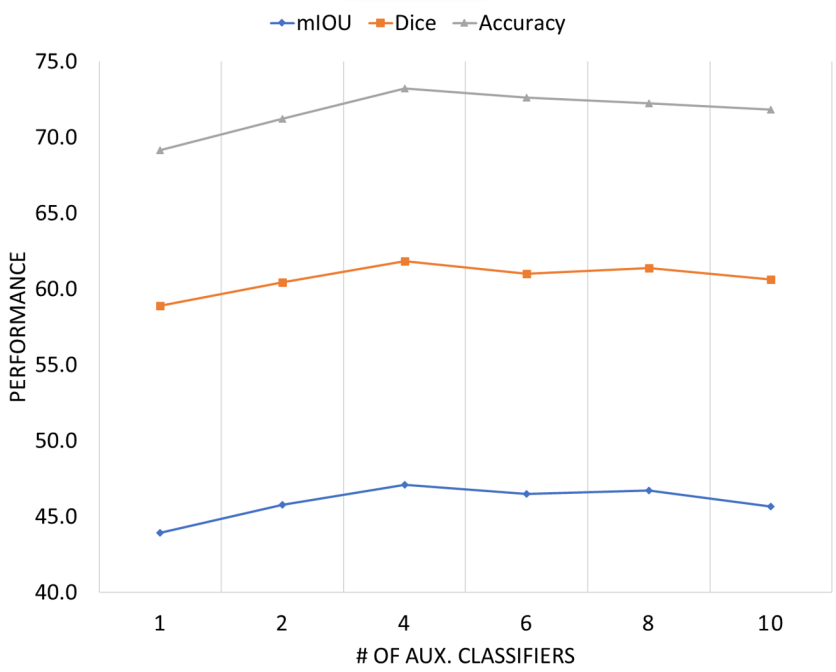


Figure 3.8: Performance graph with respect varying number of pixel classifiers used while training$\mathcal{L}^{cross}$ loss with BCSS data split of 1/8

study, have shown superior performance of CRCFP with respect to other SOTA methods. However, it's worth exploring internal representations of the learned models (i.e., feature embeddings) to account for (1) Consistency in feature space and (2) Cluster assumption for the sake of validation of aforementioned claims in the introduction section.

3.3.5 Feature Space Visualisation

In order to observe the consistency in feature space, feature embeddings were extracted from both our SSL based CRCFP trained on 1/2 proportion of the training data vs DeepLab-v3 trained on all data (i.e., fully supervised) since they achieved the same performance. Extracted feature maps were upsampled to match the size of the input image (i.e., 320 $\times$ 320) and were then mapped

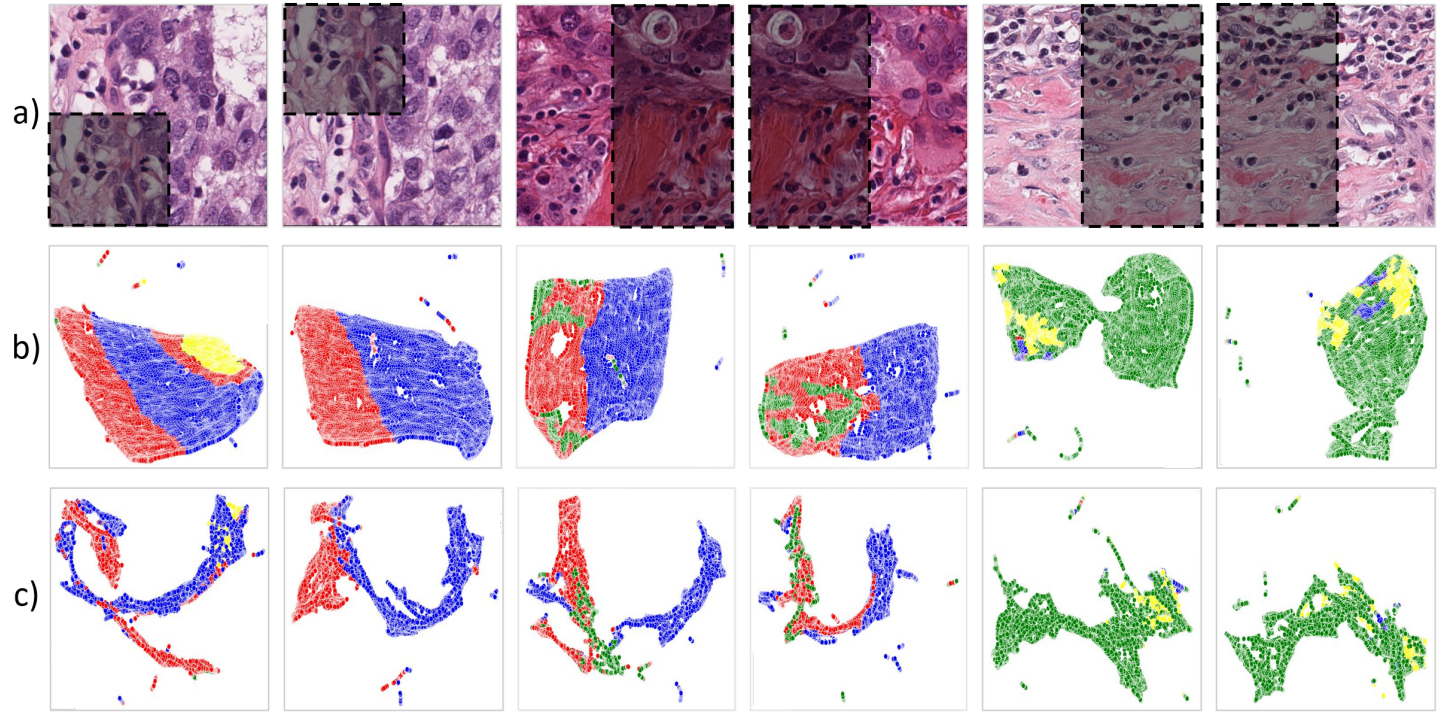


Figure 3.9: (a) BCSS dataset images pairs with overlapping regions cropped sequentially (i.e., dashed grey boxes) from the same image to mimic changing contexts. (b) UMAP visualisations of feature embedding distributions extracted from a fully supervised model. (c) UMAP visualisations of feature embedding distributions extracted from a semi-supervised model. Note that the feature embeddings are represented in the same UMAP space where dots with the same colour represents feature embedding from the same class.

to lower dimensions using UMAP [188] for visualisation purposes. It can be seen in Figure 3.9 that the feature embedding distributions are consistent with varying contexts, especially in the 1^{st} and 2^{nd} column for our CRCFP model as compared to fully supervised ones. Similarly, it can be observed in the other examples where the varying context is inherent due to the sequential overlap in the patch tessellation process. In comparison, the fully supervised model is susceptible to perturbations in contextual cues, as can be observed. It is worth noting the last two columns where the shape of feature embedding distribution changes along with the orientation of the same sample points from the same class. Especially the ones shown in yellow dots as compared to our framework, where the distributions are almost consistent under these perturbations.

3.3.6 Cluster Assumption

Consistency regularisation based methods work on the basis of cluster assumption and have achieved SOTA results in semi-supervised classification and segmentation. The main idea behind consistency regularisation is to have high and low density regions where samples closer to each other are likely to share the same label forming a high density region with a low average distance. At the same time, the class boundaries are likely to be aligned with the low density regions, i.e., high average distance. In order to observe cluster assumption, feature embeddings were extracted from CRCFP and were compared against RGB colour space as shown in Figure 3.10. Extracted feature maps were upsampled to match the size of the input image, and then the average euclidean distance between each patch of size 21×21 centred around its four immediate spatial neighbours (left, right, top and bottom) was calculated. It can be seen in Figure 3.10(d) that the class boundaries are much more aligned and apparent in the feature space as compared to the colour space where the boundaries don't align well e.g., some shown with the black arrows, thus violating cluster assumption. This can be due to the fact that the CNNs at higher layers tend to learn more semantic based features from the basic low-level features. Also, interestingly the background/fat represented in white colour in input images somewhat holds the high density regions because there is not much change in colour values for that region. In comparison, the rest of the tissue area is not very homogeneous in pixel values due to the presence of cells of various shapes and sizes.

3.4 Chapter Summary

In this chapter, I have presented a novel consistency based semi-supervised learning based semantic segmentation framework for region and nuclei seg-

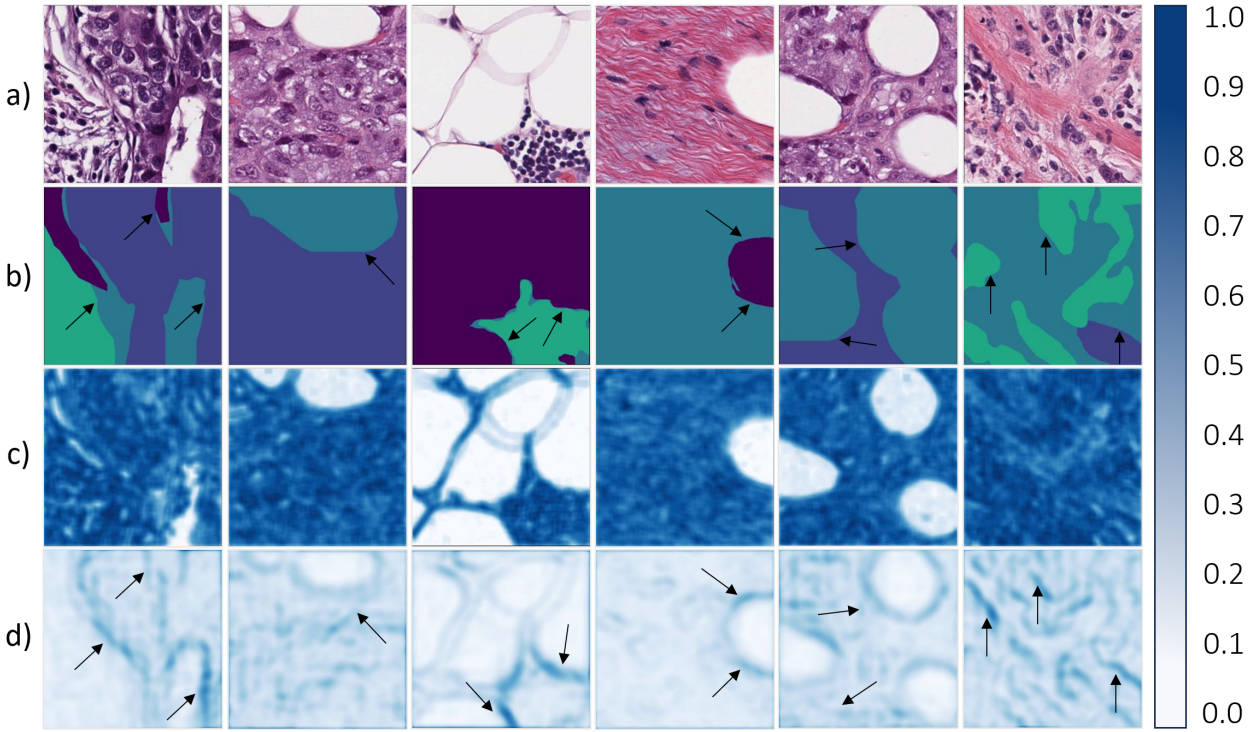


Figure 3.10: (a) Example images from BCSS test dataset. (b) Respective masks show the foreground classes and background pixels. (c,d) Average euclidean distance L^2 between the central patch of size 21×21 with four overlapping patches in the immediate neighbours in RGB colour space and feature space, respectively. Note that for feature space visualisation, encoder embeddings were upsampled to map input size. The darker blue colour represents the low density regions corresponding to the high average distance.

mentation in histology images. Our method is invariant to varying contexts and perturbations, making it efficient and robust for semantic segmentation tasks. We have shown that context-aware consistency learning can exploit unlabelled images efficiently with the help of cross-consistency training and entropy minimisation. Extensive experiments on two publicly available large histopathological datasets have shown the superiority of the CRCFP framework by achieving new SOTA results for semi-supervised semantic segmentation. Also, detailed ablation studies for different network parameters and components show the contribution of each network component, demonstrating the effectiveness of our method. Future directions include improvements to the presented method with respect to improving the context-aware loss for minor classes and finding histology specific perturbation, e.g., targeting stain variations, on a large multi-centric histopathological dataset. Large multi-centric data is vital for the validation of the study as the quality of downstream analysis is highly dependent on the segmented histology primitives.

Chapter 4

Weakly Supervised Learning for Predicting Malignancy in Oral Epithelial Dysplasia (OED)

4.1 Introduction

Oral cancer is amongst the most common cancers in the world and is considered a major health problem due to the its significant associated morbidity and mortality [189]. The 5-year survival rate has not improved over the last few decades, regardless of improvements in surgical and oncological treatments. A large majority of oral cancers (>90%) are oral squamous cell carcinoma (OSCC), with one of the biggest obstacles to improvement in prognosis being delayed presentation of disease, as evidenced by the fact that survival for stage-I OSCC is 80% which reduces to 20-30% for stage IV disease [190, 191]. OSCC is caused by a multitude of genetic and environmental factors and is preceded in a majority of cases by a potentially malignant state with the proliferation of atypical epithelium known as oral epithelial dysplasia (OED) [10]. Dysplastic lesions have been shown to have an increased risk of progression to malignant transformation [192]. Currently, there are no specific clinical tools or biological or molecular markers routinely used or recommended in clinical practice for the prognostication of dysplastic lesions. Some clinical risk predictors have been suggested to be helpful including size, clinical site (e.g., the floor of the mouth, lower gums, lateral tongue), and clinical appearance (i.e. leukoplakia, erythroplakia etc.) and can be found in a wide range of conditions collectively referred to as oral potentially malignant disorders (OPMDs) in clinical practice [193].

With the wider adaptation of digital pathology in clinical practice, AI algorithms have also evolved and have shown promise for automated detection and quantification of histological features for classification [41, 47, 50, 93, 194], detection [41, 42, 101, 195], segmentation [142, 145, 196] and survival analysis [93, 107]. Digitisation of histology slides along with AI can be used to develop algorithms to assist pathologists in diagnostic decision-making and better prognostication for improved patient management. To the best of our knowledge, there has been limited research on computational analysis of OED histology images for the prediction of malignant transformation. Existing methods in literature have used relatively small cohorts, manual elements, or region of interest (ROI) based analyses [50, 197–202]. All these methods have focused mainly on OED identification or grading and lack predictive or prognostic ability. Limited computational pathology work has been reported at the WSI level for predictive analysis of OED, including recurrence and malignant transformation potential. Dost et al. [199] examined 368 OED patients where 7.1% progressed to carcinoma and showed that there was no association of OED grade with malignant transformation. Gilvetti et al. [203] reported a study including 120 patients with a mean follow-up of 47.7 months ($\pm$29.9 SD) and showed that the recurrence rate was significant in high grade OED patients with erythroplakia with $p = 0.023$ with the mean time to recurrence of 62 months ($\pm$31.5 SD). Malignant transformation was also shown to have a significant association with age ($p = 0.034$), clinical appearance ($p = 0.030$), lesion site ($p = 0.007$) and some other clinical features with a mean transformation time of 50 months ($\pm$32.5 SD). A recent study by Mahmood et al. [204] examined the correlation between individual histological features and OED prognosis. They examined OED biopsies from 108 patients with a minimum of five-year follow-up to analyse histological features predictive of recurrence and malignant transformation. Two different prognostic models based on the presence of specific histological features (bulbous rete processes, hyperchromatism, loss of epithelial cohesion, loss of stratification, suprabasal mitoses and nuclear pleomorphism- irrespective of grade) were proposed with an area under the receiver-operator characteristic curve (AUROC) value of 0.77 for malignant transformation and 0.72 for recurrence. This highlights the usefulness of individual (grade-independent) histological features for OED prognosis prediction. A significant proportion of OED lesions can transform into malignancy (OSCC), and at present, there are no tools available for objective and reproducible prediction of malignant transformation. Early prediction of malignant transformation is crucial to aid patient care and inform appropriate treatment to improve prognosis and reduce the need for radical and disfiguring surgery later. In this chapter, we investigate the effectiveness of deep learning algorithms for prognostication from Haematoxylin & Eosin

(H&E) stained WSIs in an end-to-end manner.

4.2 Materials and Methods

4.2.1 Data

The dataset used for this study comprised 163 Haematoxylin and Eosin (H&E) stained and scanned whole slide images (WSIs) of OED cases between 2005 to 2016. WSIs were scanned at ×20 using an Aperio CS2 scanner (n = 66) and at ×40 using a Hamamatsu scanner (n = 97) after ethical approval (REC Reference- 18/WM/0335, NHS Health Research Authority West Midlands). Amongst 163 cases, 137 were OED cases with 50 transformed into malignancy. The remaining cases were non-dysplastic oral mucosal biopsies, including benign hyperkeratosis or mild epithelial hyperplasia. The mean average age in the dataset of OED cases was 64.64 (range 25-97), with the mean age for men (n = 84) was 66.3 and the mean age of women (n = 53) being 64.5. The main clinical sites of involvement were the tongue, floor of the mouth and buccal mucosa. The mean time to malignant transformation was 6.51 years (±5.35 SD). The inclusion criteria for WSIs were decided upon the following conditions:

- A histological diagnosis of OED
- Sufficient availability of tissue, i.e., (excluding tangentially cut sections, tissue with artefacts).
- Minimum five-year of follow-up data (including treatment, recurrence and transformation information) from the initial diagnosis
- All cases were independently seen by at least two certified/consultant pathologists.

The interobserver disagreement between the two pathologists was assessed using Cohen's kappa score, which resulted in a value of 0.854. The score indicates a high level of agreement between the two pathologists. Cases with disagreement were resolved through discussion within the team. More information about the cohort can be seen in Table 4.1. Epithelium masks were obtained using HoVer-Net+ [100] and then refined manually for some cases (i.e., removing blood vessels being recognised as epithelium layers). In contrast, slide-level labels were obtained for each case from patient records (i.e., clinical notes and biopsies), including histological grades, recurrence status, and malignant transformation status (i.e., OED has progressed into OSCC at the same diagnosed location within the follow-up time). The WSIs were split into train and test sets using three different stratified 5-folds on transformation

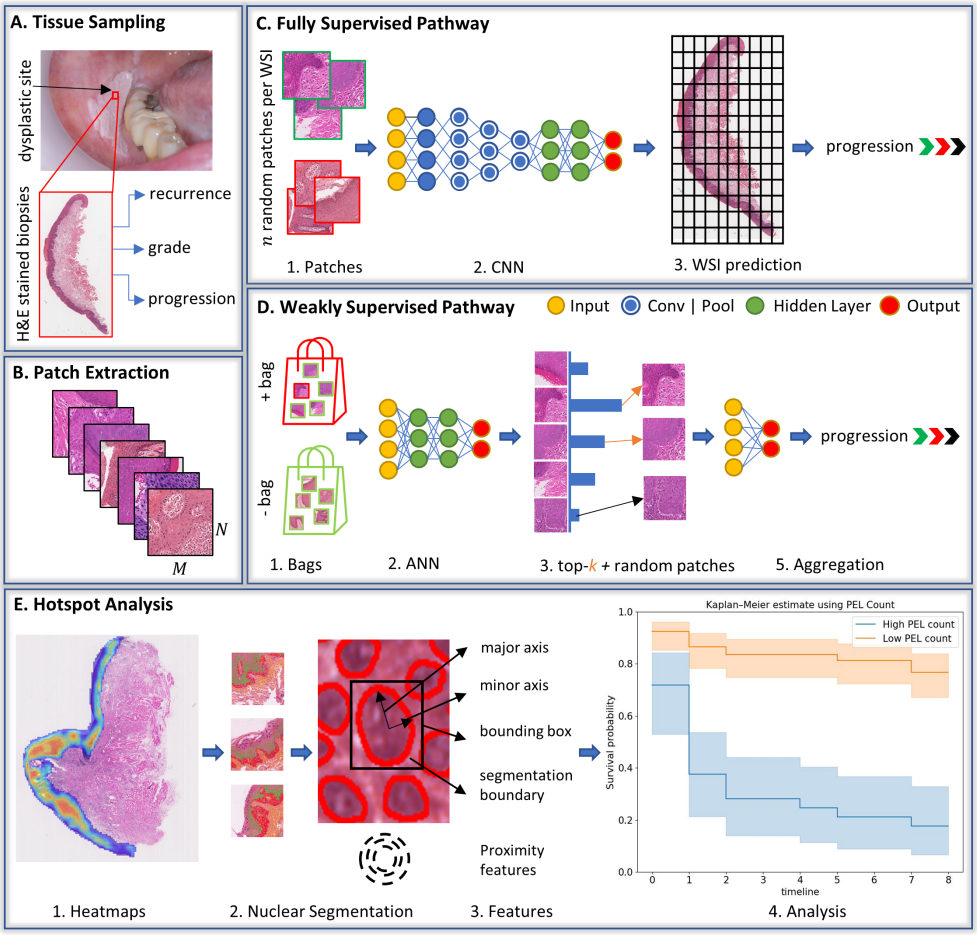


Figure 4.1: The overall workflow of the study is shown in different sections. A) the process of getting the tissue biopsies from dysplastic lesions and corresponding WSIs with their associated labels assigned by a pathologist. B) patches of size M×N were extracted from the epithelium region of WSIs. C) fully supervised pipeline where the patches were assigned the WSI level labels and trained using CNNs for the downstream tasks. D) weakly supervised pipeline where positive (+ive) and negative (-ive) batch of features/images was created and used for training. E) heatmaps were generated using IDaRS to explore the hotspot areas and their contribution towards the malignant transformation prediction using nuclear analysis. Nuclear features from different layer of epithelium i.e., keratin (blue nuclei), epithelial (green nuclei), basal (red nuclei) and tissue area (orange nuclei) from the hotspot and cold spots were used for progression free survival by using peri-epithelial lymphocytes (PELs) count.

Characteristic	**Number** (%)
OED cases	137
Cases with malignant transformation	50 (36.4%)
WHO grade	
Mild	41 (29.9%)
Moderate	53 (38.6%)
Severe	43 (31.3%)
Binary grade	
Low-risk	80 (58.3%)
High-risk	57 (41.6%)
Mean age [min-max]	64.64 [25-97]
Gender	
Male	84 (61.3%)
Female	53 (38.6%)
Clinical (intra-oral) site	
Tongue	53 (38.6%)
Floor of mouth	27 (19.7%)
Buccal mucosa	17 (12.4%)
Others	38 (27.7%)
Survival	**Mean (Standard Deviation)**
Survival (Months)	84.75 (63.03)
Survival (Year)	6.51 (5.35)

Table 4.1: Characteristic of the cohort used for the study with clinical and demographic information of OED cases.

status for all experiments. Patches of size 512×512 were extracted using the epithelium mask with an overlap of 50% from all the WSIs at 0.50µ per pixel (mpp). For extracting the deep features, ResNet-50 [135] was used as a feature extractor pre-trained on ImageNet. A feature vector of size 1024 was extracted for each patch resulting in a bag of shape xR^{n1024} for all WSIs (where n is the number of patches extracted).

4.2.2 Methods

Malignant Transformation Prediction

Figure 4.1 shows the overall pipeline, which involves initially extracting X patches of size M×N with slide level labels Y from WSIs with an overlap of O using the epithelium mask. Extracted patches were utilised for training the deep learning models for predicting malignant transformation. In this chapter, we used iterative draw-and-rank sampling (IDaRS) [54], which works by ranking and selecting the top and random patches from a WSI, assuming that not all patches are equally important and predictive of the outcome. IDaRS selects two subsets of patches for training, including random patches r and top-ranked patches k for each WSI. Both subsets are then pre-processed using

the standard set of augmentations and train a CNN with weak labels. We have also compared the IDaRS with other fully supervised and weakly supervised algorithms, e.g., multi-layer perceptron (MLP), Attention-MIL (A-MIL) [205], clustering constrained attention multiple instance leaning (CLAM) [104], and CNN based benchmark classification models (ResNet [135], DenseNet [206] and Vision Transformers [146] with max pooling as an aggregator for the final WSI label).

Table 4.2: Nuclear features extracted from layer wise nuclei and their explanations.

Feature	**Explanation**
Extent (EX)	Ratio of bounding box pixels to total region
Equivalent diameter (ED)	Diameter of the circle in the bounding box
Eccentricity (ECC)	Ratio of focal distance over major axis
Convex area (CA)	Number of pixels in the convex hull
Centroid (C)	Centre location of bounding box
Major axis length (MJL)	Length of the major axis
Minor axis length (MNL)	Length of the major axis
Nuclei count (NC)	Total number of nuclei in the patch
Cellularity per micron (ϕ)	Nuclei density in patch per micron
Nearest neighbour distance (NND)	Nearest nucleus distance from nucleus of interest

Cellular Composition Analysis

To further analyse and validate the hotspots being identified by the IDaRS model, cellular compositions of top tiles (i.e., hotspots and coldspots) from transformed and non-transformed cases were analysed. Nuclear features were extracted from each layer (i.e., keratin, epithelial, and basal see Fig 1.1) and associated connective tissue in an automated manner using nuclear segmentation and classification. For this purpose, input patches were first stain normalised using a sample from The Cancer Genome Atlas (TCGA) cohort before being fed into HoVer-Net [42], which was pre-trained on the PanNuke dataset [196] for nuclear instance segmentation and classification. For segmentation of the keratin, epithelial and basal layers within the epithelium, HoVer-Net+ [100] was used. Figure 4.2 and Table 4.2 shows a range of morphological and proximity features extracted from the segmented image patches and aggregated statistically using the minimum $\wedge$, maximum $\vee$, mean μ, median m and standard deviation σ. Here, ordinary least square (OLS) was used with post-hoc t-tests for calculating the statistical significance with Benjamini/Hochberg adjustment

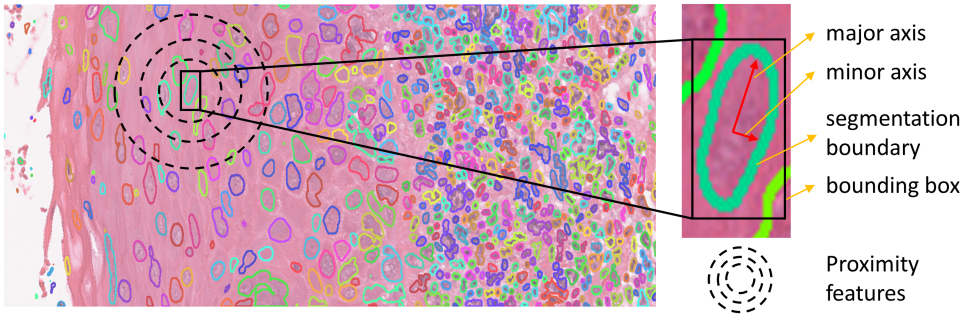


Figure 4.2: Shows the patch with nuclei instance segmentation (left) and segmented region of a nucleus in green boundary (right) where black box represents the bounding box and red lines represent the major and minor axis while the green area represents the segmentation boundary. The black concentric circles (left) represented the neighbourhood of the nucleus and were used for extracting spatial features, e.g., distance to closest nuclei (proximity).

[2]. Cellular composition helps understand/interpret the results of IDaRS and differentiates transformed cases from non-transformed ones in an objective manner.

Peri-epithelial Lymphocytes (PELs) Count

Elevated PEL counts can be linked to a higher risk of malignant transformation in oral epithelial dysplasia (OED),. To further explore the role of PEL count in transformed and non-transformed cases, a Wilcoxon rank-sum test was performed where $p < 0.05$ was considered significant. Moreover, we also analysed the distributions of PEL count in subgroups based on two clinical features, i.e., gender and age. Gender was divided into male and female groups. The age subgroups were separated into ranges between 0-50, 51-70, and 71-100.

Survival Analysis

To investigate the prognostic significance of the clinical, pathological, and nuclear features for progression free survival (PFS), Kaplan–Meier (KM) curves and Cox proportional hazard (CPH) models were used for univariate and multivariate analysis. To distinguish between the high-risk (short term survival) and low-risk (long term survival) groups, the optimal cut-off value was calculated by taking the mean of hazard value for each instance using the CPH model where the statistical significance is large between the high and low-risk groups. Further, a long-rank test was performed to determine the statistical significance and $p < 0.05$ was considered statistically significant.

4.2.3 Experimental Settings

For IDaRS, we set the random patches r=30 and top-patches k=5 and trained a pretrained ResNet-34 on ImageNet with a batch size of 16 and patch size of 256. IDaRS was trained for 30 epochs with binary cross-entropy loss and optimised using the Adam optimiser. For training, MLP and CLAM deep features were then fed as input to the models for generating WSI-level outputs. MLP and CLAM were trained for 1000 epochs using the default configurations from the CLAM. For A-MIL and CNN models' the same input and configurations as IDaRS were used for the training and test purposes. All models were trained and tested on a system with two Nvidia Titan-X with 12 GB of memory, dedicated RAM of 128GB, and an Intel® Core i9 processor. Where, an average epoch takes 10 minutes, and downstream analysis for a single WSI takes 1 minute.

4.2.4 Evaluation metrics

To validate the results, stratified on transformation status, 5-fold cross-validation was performed three times with different random seeds. Patch-wise AUROC and F1-score (macro) aggregated at WSI level were used as performance metrics and are averaged across the folds. F1-score (macro) computes the arithmetic mean of the F1-score per class, treating all classes equally and regardless of their number. AUROC evaluates the binary problems by plotting the true positive rate (TPR) against the false positive rate (FPR) at various thresholds. The area under the ROC curve (AUROC) measures the ability of the classifier to differentiate between the two classes where the TPR and FPR are calculated as:

$$TPR = \frac{TruePositives}{TruePositives + FalsePositives}$$

$$FPR = \frac{FalsePositives}{FalsePositives + TrueNagatives}$$

4.3 Results and Discussion

4.3.1 Malignant Transformation

The results of our experiments are shown in Table 4.3 indicate that the performance of IDaRS is comparatively better than the other weakly and fully supervised algorithms with an AUROC of 0.78 (±0.07 SD) and F1-score of 0.69 (±0.05 SD) as compared to MLP, CLAM, and A-MIL. It can also be observed from the ROC plots in Figure 4.3 that the standard deviation across different folds for IDaRS is smaller as compared to the other weakly supervised algorithms. It is worth noting that the performance of CLAM is competitive to

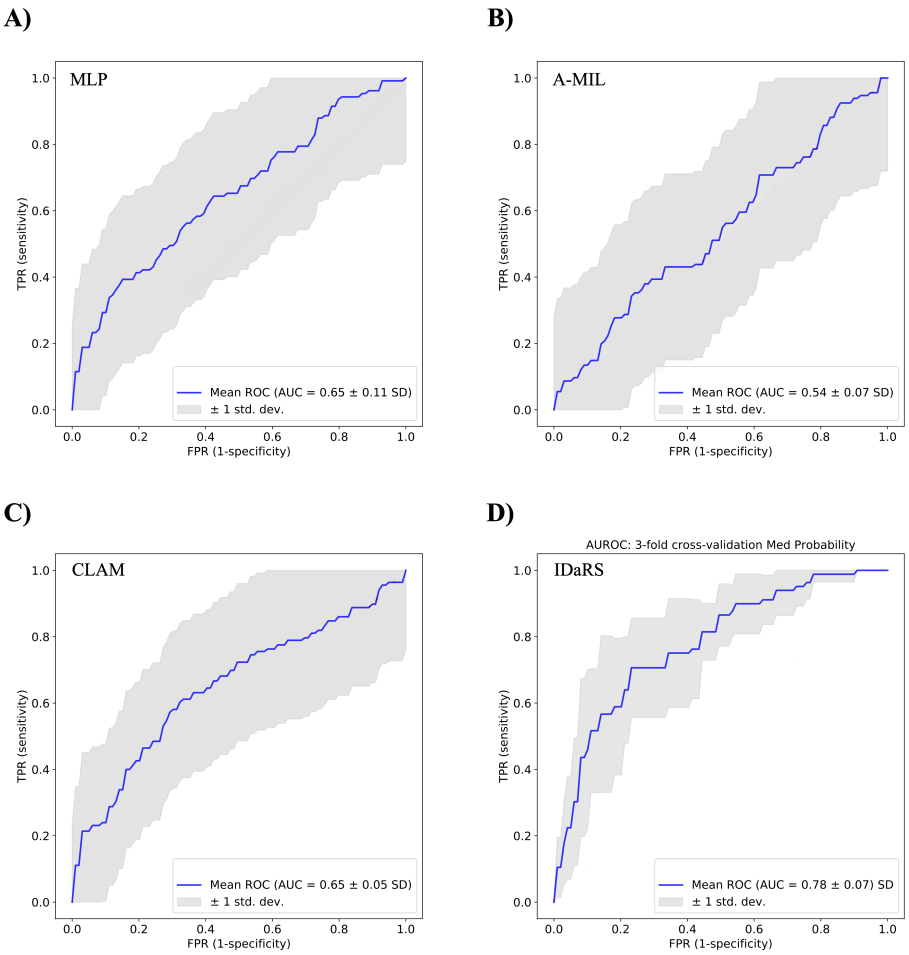


Figure 4.3: ROC curve plots on 5-fold cross-validation for OED malignant transformation prediction using (A) MIL (B) A-MIL (C) CLAM and (D) IDaRS.

Table 4.3: Performance of IDaRS model as compared to other weakly supervised and fully supervised models with deep features, IDaRS is achieving high performance in terms of AUROC. SD = Standard Deviation

Model	**top-*k***	**AUC ± SD**	**F1-score ± SD**
MLP	1	0.65 ± 0.09	0.56 ± 0.11
	5	0.64 ± 0.11	0.55 ± 0.01
A-MIL [205]	-	0.54 ± 0.07	0.44 ± 0.30
CLAM [104]	1	0.65 ± 0.04	0.64 ± 0.04
	5	0.65 ± 0.05	0.63 ± 0.01
IDaRS [54]	5	**0.78 ± 0.07**	**0.69 ± 0.05**
ResNet-50 [135]	-	0.54 ± 0.10	0.43 ± 0.11
ViT [146]	-	0.55 ± 0.01	0.44 ± 0.08
DenseNet [206]	-	0.56 ± 0.05	0.44 ± 0.01

IDaRS as compared to the MIL in terms of the F1 score. The performance of CLAM was competitive to IDaRS as compared to the MIL in terms of F1-score. The fully supervised networks performed worse than other weakly supervised models due to the inherent nature of the problem which introduces noise in the labels and corrupting the model's training.

Feature	$p > \|t\|$	$p > \|t\|$ **(adjusted)**
Tissue NC	0.0013	0.0481*
Tissue σ Nuclei in 100 mpp	0.0.289	0.2755
Tissue max ECC	0.0428	0.3491
Basal μ minor axis length	0.0436	0.3491
Basal σ ED	0.0090	0.1672
Basal NC	< 0.0001	< 0.0001*
Epithelium μ ECC	0.0015	0.0487*
Epithelium μ NND	0.0099	0.1672
Epithelium μ Nuclei in 100 mpp	0.0125	0.1273
Epithelium σ ECC	0.0028	0.0729
Epithelium σ Bounding Box	0.00106	0.0487*
Epithelium NC	< 0.0001	< 0.0001*

Table 4.4: Ordinary least square regression for malignant transformation with t-test significance of nuclear features with Benjamini/Hochberg (Benjamini & Hochberg, 1995) adjustment. Only the top nuclear features are shown here where the significant p-value is highlighted using *. σ represents the standard deviation and μ represents the mean of a distribution.

4.3.2 Exploring the visual patterns

To validate and further investigate the features learnt by the top performing IDaRS, we explored the top tiles from the heatmaps of transformed and non-transformed WSIs. For correlating the hotspot/coldspots with the clinical features, heatmaps were also analysed manually for corroboration purposes

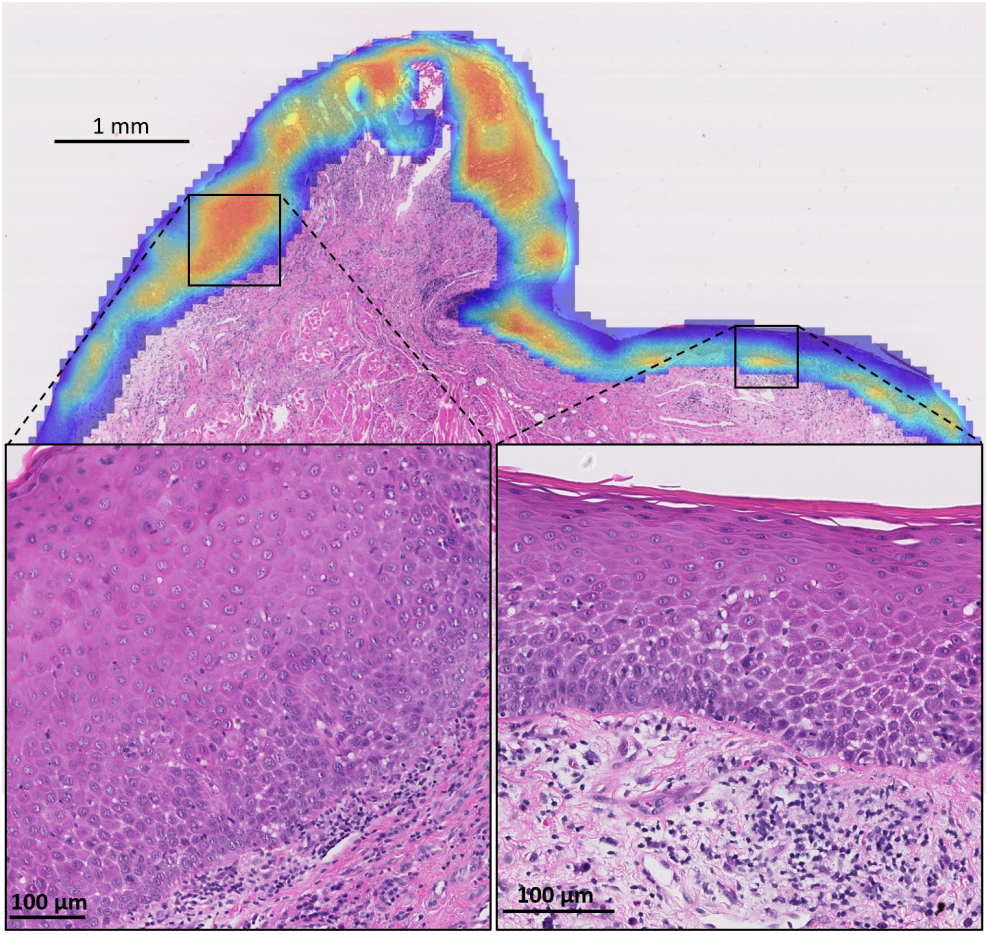


Figure 4.4: Heatmap of high-risk OED case for the malignant transformation predicting using IDaRS. Red regions in the heatmap overlay shows a high probability of malignant transformation in respective areas. From those high probability region two of them are being shown in more detail in the two black boxes.

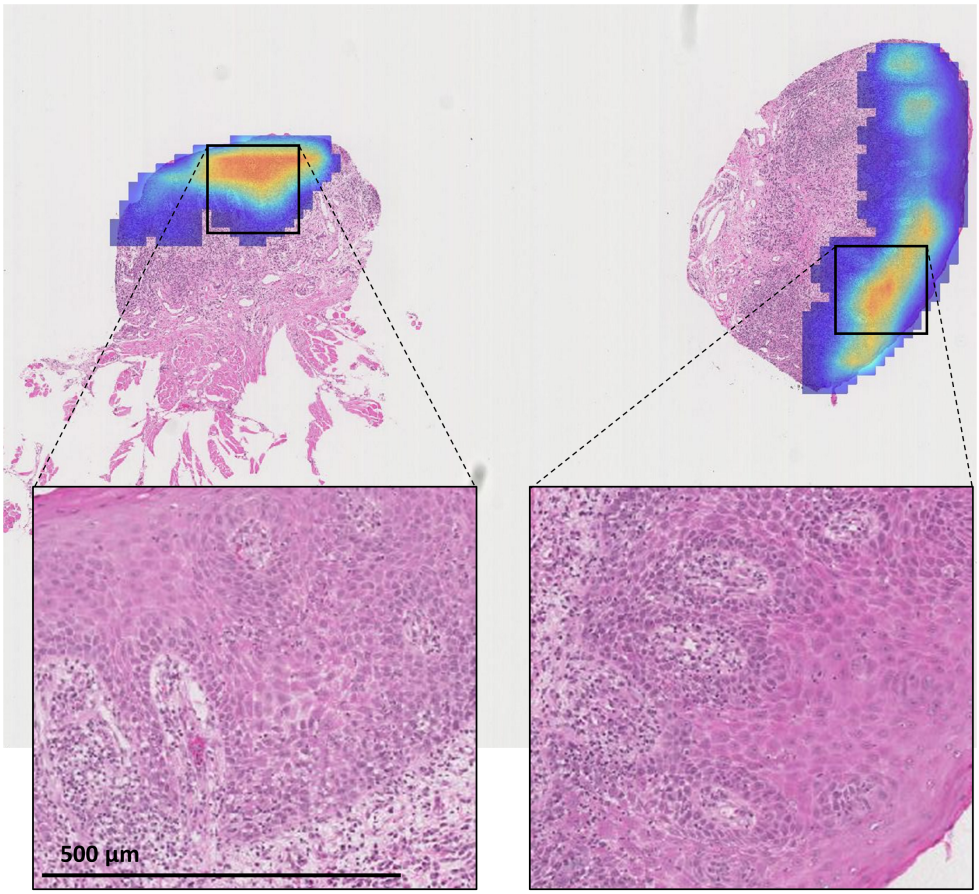


Figure 4.5: Heatmap of low-risk OED case for the malignant transformation predicting using IDaRS. The red region shows a high probability of malignant transformation in those areas. From those high probability region two of them are being shown in more detail in the two black boxes.

by an expert pathologist. Figure 4.4 shows the heatmap for a histologically high-risk case where the red (hotspot) colour represents a region with a higher probability of malignant transformation. In contrast, the blue (coldspot) colour corresponds to a region with a low probability of transformation. Closer examination of hotpots shows evidence of disordered stratification, dyskeratosis, as well as nuclear and cellular pleomorphism with a dense lymphocytic infiltrate in the adjacent peri-epithelial connective tissue. Similarly, Figure 4.5 shows the heatmap for a histologically low-risk case where cellular pleomorphism with a lymphocytic infiltrate can be seen in the hotspot regions. The dense lymphocytic infiltrate is referred to as peri-epithelial lymphocytes (PELs) for the rest of the analysis.

4.3.3 Cellular Composition Analysis

Following the manual analysis of the heatmaps, automated cellular composition analysis was performed to uncover significant hidden patterns/features in transformed vs non-transformed cases. Table 4.4 shows the prognostic significance of the extracted nuclear features for predicting malignant transformation. For the epithelial layer, variation in eccentricity ($p = 0.048$), bounding box ($p = 0.0487$) and total nuclei count ($p < 0.0001$) showed significance along with Basal layer NC ($p < 0.0001$). An increase in cell count (hyperplasia or crowding) is an important feature observed in high-risk dysplasia in both the central epithelium layer and specifically within the basal layer. Other features in the epithelium, e.g., variation in nuclei count (100µm per pixel) and nearest nuclei distance, correspond to congestion in spatial arrangements of epithelial nuclei and require more data for validation. Similarly, changes in basal layer nuclei minor axis, equivalent diameter corresponds to the nuclear pleomorphism and are observed in high-risk OED cases. Interestingly, the nuclei count in the connective tissue area also showed significance for predicting the transformation ($p = 0.0004$), which corresponds to the previous observation regarding the dense lymphocytic infiltrate in the adjacent peri-epithelial connective tissue.

4.3.4 Peri-Epithelial Lymphocytes (PELs)

Figure 4.6 shows examples of patches from both hotspots (red) and coldspots (blue) regions of the transformed and non-transformed cases with their corresponding layer-wise cellular compositions. For most of the coldspots, the epithelium and basal nuclei are dominant, whereas in the hotspots (red) PELs are in abundance in the transformed cases compared to non-transformed cases (Figure 4.6). As a whole, PELs were statistically significant ($p = 0.02$) for differentiating between the transformed vs non-transformed cases. Gender based subgrouping showed no significance between male and female groups.

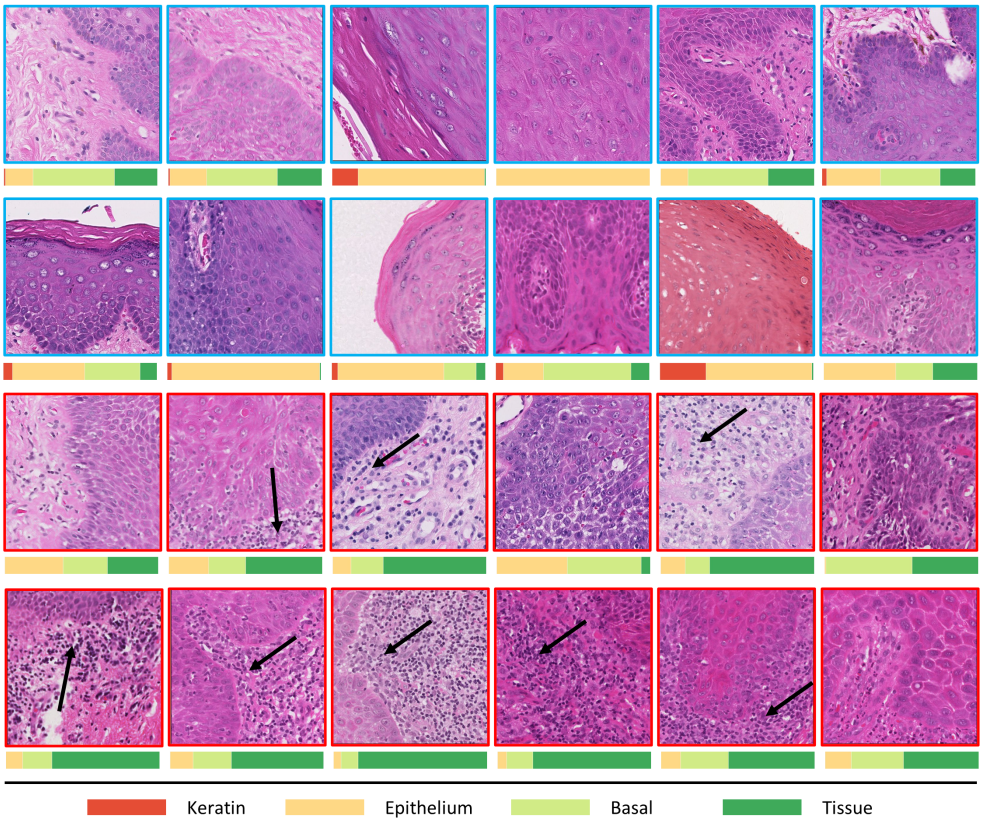


Figure 4.6: Patches extracted from the hotspot (red) and coldspots (blue) of the WSIs with their layer wise nuclear composition. Most of the coldspot regions have dominant epithelial nuclei as compared to the hotspots where PEL can be seen dominating the overall ratio.

However, for age, the 0-50 group showed prognostic significance with respect to malignant transformation with $p = 0.001$. Figure 4.7 shows the boxen plots for (A) the overall distribution of the PEL ratio in transformed cases versus non-transformed cases and (B) the distribution of the PEL ratio in transformed cases versus non-transformed cases, including age subgrouping.

Features	**Aggregation**	***p***	**C-index**	**Lower 95%**	**Upper 95%**
Clinical Parameters					
Gender	-	>0.05	0.52	0.52	0.53
Age	-	>0.05	0.59	0.59	0.60
Pathological Parameters					
WHO Grading (Mild vs Mod + Severe)	-	<0.05	0.68	0.68	0.69
WHO Grading (Mild + Mod vs Severe)	-	<0.05	0.68	0.68	0.68
Binary Grading	-	<0.05	0.68	0.68	0.69
Nuclear Features					
PEL count	μ	>0.05	0.45	0.45	0.46
	σ	<0.05	0.60	0.59	0.60
	m	>0.05	0.57	0.56	0.58
	$\wedge$	<0.05	0.73	0.72	0.73
	$\vee$	>0.05	0.53	0.52	0.54
Basal NC	μ	>0.05	0.45	0.44	0.46
	σ	<0.05	0.66	0.65	0.67
	m	>0.05	0.52	0.51	0.53
	$\wedge$	<0.05	0.70	0.69	0.71
	$\vee$	>0.05	0.53	0.52	0.54
Epithelium NC	μ	<0.05	0.65	0.64	0.65
	σ	<0.05	0.72	0.71	0.73
	m	<0.05	0.66	0.65	0.67
	$\wedge$	<0.05	0.73	0.73	0.74
	$\vee$	>0.05	0.46	0.45	0.47

Table 4.5: Univariate analysis of the clinical, pathological and digital features where p is calculated using the log-rank method, and C-index is calculated using the Cox Proportional Hazard model bootstrapped 1000 times for lower and upper confidence intervals. $\wedge$ represents minimum, $\vee$ represents maximum, μ represents mean, m represents median and σ represents standard deviation.

4.3.5 Survival Analysis

Table 4.5 shows the univariate analysis of the aforementioned nuclear features mentioned in Cellular Composition Analysis with clinical and pathological features, where it can be seen that both clinical features, age ($p > 0.05$, C-

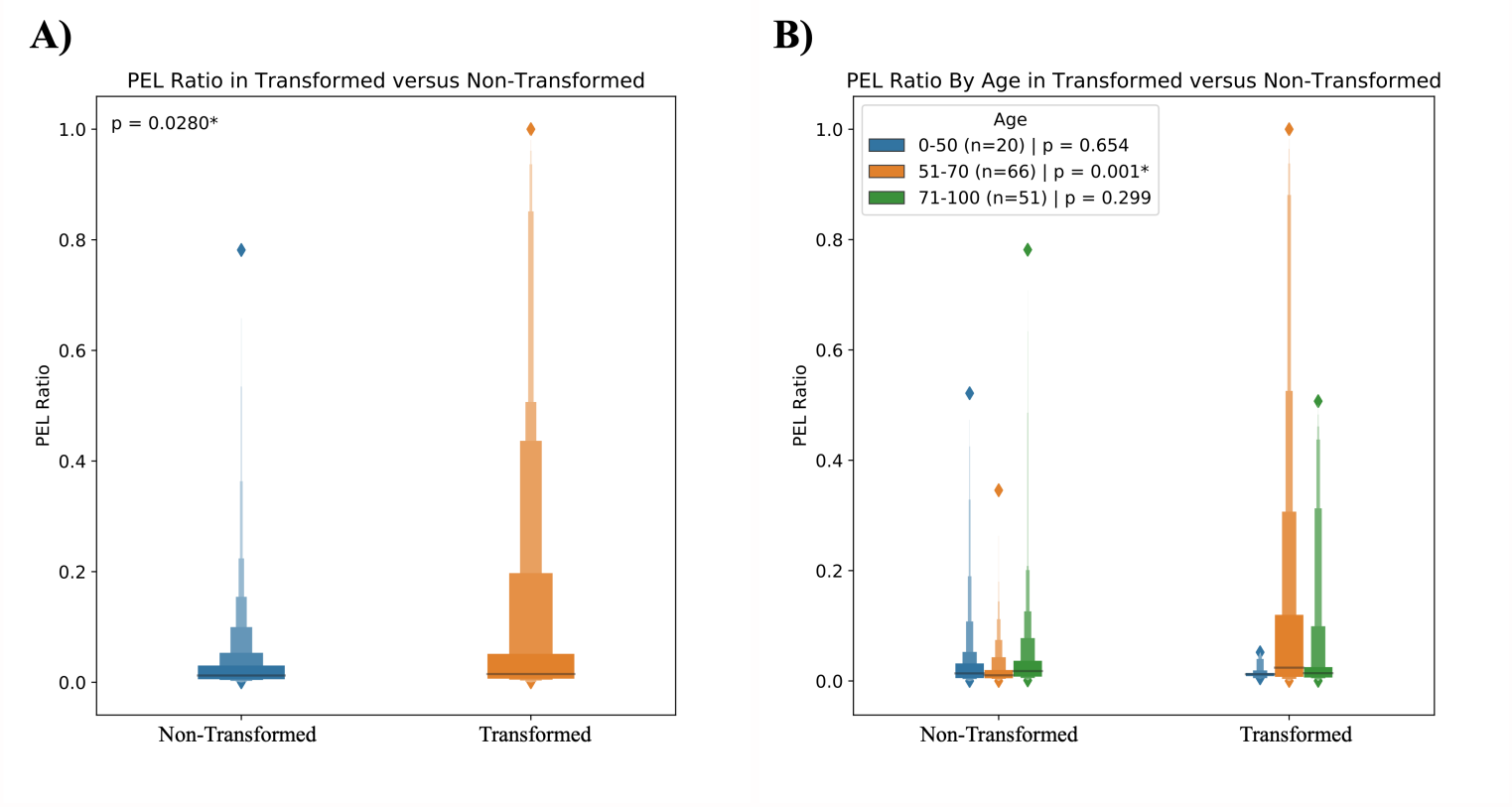


Figure 4.7: (left) Shows the boxen plot for the ratio of PELs present in both transformed and non-transformed patches and (Right) Shows the further breakdown of the PEL ratio in age groups where it can be seen that the 0-50 age group has a distinct difference in PEL ratio as compared to the other groups.

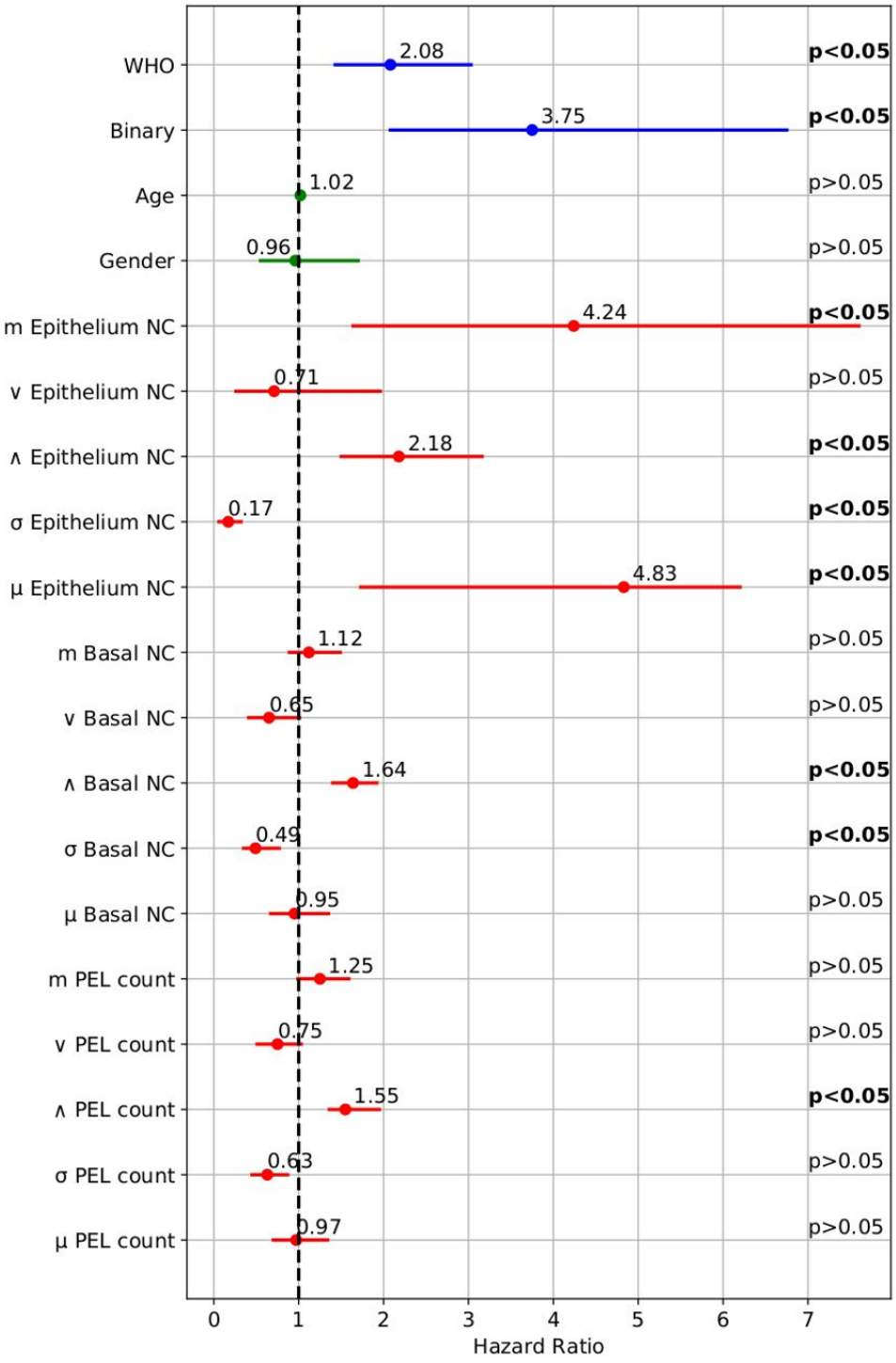


Figure 4.8: Univariate analysis of different features (blue) pathological grades i.e., WHO grading and binary grading, (green) clinical and (red) top most significant nuclear. For each feature, the dot represents the hazard ratio, and the filled line shows the lower and upper confidence interval of 95%. p-values were shown at the right, calculated using the Wald test. $\wedge$ represents minimum, $\vee$ represents maximum, μ represents mean, m represents median and σ represents standard deviation.

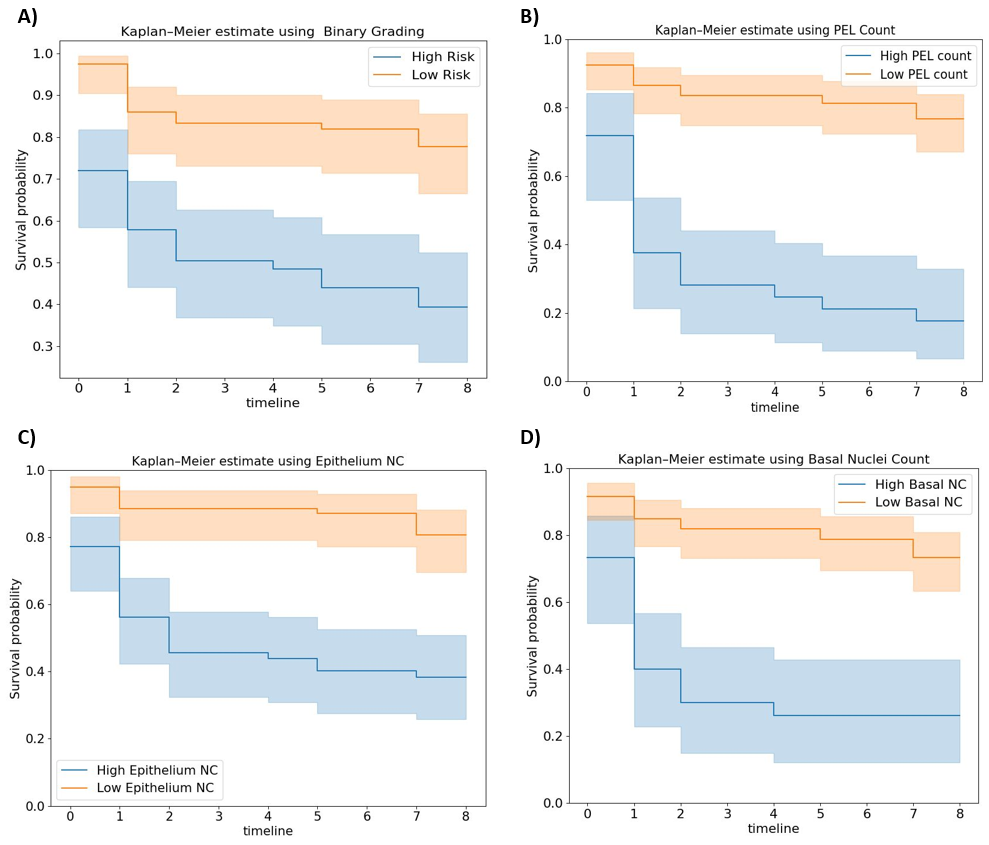


Figure 4.9: Kaplan–Meir (KM) curve for progression free survival of OED using (a) Binary Grading, (b) PEL count, (c) Epithelium layer NC and (d) represents the KM curve using the basal layer nuclei count.

index = 0.59 [95%, 0.59 – 0.60]) and gender ($p > 0.05$, C-index = 0.52 [95%, 0.52 – 0.53]) are nonsignificant. Conversely, the pathological features showed significance for binary grading ($p = 0.004$, C-index = 0.68 [95%, 0.67 – 0.69]) and WHO based grading when moderate and severe cases were combined against mild grade ($p = 0.04$, C-index = 0.68 [95%, 0.67 – 0.68]). When mild and moderate cases were combined and compared against severe, they showed the same significance as ($p = 0.04$, C-index = 0.68 [95%, 0.67 – 0.68]). The nuclear features extracted from the epithelial layer, basal layer and connective tissue area also showed significance for the minimum number of nuclei count (NC) in the basal layer ($p < 0.05$, C-index = 0.70 [95%, 0.69 – 0.71]), epithelial layer ($p < 0.05$, C-index = 0.73 [95%, 0.73 – 0.74]) and PELs ($p < 0.05$, C-index = 0.73 [95%, 0.72 – 0.73]). Figure 4.9 shows the KM curves for (A) binary grades, (B) PEL count, (C) epithelium layer NC and (D) basal layer NC, where it can be seen that all features are statistically significant in differentiating the high risk and low-risk lesions with a clear separation between the two groups, especially the PEL count. Figure 4.8 shows the hazard ratio (HR) for variation in basal layer NC and epithelium layer NC appears to be associated with improved survival, whereas the minimum PEL count, epithelium layer NC and basal layer NC are the adverse predictors of PFS. Furthermore, Table 4.6 shows the multivariate analysis of the most significant nuclear and pathological features (i.e., binary grading, *min* epithelial layer NC, *min* basal layer NC and *min* PEL count) to examine their combined effect on the PFS. When these features are combined, the C-index improves by reaching 0.79 [95%, 0.78 – 0.80], with binary grading, epithelium layer NC and PEL being the most significant prognostic features for malignant transformation. In the absence of binary grading, the C-index achieved using nuclear features only is competitive, reaching 0.78 [95%, 0.77 – 0.78]. Similarly, combined binary grading with PEL counts reached the same C-index of 0.78 [95%, 0.77 – 0.78] as compared to the other two features with binary grading, i.e., epithelium layer NC 0.76 [95%, 0.75 – 0.77] and basal layer NC 0.77 [95%, 0.76 – 0.77]. This highlights the importance of using PEL counts as a prognostic feature for predicting malignant transformation. Further, the combined performance of basal layer NC and epithelium layer NC with PEL count also shows the significance of using PEL in conjunction with other clinical and nuclear features.

In this chapter, we explored the potential of deep learning for predicting malignant transformation from digitised OED histology slides. We trained a weakly supervised learning framework for malignant transformation prediction and further analysed the predictive "hotspots" in epithelial and peri-epithelial tissue regions. We have demonstrated that deep learning based weakly supervised IDaRS can predict malignant transformation with an AUROC of 0.78 ($\pm$0.07 SD) on stratified 5-fold cross-validation using three different random

Feature	p	HR	Lower 95%	Upper 95%
C-index = 0.79, 95% CI [0.78 – 0.80]				
Binary Grading	<0.05	2.43	1.30	4.54
Basal NC	>0.05	1.04	0.79	1.37
PEL count	<0.05	1.72	1.24	2.37
Epithelium NC	<0.05	1.48	1.07	2.05
C-index = 0.78, 95% CI [0.77 – 0.78]				
Basal NC	>0.05	1.08	0.81	1.43
PEL count	<0.05	1.72	1.23	2.39
Epithelium NC	<0.05	1.67	1.20	2.32
C-index = 0.77, 95% CI [0.76 – 0.77]				
Binary Grading	<0.05	2.97	1.62	5.43
Basal NC	<0.05	1.54	1.31	1.81
C-index = 0.78, 95% CI [0.77 – 0.78]				
Binary Grading	<0.05	3.10	1.70	5.65
PEL count	<0.05	1.81	1.50	2.18
C-index = 0.76, 95% CI [0.75 – 0.77]				
Binary Grading	<0.05	2.76	1.44	4.93
Epithelium NC	<0.05	1.84	1.27	2.66
C-index = 0.73, 95% CI [0.72 – 0.74]				
Basal NC	>0.05	1.13	0.84	1.52
PEL count	<0.05	1.68	1.20	2.34
C-index = 0.77, 95% CI [0.77 – 0.78]				
Epithelium NC	<0.05	1.67	1.19	2.35
Basal layer NC	<0.05	1.54	1.29	1.83
C-index = 0.78, 95% CI [0.77 – 0.78]				
Epithelium NC	<0.05	1.68	1.21	2.34
PEL count	<0.05	1.83	1.50	2.25

Table 4.6: Multivariate analysis of the pathological and digital features where p is calculated using the Wald test, and C-index is calculated using the Cox Proportional Hazard model bootstrapped 1000 times for lower and upper confidence intervals.

seeds. The higher performance of IDaRS as compared to other MIL algorithms is because it dynamically learns important feature representations from the patches internally, as compared to fixed feature representation of a patch as an input limiting the learning possibilities of the model. Mahmood et al. [204] also reported the AUROC of 0.77 for transformation using a similar but smaller cohort with the nuclear features subjectively assessed by three pathologists.

We have also explored the cellular compositions (i.e., nuclear features) and their role in potentially malignant areas (i.e., hotspots) of transformed cases and compared them to the non-transformed areas (i.e., coldspots). Nuclear features from the epithelial layer and associated connective tissue area were found to be the most significant prognostic features for predicting malignant transformation. Other important features found in the epithelial and basal

layer during the experiments were variations in the number of nuclei in 100µm per pixel (mpp), the standard deviation in cell eccentricity, mean major and minor axis length etc. These nuclear features also correspond to the aberration of nuclei (i.e., variation in size of nuclei captured as a variation in the minor axis of the nuclei and convexity of the nuclear shape) and congestion due to proliferation of nuclei in the epithelial and basal layer. However, in order to verify the significance of these features we require more data to test these features' ability to indicate prognostic significance for malignancy. It has also been reported in the literature that the PELs can play an important role in transforming dysplasia into carcinoma [207]. There is a possible explanation for the transformation that the epithelium is affected by the PEL. This can be due to the release of cytokines linked with oxidative stress, transforming the epithelial cells into premalignant ones [208, 209] as we have seen that PELs showed significance for predicting the transformation with $p < 0.05$. For PFS, we examined clinical, pathological, and nuclear features of oral epithelial dysplasia. Our findings indicated that, in addition to binary grading, the variation in Basal layer NC and Epithelial layer NC were associated with improved PFS. On the other hand, we observed that the minimum number of nuclei in Basal layer, Epithelial layer, and PEL were linked to a higher risk of malignant transformation or poor survival. Gan et al. [207] have also investigated the potential role of lymphocytic infiltration in malignant transformation by analysing the RNA sequencing of the immune infiltration sites in moderate and severe OED. The authors highlighted the importance of immune signatures established from oral cancer to identify three distinct subtypes of moderate and severe OED: immune cytotoxic, non-cytotoxic and non-immune reactive from transcriptional data. Their findings suggest that the lack of CD8 T-cells in the non-cytotoxic subtype and non-immune reactive subtype can lead to progression in moderate and severe dysplasia. The chapter identified binary grading as a significant indicator for malignant transformation in oral epithelial dysplasia (OED), whereas the study performed by Dost et al. [199] did not find any association between grading and transformation. However, Mahmood et al. [204] demonstrated an association between nuclear features used for OED grading (e.g., bulbus rete pegs, loss of epithelial cohesion etc.,) and malignant transformation. Similarly, Gilvetti et al. [203] demonstrated the importance of various clinical features, including age, in predicting outcomes for oral epithelial dysplasia (OED). Our study also found age to be a significant prognostic factor in one of the subgroups (0-50) with a p-value of 0.001, corroborating the findings of. We also found in our multivariate analysis that when we combined these pathological and nuclear features for PFS, it improved the results specifically due to the addition of epithelium layer NC and PEL count. However, an interesting avenue in future would be to analyse and

investigate the role of dysplasia infiltrating lymphocytes (DILs) in malignant transformation. Although the cohort is small and uni-centric, the department in question is a regional and national referral centre in the UK. Nonetheless, the practical application and adaptation of these methods in clinical practice require substantially large and truly multicentric cohort data allowing more rigorous validation of the proposed algorithms.

4.4 Chapter Summary

To best of our knowledge this is the first study to propose and show the association of peri-epithelial lymphocytes (PELs) count in malignant transformation along with other digital biomarkers, e.g., epithelium layer NC and basal layer NC. Our multivariate feature analysis has shown that PELs and epithelial NC have shown to improve the prognostic value in conjunction with binary OED grading for predicting malignant transformation. Our proposed methodology for predicting malignancy in an end-to-end manner has the potential to play an important role in precision medicine and personalised patient management for early prediction of malignancy risk with the potential to guide treatment decisions and risk stratification.

Chapter 5

Coarse Segmentation for OED grading using Graph CNNs

5.1 Introduction

There has been a limited amount of literature available for computational research on whole slide image (WSI) level predictive analysis of OED, as well as on the correlation between OED grade and prediction of malignant transformation. Dost et al. [199] found no association between OED grade and malignant transformation in their study of 368 patients, where 7.1% of the patients had progression of OED to malignancy. While Gilvetti et al. [203] conducted a study involving 120 patients with a mean follow-up of 47.7 months ($\pm$29.9), which showed that patients with erythroplakia had a significant recurrence rate (i.e., $p = 0.02$) and a mean time to recurrence of 62 months ($\pm$31.5). They also found that malignant transformation was significantly associated with age ($p = 0.03$), clinical appearance ($p = 0.03$), lesion site ($p = 0.01$), and some other clinical features, with a mean transformation time of 50 months ($\pm$32.5). Shephard et al. [51] employed nuclear size and shape characteristics to forecast OED progression or transformation in H&E images, achieving mixed outcomes. Mahmood et al. [204] conducted a study on OED biopsies from 109 patients with a minimum of five-year follow-up to identify histological features that could predict recurrence and malignant transformation. They proposed two prognostic models based on specific histological features such as bulbous rete pegs, hyperchromatism, loss of epithelial cohesion, loss of stratification, suprabasal mitoses, and nuclear pleomorphism. The models had an area under the receiver-operator characteristic curve (AUROC) ≥ 0.77 for malignant transformation and AUROC ≥ 0.72 for recurrence. The study highlighted the significant link between OED features and clinical outcomes, although automation of the feature extraction at the WSI level without the aid of a pathologist is yet to be achieved. Bashir et al. [46] developed a

weakly-supervised framework to predict the malignant transformation status in oral epithelial dysplasia (OED) at the whole slide image (WSI) level on a cohort of 163 cases. Their study found that peri-epithelial lymphocytes (PELs) were a significant prognostic feature in correctly predicted cases. However, the model required manually refined epithelial masks to extract epithelial patches rather than additional processing for delineating the different epithelial layers for the downstream analysis.

This highlights the significance of objective OED grading and prediction of malignant transformation. Regardless of the success of deep learning methods in slide-level prediction tasks, these methods struggle to capture the overall organisation and structure of the tissue at both global and local levels. This limitation has led to graph-based approaches in this field, which offer a more principled way of modelling this problem by considering the relationships between individual elements within the tissue. By representing the tissue as a graph, where nodes represent individual elements such as cells or regions and edges represent relationships between them, graph-based methods can capture the complex interactions and dependencies within the tissue. One advantage of Graph Neural Networks (GNNs) is that they are naturally resistant to changes in the rotation and translation of nodes in a graph [210] and can learn increasingly abstract feature embeddings for each node in the graph through message passing between adjacent nodes as the network layers are traversed.

In this chapter, for segmenting the epithelium into sub-layers, we propose a coarse segmentation method addressing the challenges associated with context, accuracy, labelling and complexity. Unlike patch-based classification, it outputs a denser prediction map but is coarser than pixel-based segmentation. We also investigate the effectiveness of using graph neural networks (GNN) with deep, nuclear and fusion features to predict the grade and malignant transformation of OED from digitised WSIs of routine H&E stained histology sections in a comprehensive manner. The tasks included differentiating between high- vs low- risk patients (according to the binary grading system) along with the prediction of malignant transformation. For each task, we compared GNN with other MIL based algorithms and validated it on 5-fold cross-validation using three random trails. We propose a simple two-layer GNN network with an edge convolutional layer outperforming other techniques by achieving high F1-scores and AUROCs for binary grading and malignant transformation. The aims of our work are as follows:

- We propose a coarse segmentation method with feed forward convolutional network only, that can be trained on sparsely annotated data, incorporate more context than patch-based classification and is also faster than pixel-

based segmentation.

- We conduct extensive experiments to compare the efficacy of the coarse segmentation approach with both patch-based classification and pixel-based segmentation methods in terms of accuracy and run time.
- Explore nuclear and deep features based graphs for prediction of OED grades and transformation in histology WSI of lesions.
- Analysis of nuclear features found in the hotspot of heatmaps for OED grading and transformation.
- For progression free survival (PFS), we analysed the clinical, pathological, GNN scores and nuclear features.

Characteristic	**Number** (%)
OED cases	241
Cases with malignant transformation	50 (20.7%)
WHO grade	
Mild	80 (33%)
Moderate	95 (39.5%)
Severe	66 (27.5%)
Binary Grade	
Low-risk	154 (63.7%)
High-risk	87% (36.3%)
Mean age [min-max]	64.2 [25-97]
Gender	
Male	127 (52.6%)
Female	114 (47.3%)
Clinical (intra-oral) site	
Tongue	103 (42.7%)
Floor of mouth	49 (20.3%)
Buccal mucosa	30 (12.4%)
Others	59 (24.4%)
Survival	**Mean (Standard Deviation)**
Survival (Months)	84.95 (49.4)
Survival (Year)	6.74 (4.18)

Table 5.1: Characteristic of the cohort used for the study with clinical and demographic information of OED cases.

5.2 Materials and Methods

5.2.1 Data

The dataset used for this study comprised 241 Haematoxylin and Eosin (H&E) stained and scanned whole slide images (WSIs) of OED cases between 2005 to

2016. WSIs were scanned at ×20 using an Aperio CS2 scanner (n = 143) and at ×40 using a Hamamatsu scanner (n = 98), with more samples available at the time of completion of work on this Chapter. Ethical approval reference is the same as in Chapter 4 (REC Reference- 18/WM/0335, NHS Health Research Authority West Midlands). The mean average age in the dataset of OED cases was 64.64 (range 25-97), with the mean age for men (n = 127) was 66.3 and the mean age of women (n = 114) was 64.5. The main clinical sites of involvement were the tongue, floor of the mouth and buccal mucosa. The mean time for malignant transformation was 6.51 years (±5.35 SD). The inclusion criteria for WSIs were decided upon the following conditions:

- A histological diagnosis of OED
- Sufficient availability of tissue.
- Minimum five-year follow-up data (including treatment, recurrence and transformation information) from the initial diagnosis.
- Review of the pathology by two independent pathologists.

More information about the cohort can be seen in Table 5.1. For coarse segmentation 16282 ROIs for training, 3617 for validation, and 3710 for test of size 512×512 at 10× magnification were extracted from 43 OED WSIs. ROIs were annotated at a pixel-level by the pathologist. The coarse mask for each ROI was generated by aggregating the pixel-level annotation of $k \times k$ mini-patches. From the HNSCC dataset, 24 WSIs (12 from TCGA-HN and 12 from the inhouse dataset) were annotated at mini-patch ($k = 32$) level, which resulted in 141541 training and 38893 test ROIs at 10× magnification. Whereas slide-level labels were obtained for each case from patient records (i.e., clinical notes and biopsies), including histological grades, recurrence status, and malignant transformation status (i.e., OED has progressed into OSCC at the exact diagnosed location within the follow-up time). For the training of coarse segmentation network, WSIs were manually annotated by an expert oral and maxillofacial pathologist. Epithelium masks were obtained using coarse segmentation trained on a set of annotated OED cases and were then refined manually for few cases.

Table 5.2: Frequently used mathematical notations in Chapter 5

Notation	Explanation
x^i	Patch from a dataset
y^i	Coarse segmentation mask corresponding to patch x^i
D	Dataset size
L	Number of patches in the dataset
H	Height of patch x^i
W	Width of patch x^i
h	Height of mini-patch x^i
w	Width of mini-patch x^i
k	Kernel mini-patch in x^i
r^i	Annotated region for the i^{th} class
C	Total number of classes in the dataset
$C^{'}$	Number of classes with regions annotated
P	Expected count of pixels per class
W^i	Weight for the i^{th} class
$\mathcal{L}^{WSCE}$	Weighted sparse cross entropy loss
B	Bag of WSI instances
Z	WSI level labels associated with the bags
T	Total number of WSIs
G^b	Graph representation of a WSI
$F(G^b;\theta)$	Slide level prediction
θ	Trainable parameters in the GNN
h	Deep or nuclear features of patches
g	Geometric coordinates of patches
v	Node representation in a graph
V	Set of vertices in a graph
E	Set of edges in a graph
u_l^o	Feature embedding of node o in layer l
$H^{(l)}$	MLP with trainable parameters θ^l
$\mathcal{L}^{pair}$	Pairwise ranking loss
w^T	Transpose of MLP weight vector w

Table 5.3: Mathematical notations used in this chapter.

To facilitate the reader's comprehension, frequently used mathematical notations are listed and defined in Table 5.3.

5.2.2 Methods

Coarse Segmentation

The proposed framework takes a whole-slide image (WSI) as an input and processes it in a tessellated manner through CSNet to generate a WSI-level segmentation masks. In the training phase, the CSNet model only requires sparsely annotated coarse segmentation masks with their respective image patches, as shown in Figure 5.1. The following sections will explain network

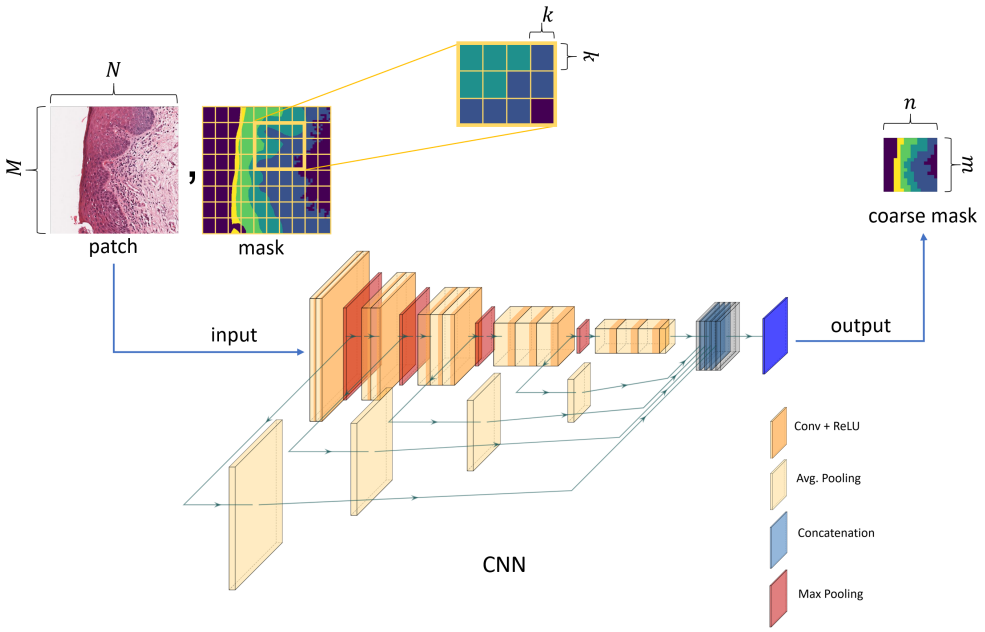


Figure 5.1: The architecture of the proposed method where the input of size $M \times N$ is fed into the coarse segmentation network outputting coarse segmentation mask of size $m \times n$ and size of output depends on the size of mini-patch k. This can be regarded as a type of subsampling.

input, architecture, and training in detail.

Network Input

The input to coarse segmentation network is a patch (x^i) from a dataset, $D = \{x^i, y^i; i = 1, \dots L\}$, containing L patches extracted from WSIs with corresponding coarse level segmentation masks. Each patch x^i is of size $M \times N$, and its corresponding coarse mask y^i is of size $m \times n$ where each pixel in y^i represents a class label for mini-patch of the size of $k \times k$ in x^i. In the case of pixel-level ground truth, the coarse mask can be generated using majority voting of the pixel-level labels in mini-patch. Generated coarse mask is m times smaller than the original mask where $m = \frac{M}{k}$, e.g., a patch x^i of size 512×512 and mini-patch of size $k = 32$ will yield a coarse segmentation mask y^i of size 16×16 where each pixel represents 32×32 pixels in original mask. Choice of k depends on how much accuracy and context is needed, i.e., if $k = M$, then it is equivalent to patch-based classification, whereas $k = 1$ is the same as pixel-wise segmentation as seen in Figure 5.2.

Network Architecture

Any standard convolutional neural network (CNN) can be used as a coarse segmentation network with small modifications in the network architecture, where instead of using the fully connected (FC) layers at the end, a resize

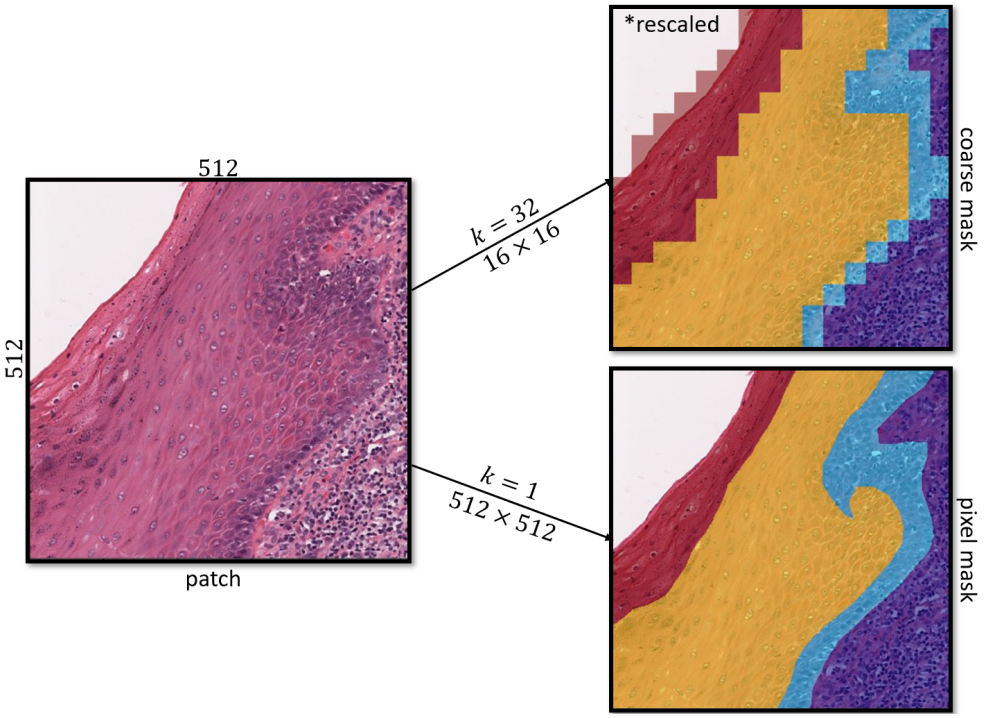


Figure 5.2: Coarse and pixel-wise masks: (*Left*) a visual field of 512×512 pixels showing the epithelium in oral tissue on the left and its corresponding coarse and pixel-wise masks (*Right*). A coarse mask is generated using the mini-patch window of $k = 32$ pixels resulting in the coarse mask of 16×16 pixels.

layer of size of $m \times n$ is used followed by 1×1 convolution to output the coarse prediction map unlike fully convolutional network (FCN) [120] where the output is resized $M \times N$. We used DenseNet [206] based CSNet, where we replaced the last average pooling layer of DenseNet with 1×1 convolution. To further improve the prediction accuracy, we introduced additional skip connections (SC) in the network from each dense block concatenated at the end of the network, as shown in Figure 5.1. These skip connections help the model to improve the spatial context in the final prediction map, as with the use of pooling layers, the spatial context is lost in the final layers. Although the CSNet output is a segmentation mask, there is no encoder-decoder involved in our network design as it is a simple CNN with final classification layers replaced with convolutional layers.

Weighted Sparse Loss

To train CSNet, instead of a trivial cross-entropy (CE) loss function, a sparsely weighted cross-entropy loss function is used to handle the sparsely annotated data. As annotating the entire WSI is a tedious and laborious task, a WSI is often annotated sparsely where there are some un-annotated regions that

can act as noise if being treated as background during training. Also, this can sometimes help in the decision where there is ambiguity for annotators due to inter- and intra- observer variability to let the model decide the labels during training without incorporating the loss of these un-annotated regions. As this incorporates the loss from the annotated parts only, it can introduce a class imbalance, which is catered by using the weights for each class calculated during training epochs. Higher weight is assigned to the class with less number of pixels, and less weight is assigned to the class with more pixels using the count per class as $P = \frac{\sum_{i=1}^{C} r_i}{C^{'}}$ where P is the expected count and r_i is the annotated region for i^{th} class, C is the total number of classes in the dataset and $C^{'}$ is the number of classes with regions annotated. Final weights W^i are calculated from expected count P as $W^i = \frac{P}{r^i}$ where W^i is the weight for r^i, it will be greater than one if the number of pixels in r^i region is less than the expected count and vice versa. Finally, the weighted sparse cross entropy (WSCE) loss is calculated as.

$$\mathcal{L}^{WSCE} = -\sum_{x \in X} p(x) \log q(x) \odot W \tag{5.1}$$

OED Grading and Malignant Transformation

Once we have the epithelium segmented into sub layers the next task is to predict the grade and malignant transformation from WSIs. Figure 5.3 shows the overall framework of the proposed pipeline for OED diagnosis and prognosis. OED diagnosis and prognosis can be modelled as multiple instance learning (MIL) problem, where a bag of instances is assigned a positive label if it contains at least one positive instance; otherwise is assigned a negative label. We modelled OED diagnosis and prognosis as individual MIL problem where a single MIL model is trained to predict the binary grade of OED, i.e., low-risk and high-risk. At the same time, another MIL model is trained to predict malignant transformation from OED with the help of ranking based loss. For our work, we used graph representation of a WSI as a bag and trained a graph neural network where $B = \left\{b^1, b^2, b^3 \dots \ b^t\right\}$ and $Z = \left\{z^1, z^2, z^3 \dots \ z^t\right\}$ represents t WSIs and their associated labels where $t = \{1, 2, 3, ...T\}$ and T is total number of WSIs. A graph representation is built using the $G^i = G\left(b^i\right)$ for each b in B, and the graph neural network is trained with trainable parameters θ to generate slide level prediction as $F\left(G\left(b^i\right); \theta\right)$. Finally, trained models were used for inference at test time for diagnosis and prognosis. The whole process can be broken down into the following steps, i) feature extraction, ii) graph construction, iii) graph neural network, iv) training, v) hotspot analysis and vi) survival analysis.

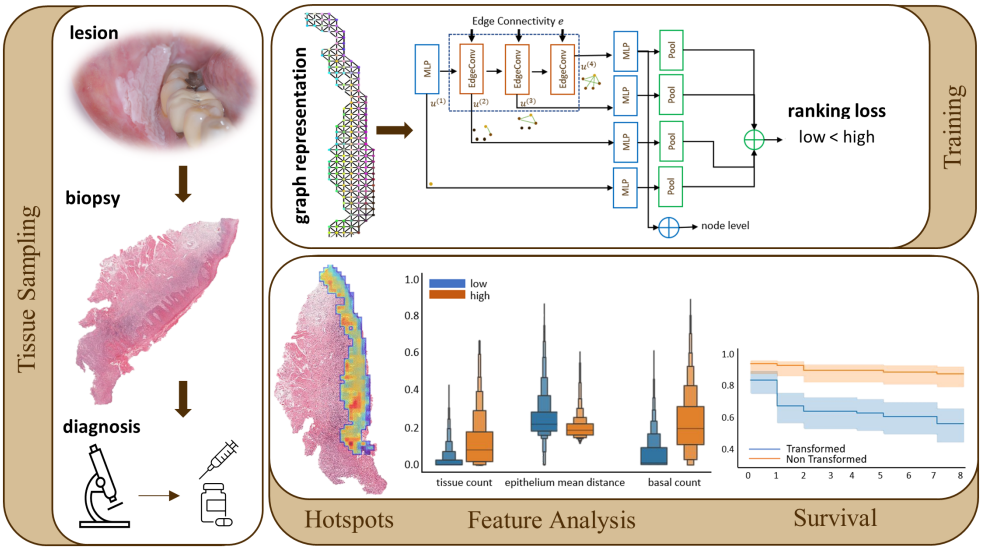


Figure 5.3: OED diagnosis and prognosis pipeline. Traditionally, a digitised biopsy of dysplastic lesions is analysed by the pathologist for grade prediction and treatment decisions. On the other hand, our pipeline creates a graph representation of the WSI for training a graph neural network using ranking loss for diagnosis, i.e., OED grade prediction and prognosis, i.e., malignant transformation. Hotspot analysis revealed that the nuclei count in tissue area and basal layer along with crowdedness of nuclei in the epithelium and peri-epithelium tissue area were found to be significant nuclear features for differentiating between the low-risk and high-risk cases along with the progression free survival in OED cases.

Feature Extraction

We extracted deep and nuclear features from each WSI in the form of patches for graph construction. Patches of size 512×512 were extracted using the epithelium masks generated automatically using coarse segmentation with an overlap of 50% from all the WSIs at 0.50µm per pixel (mpp). For extracting the deep features, ResNet-50 [135] was used as a feature extractor pre-trained on ImageNet. A feature vector of size 1024 was extracted for each patch resulting in a bag of shape $h \in \mathbb{R}^{n \times 1024}$ for all WSIs (where n is the number of patches extracted) along with their top-left corner geometric coordinates as $g \in \mathbb{R}^{n \times 2}$. Afterwards, the patches were first stain normalised for extracting nuclear features using a sample from the training cohort before being fed into HoVer-Net+ [100] for nuclear instance segmentation. Twenty-four morphological and spatial features (i.e., eccentricity, convex area, contour area, extent, perimeter, solidity, radius, major/minor axis, equivalent diameter, nearest neighbours distance (NNC) etc.) were extracted form nuclei in each layer per patch. They were aggregated statistically using the mean μ, min $\wedge$, max $\vee$ and standard deviation σ. This resulted in a bag of shape $h \in \mathbb{R}^{n \times 384}$ (i.e., 24 features $\times$ 4

layers × 4 aggregations) for all WSIs along with their top-left corner geometric coordinates as $g \in \mathbb{R}^{n\times 2}$.

Graph Construction

After extracting deep and nuclear features, we constructed three types of graph representations, i.e., deep graphs, nuclear graphs and fusion (deep + nuclear) graphs. Nodes were represented as $v^j = \left(g^j, h^j\right)$ where h and g are the node features and their geometric coordinates. A graph representation of a WSI consists of a set of vertices V with each patch as a node and their edges as E where $G = (V, E)$. The set of edges, denoted as $E \in V \times V$, captures the interaction and interconnectedness of the nodes. Delaunay's triangulation [211] establishes the edge set based on the geometric coordinates of patches to effectively represent communication patterns among tissue components. This process involves setting a maximum distance connectivity threshold of $d^{max} = \{500, 1000, 3000, 50000, 7000, 10000\}$ pixels to ensure that the graph resulting from the triangulation is planar, which means that no two edges in the graph intersect each as seen in Figure 5.4.

Graph Neural Network (GNN)

We built a graph neural network inspired from SlideGraph$^\infty$ [212] where both graph level and node level predictions from graph representations can be generated from the input as seen in Figure 5.5. The graph neural network (GNN) consists of multiple EdgeConv layers along with their MLP layers, as this configuration helps in the generation of a feature embedding of a graph node by differentiating its features from the neighbour node's features. The initial layer uses the original node level features, while subsequent layers use the node embedding generated using MLP, which allows the GNN to accumulate information from increasingly higher order neighbours of each node, leading to progressively more abstract feature representations. Mathematically, the output feature representation of an EdgeConv layer for a given node can be written as a function of the node's index and the layer index within the GNN architecture as $u_s^o = \sum_{j\in N^o} H^{(s)}\left(u_{(s-1)}^o, u_{(s-1)}^j - u_{(s-1)}^o; \theta^s\right)$ where $s = 1 \ldots S, u_{(0)}^o = u^o, N^o$ represents the oth node neighbourhood while $H^{(s)}$ denotes the MLP with trainable parameters θ^s. u_s^o represents the feature embedding of node k where $v^o = (g^o, u^o) \in$ V passes through the EdgeConv and corresponding MLP for node level features as $f^s\left(v^o\right) = w_T^s u_s^o$. In order to generate node level scores, these node level features were aggregated using either sum, min and max. Whereas for generating WSI level score, the node scores were further pooled to generate layer wise WSI score as $F^s\left(G\right) = \Sigma_{\forall v\in V} f^s$. Finally, these scores were summed to generate the final WSI score as $F\left(G;\theta\right) = \Sigma_{s=0}^S F^s\left(G\right)$ where

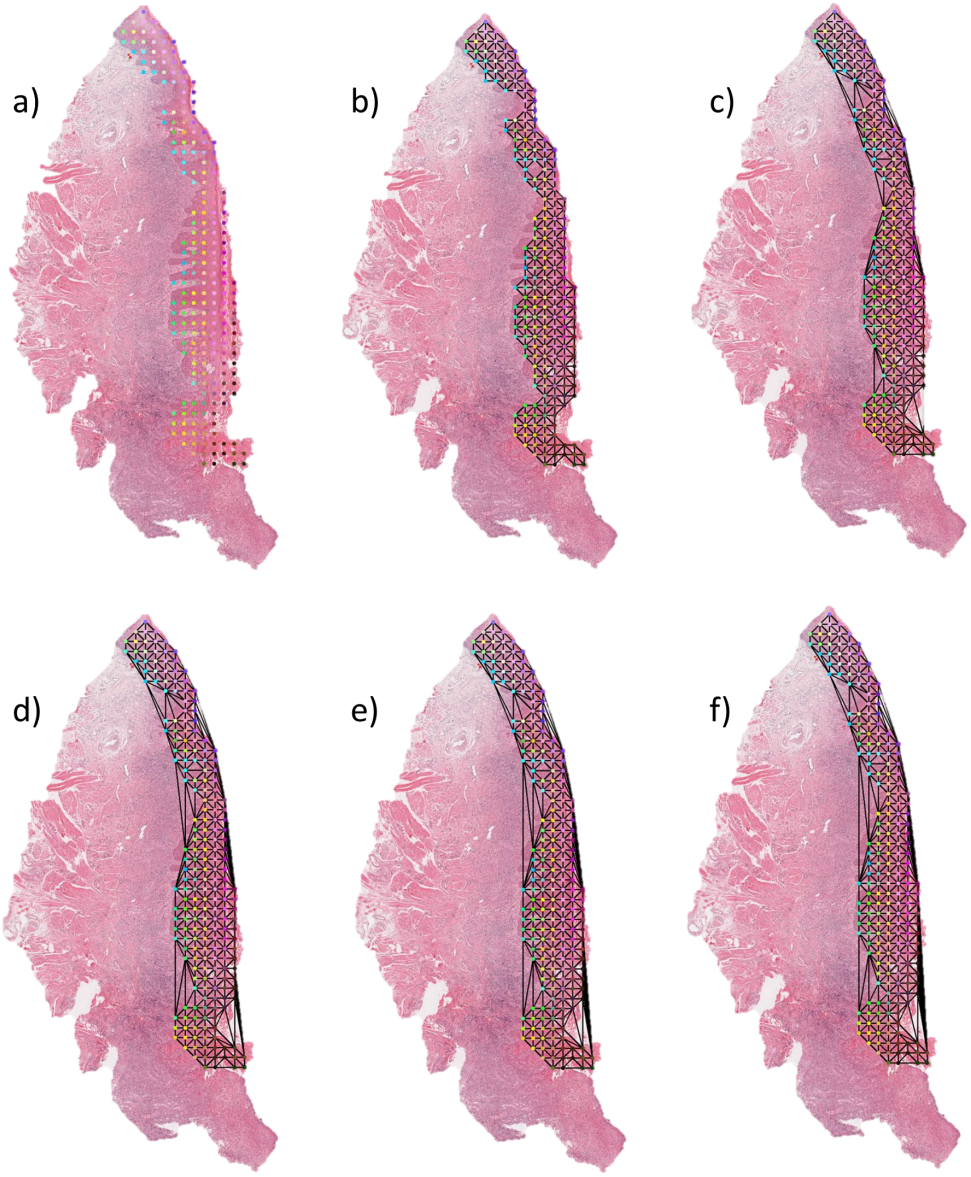


Figure 5.4: Graph construction using different thresholds d^{max} in Delaunay's Triangulation. a) shows the graph construction with no edges between the patches due to a small threshold of $d^{max} = 500$ pixels. b) shows the graph construction with edges between immediate neighbours with $d^{max} = 1000$ pixels. Similarly, c), d), e) and f) show the graph construction with $d^{max} = 3000, 5000$ and $10,000$ pixels where it can be seen by the black lines representing edges connecting distant nodes as we increase the threshold.

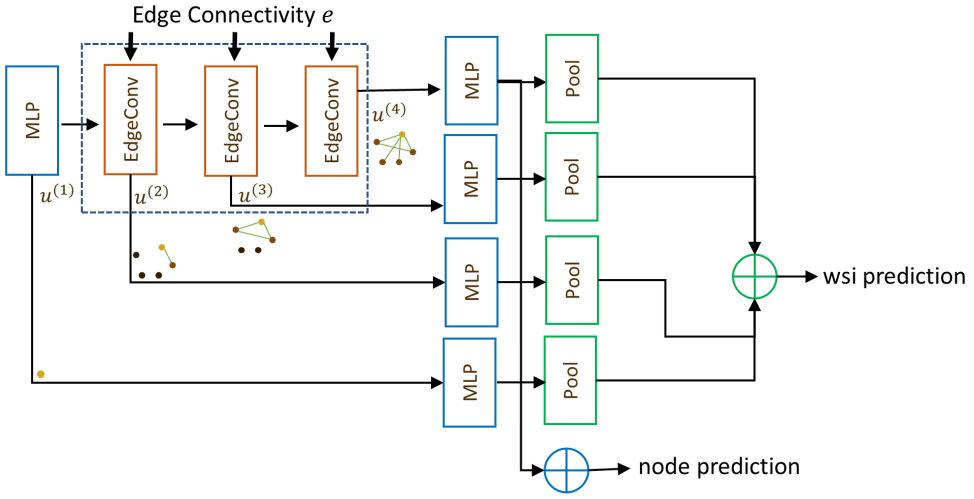


Figure 5.5: GNN architecture for graph based node and WSI prediction composed of EdgeConv subsequent MLP layers. Node level predictions were generated by aggregating the output of MLP layers, while for WSI level prediction, MLP output is first pooled and then aggregated.

θ represents all trainable parameters in the GNN.

Hotspot Analysis

Once the models are trained for diagnostic and prognostic purposes, their node level prediction can be used to identify hotspots and coldspots. In the case of OED grading, the coldspots would correspond to low-risk areas, while hotspots would correspond to high-risk areas. Whereas for OED malignant transformation prediction, the coldspots would correspond to the areas with a low chance of transformation, while hotspots would represent the higher chance of malignant transformation. In this regard, we extract the top 15% of the patches from the true positive and true negative cases to analyse their unique differentiating signal. After extracting the top 15% patches, their nuclear features extracted earlier were re-used to find out the important features, i.e., unique differentiating signals in coldspots and hotspots. The statistical significance of the nuclear features was calculated using ordinary least squares (OLS) with post-hoc t-tests and Benjamini/Hochberg [2] adjustment. The cellular composition was analysed to better understand the results obtained from GNN and to objectively differentiate between low-risk, high-risk, transformed, and non-transformed cases.

Survival Analysis

To determine the prognostic significance of clinical, pathological, GNN diagnostic/prognostic score, and nuclear features with respect to progression-free

survival (PFS), univariate and multivariate analyses were conducted using Kaplan-Meier (KM) curves and Cox proportional hazard (CPH) model. To differentiate between the high-risk (short-term survival) and low-risk (long-term survival) groups, the optimal cut-off value was computed by averaging the hazard value for each instance, considering the large statistical significance between the high- and low- risk groups in the CPH model. A log-rank test was also performed to assess the statistical significance, and a $p < 0.05$ was deemed statistically significant.

5.2.3 Training

To train our GNN, pairwise ranking loss was used where during training, positive and negative cases were chosen in a stratified manner in order to compare them with each other. The mathematical formulation of the loss is as follows $\mathcal{L}^{pair} = \Sigma_{i \in Batch} \Sigma_{j \in Batch} \max\left(0, 1 - \left(F\left(G^i; \theta\right) - F\left(G^j; \theta\right)\right)\right)$. The loss is backpropagated to adjust the weights of EdgeConv and MLP layers during training.

5.2.4 Experimental Settings

Coarse Segmentation

Input patches were pre-processed using standard pre-processing steps of normalisation and augmentation (i.e., random rotation [0, 90, 180, and 270 degrees], random clipping [horizontal, vertical], random jittering [0-128] pixels and random colour perturbation). Then the model is trained using RMSProp optimiser with an adaptive learning rate starting from 0.001 for a minimum of 100 epochs. Using a system equipped with two Nvidia Titan-X GPUs, each with 12GB of memory, 128GB of dedicated RAM, and an Intel® Core i9 processor. Python language with Tensorflow deep learning framework was used to develop this framework.

We conducted the following experiments to compare and validate the proposed coarse segmentation approach and used pixel-wise F1-score for evaluation purposes.

- DeepLab-v3 was trained for pixel-wise as well as for coarse segmentation using OED subset as pixel-wise annotations were only available for this dataset and were not available for the HNSCC dataset.
- To compare patch-based classification and coarse segmentation, various standard CNNs were trained for patch-based classification. The HNSCC dataset for tissue classification was selected for patch-based classification.

- To compare the inference time of patch-based classification and pixel-wise segmentation with coarse segmentation, an average sized WSI was processed with all methods and the final time was reported.

GNN Diagnosis and Prognosis

For GNN, we used three EdgeConv layers where the first layer had the same neurons as input data (i.e., nuclear = 388, deep = 512, fusion = 900), second and third layers had 1024 neurons in their respective MLPs. Each layer is followed by batch normalisation (BN) and exponential linear unit (ELU) as activation layer. GNN is optimised using adaptive momentum based optimisation [213] (Adam) with a learning rate of 0.0001 and a weight decay of 0.0001. The model was trained for 300 epochs with a batch size of 96, using a system equipped with two Nvidia Titan-X GPUs, each with 12GB of memory, 128GB of dedicated RAM, and an Intel® Core i9 processor. In order to compare the performance of GNN with other frameworks, we used IDaRS and CLAM. Iterative draw-and-rank sampling (IDaRS) [54] employs a ranking strategy to select top and random patches from a whole slide image (WSI) based on their predictive importance. This approach acknowledges that not all patches contribute equally to the final outcome. For each WSI, IDaRS selects two sets of patches for training, which include a random set of patches and a set of top-ranked patches. Both subsets undergo standard augmentations and are used to train a CNN with weak labels. Clustering constrained attention multiple instance learning (CLAM) [104] was designed to overcome the challenges of domain adaptation, interpretability, and visualisation by using the attention-based sub regions of highly diagnostic values for the classification of WSIs. CLAM initially extracts latent feature vectors from each patch using a ResNet-50 CNN encoder and then uses those feature vectors instead of image patches for further computation. This saves a lot of computation and time for the model's training. For this chapter, we used GNN to predict two different clinical variables/outcomes.

- Low-risk vs High-risk OED (Diagnostic)
- Malignant Transformation vs No Transformation (Prognostic)

Stratified 5-fold cross-validation was performed three times with different random seeds to validate the results. Patch-wise AUROC and F1-score (macro) were used as performance metrics at WSI level and are averaged across the folds.

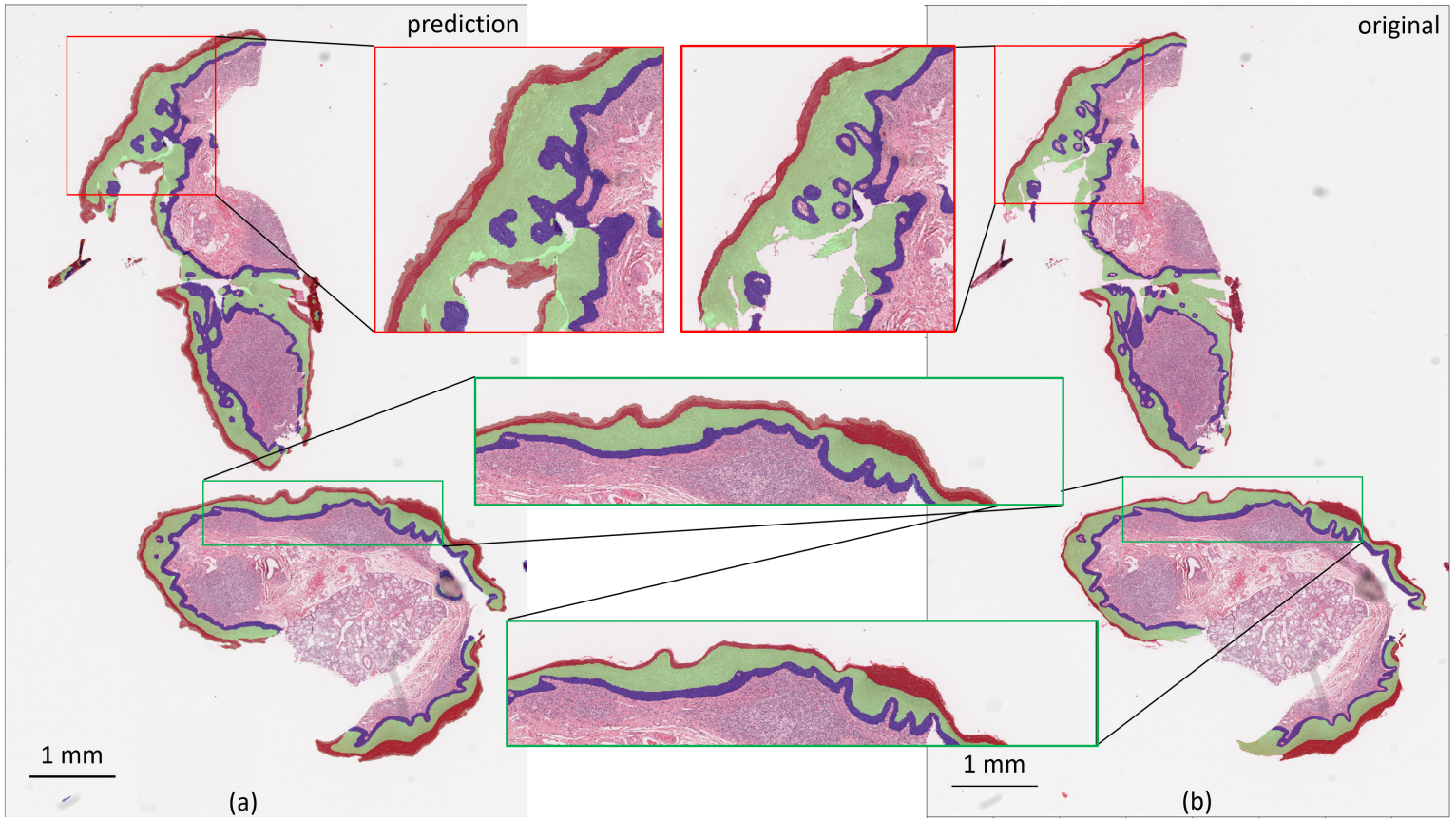


Figure 5.6: **a)** shows the prediction overlay of coarse segmentation using our proposed CSNet model with mini-patch of size $k = 32$ as compared to **b)** which is pixel-wise ground truth overlaid on WSI. Red boxes show some of the areas of false predictions while the green boxes show some of the true predictions areas in the WSI

Table 5.4: F1-score for Patch-based classification and Coarse segmentation in HNSCC for patch size of 256×256

Method	k	Stride	F1-score
ResNet-50	$k = 32$	32	73.23
MobileNet	$k = 32$	32	74.78
DensetNet	$k = 32$	32	78.76
CSNet	$k = 32$	256	**83.11**

5.3 Results and Discussion

5.3.1 Pixel-wise vs Coarse Segmentation

Table 5.5 compares the results of coarse segmentation and pixel-wise segmentation methods using DeepLab-v3 [114] and CSNet on various mini-patch sizes k. For a fair comparison, we compared the methods with the same mini-patch, e.g., it can be seen that our method with mini-patch sizes of 16 and 32 performed superior to DeepLab-v3, which shows that for coarse segmentation, we can use CSNet like methods rather than using the pixel-wise segmentation approaches. WSI level masks were generated using the sliding window approach with an overlap of 80% and resizing the output size to input using linear interpolation as it's already a coarse output. The result is shown in Figure 5.6, where it shows the prediction of our proposed CSNet with and mini-patch of $k = 32$.

Table 5.5: F1-score for pixel-wise and coarse segmentation in OED layer segmentation for a patch size of 512×512

Method	mini-patch k	F1-score
DeepLab-v3	$k = 1$	78.82
DeepLab-v3	$k = 16$	80.57
CSNet	$k = 16$	**81.24**
DeepLab-v3	$k = 32$	78.32
CSNet	$k = 32$	**80.54**

5.3.2 Patch-based Classification vs Coarse Segmentation

Table 5.4 compares the results of patch-based classification and coarse segmentation where for patch-based classification, simple standard CNNs (i.e., ResNet[135], MobileNet [214] and DenseNet [206]) were used. To make the CNN's output comparable with coarse segmentation results, the stride was set to 32, and only 32×32 region was assigned a label from the CNN's output, as shown in Figure 5.7, because assigning a single label to 256×256 will result in very low accuracy. It can be seen that CSNet performs superior by 5-10% as compared to standard CNNs due to the additional skip connections added to it as seen in Figure 5.1. Moreover, if we lower the patch size of standard CNN,

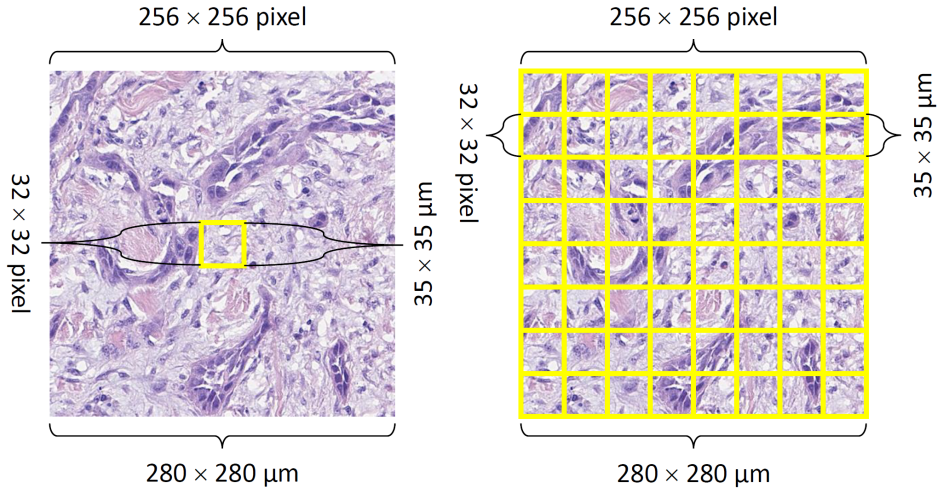


Figure 5.7: Yellow boxes show the region to which label is assigned in a 32×32 window, where the left one shows the output label to be assigned using standard CNN while the right one shows the output to be assigned from coarse segmentation.

the accuracy further drops below the current one due to the lack of context in that patch as we experimented with a smaller patch size of 128×128, and the F1-score achieved was 66.10 which is lower than the previous one. Figure 5.8 shows the prediction of our proposed method, where it can be seen that the CSNet model performs better in most of the tissue regions except for some highlighted in black circles.

5.3.3 Inference time comparison

Simple performance based comparisons are not enough to show that our proposed CSNet is much better than the simple patch-based classification and pixel-wise segmentation until we compare the inference time for these methods. To compare the inference time for these approaches, we processed an average sized WSI and calculated the total time in minutes as shown in Table 5.6. It can be seen that the simple CNN architecture DenseNet took 20 hours to process a WSI because it has to assign a single label to 32×32 each window, which increases the time required to complete the WSI as compared to pixel-wise and the proposed CSNet where it took only 21 minutes and is $60\times$ faster than the normal patch-based classification and $1.35\times$ faster than the pixel-wise segmentation.

5.3.4 Network Variations

Table 5.7 shows our network variations and their performance on the patch-based classification of the HNSCC dataset, as we have used DenseNet-121 for

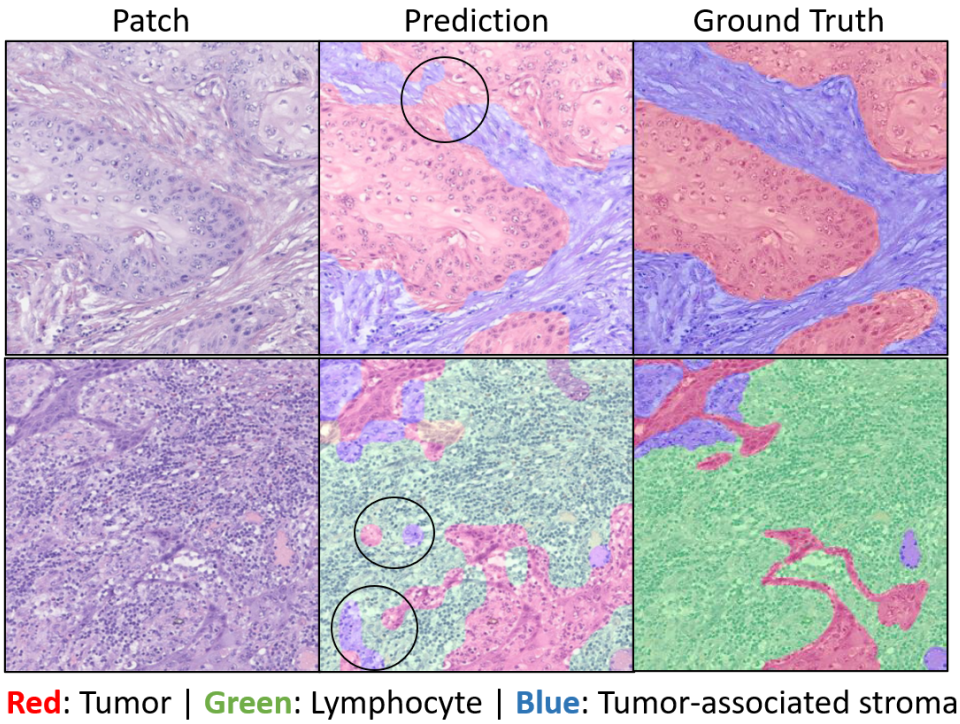


Figure 5.8: Overlay of two visual fields from HNSCC internal data for coarse segmentation where ground truth is smoothed before overlaying for display. It can be seen that most of the tissue regions are being segmented correctly, with some false predictions highlighted in black circles.

the baseline of coarse segmentation network, which is further modified with additional skip connections between the dense blocks. The intuition was to increase the spatial context in the final output maps, which in return, increases the overall accuracy of the network. As it can be seen that with the additional skip connections (SC), the F1-score of the model was improved by almost 4 points margin, and to further justify the addition of SC, the network size was reduced to half as DenseNet-61 was used as a baseline with skip connections and it can be seen that the smaller model still performs better than the larger DenseNet-121 by a margin of one point.

5.3.5 Mini-Patch Variation

Table 5.5 shows the variations in mini-patch k in our proposed CSNet, where it can be seen that using bigger k doesn't affect the performance drastically as the difference is only 1 point. Also, using a smaller mini-patch increases the performance due to the fact that there is less noise and more precise ground truth. Whereas in the pixel-wise segmentation model, DeepLab-v3, the accuracy dropped by 2 points in using a bigger mini-patch as compared to CSNet segmentation which shows that pixel-wise methods are not suitable for this coarse segmentation method.

Table 5.6: Inference time comparisons for different segmentation methods for processing one WSI

Method	Patch Size	k	Stride	Prediction	Time (min)
DenseNet	256×256	256	32	1×1	1208.28
DeepLab-v3	512×512	1	512	512×512	27.47
CSNet	256×256	8	256	32×32	**20.98**

Table 5.7: Performance comparison of different network variants for coarse segmentation

Network	Variation	F1-score
CSNet-121	Standard	79.28
CSNet-121	Skip Connections	**83.11**
CSNet-61	Skip Connections	80.56

5.3.6 OED Grade and Malignant Transformation Prediction

OED lesions have the potential to transform into malignancy, i.e., oral squamous cell carcinoma (OSCC). There are no effective tools to predict the likelihood of such transformation with confidence. Therefore, predicting OED grade along with the malignancy is critical to assessing the malignant potential of dysplastic lesions as it can help in appropriate treatment plans, e.g., surgical excision or close monitoring. In this regard, we train a GNN for predicting the binary grade of OED and another for predicting malignant transformation in OED. Figure 5.9 and Table 5.8 show the performance of GNN as compared to recent MIL models in OED grade prediction, where our proposed GNN trained with deep feature graphs surpasses the IDaRS's AUROC of 0.76 ($\pm$0.06 SD) and CLAM's AUROC of 0.68 ($\pm$0.06 SD) by achieving an AUROC of 0.81 ($\pm$0.05 SD) and F1-score of 0.74 ($\pm$0.06 SD). Similarly, we can see that our proposed GNN trained with fusion feature graphs outperforms IDaRS by 2% and CLAM by a large margin for predicting OED malignant transformation while keeping the standard deviation (SD) lower than the others. One of the main advantages of using GNN over IDaRS and CLAM is the explainability of the graphs and graph based features. Apart from node level prediction scores, GNN outputs feature importance scores for each feature which can be used to explain a particular decision. Also, IDaRS takes a lot of time to train, while graph based models are quite efficient in terms of training time due to pre-built graphs being used for training purposes. See Figure 5.10 for GNN performance with different graph feature types and connectivity thresholds where a) we can see that for OED grading, the GNN trained using graphs built on deep features performed better than the nuclear and fusion based graphs. Whereas

for transformation b) nuclear and fusion based graphs performed better in terms of AUROC.

Model	Task	AUC ± SD	F1-score ± SD
Attention-MIL	OED Grading	0.56 ± 0.08	0.43 ± 0.02
CLAM		0.68 ± 0.06	0.60 ± 0.04
IDaRS		0.76 ± 0.06	0.72 ± 0.06
Proposed GNN		0.81 ± 0.06	0.74 ± 0.06
Attention-MIL	OED Malignant Transformation	0.57 ± 0.06	0.45 ± 0.03
CLAM		0.65 ± 0.04	0.64 ± 0.02
IDaRS		0.74 ± 0.11	0.67 ± 0.08
Proposed GNN		0.76 ± 0.06	0.67 ± 0.07

Table 5.8: Performance of GNN model as compared to other weakly supervised where GNN achieves high performance in both the tasks of grading and malignant prediction in terms of AUROC and F1-score on a 5-fold cross-validation bootstrapped three times with random seeds.

5.3.7 Cellular Composition Analysis

Heatmaps were used to investigate and interpret the potential differentiable hidden signatures within the WSIs. In this regard, we used the node prediction from our diagnostic and prognostic models and built the heatmaps as seen in Figure 5.11. It can be seen in a) high-risk case is being identified by the diagnostic GNN trained to predict the binary grading and b) shows the same case processed by the prognostic GNN for predicting malignant transformation, and we can see in the hotpots the common features of high-risk OED cases, i.e., nuclear pleomorphism, dyskeratosis and irregular epithelial stratification with a dense lymphocytic infiltrate in the adjacent peri-epithelial connective tissue. Similarly, c) shows an interesting case where the patient was identified as low-risk by pathologists but, later on, transformed into carcinoma, which is being identified by the prognostic GNN. Following the manual heatmap analysis, we further analysed the nuclear features for all the true positive and true negative cases to uncover some unique signals for diagnosis and prognosis. Table 5.9 shows the significant nuclear features extracted from the top 15% of the patches from diagnostic and prognostic GNN models. For OED grading, we can see that the basal nuclei count ($p = 0.006$) and epithelium nuclei count ($p = 0.006$) were the significant nuclear features. In contrast, variation in mean epithelium crowdedness (i.e., average distance of nearest ten nuclei) showed potential and requires larger cohort for validation.

For OED malignant transformation prediction, basal nuclei count ($p = 0.00001$) and tissue area nuclei count showed the most significance ($p = 0.0001$). Mean solidity representing the shape of nuclei also showed significance in the basal layer ($p = 0.0001$) along with the mean basal crowdedness ($p = 0.001$). In

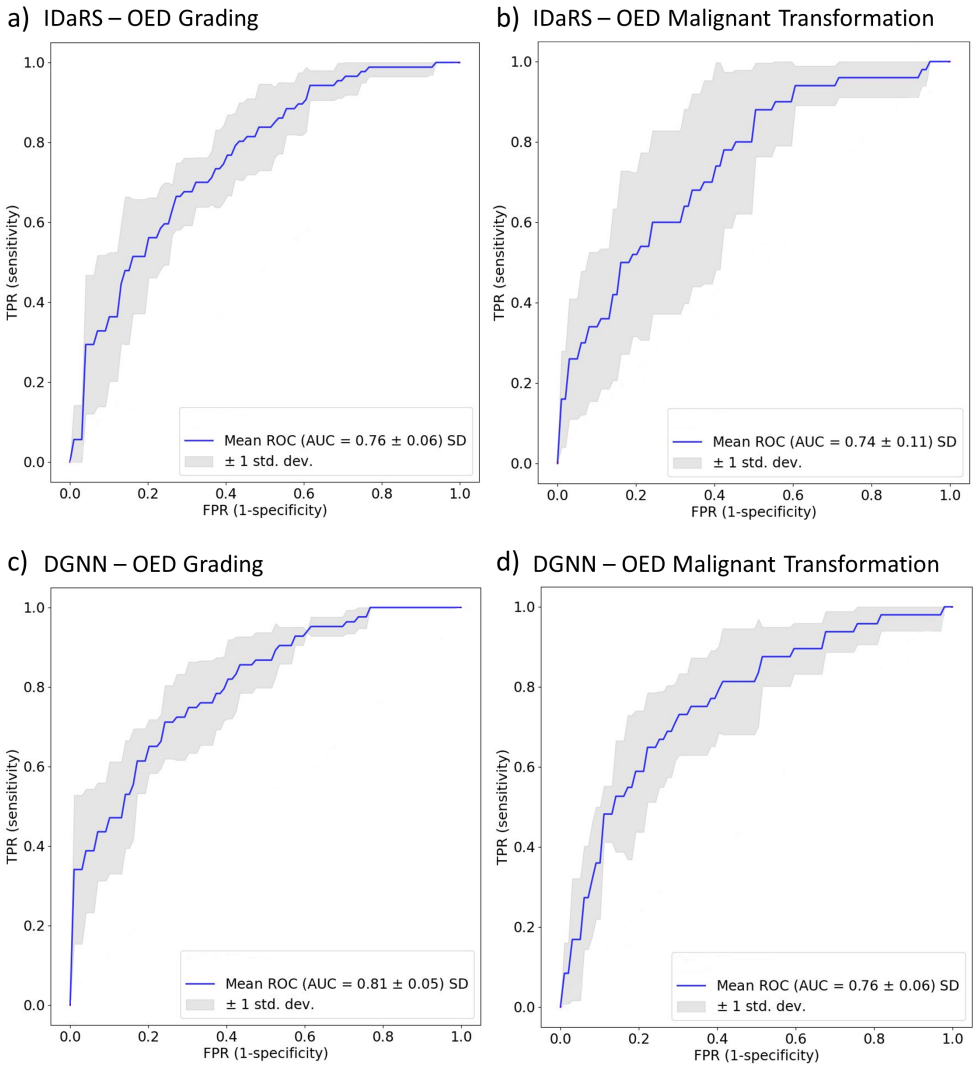


Figure 5.9: AUROC curves plots on 5-fold cross-validation for OED grading and malignant transformation for top two performing MIL methods, i.e., IDaRS (a, b) and GNN (c, d).

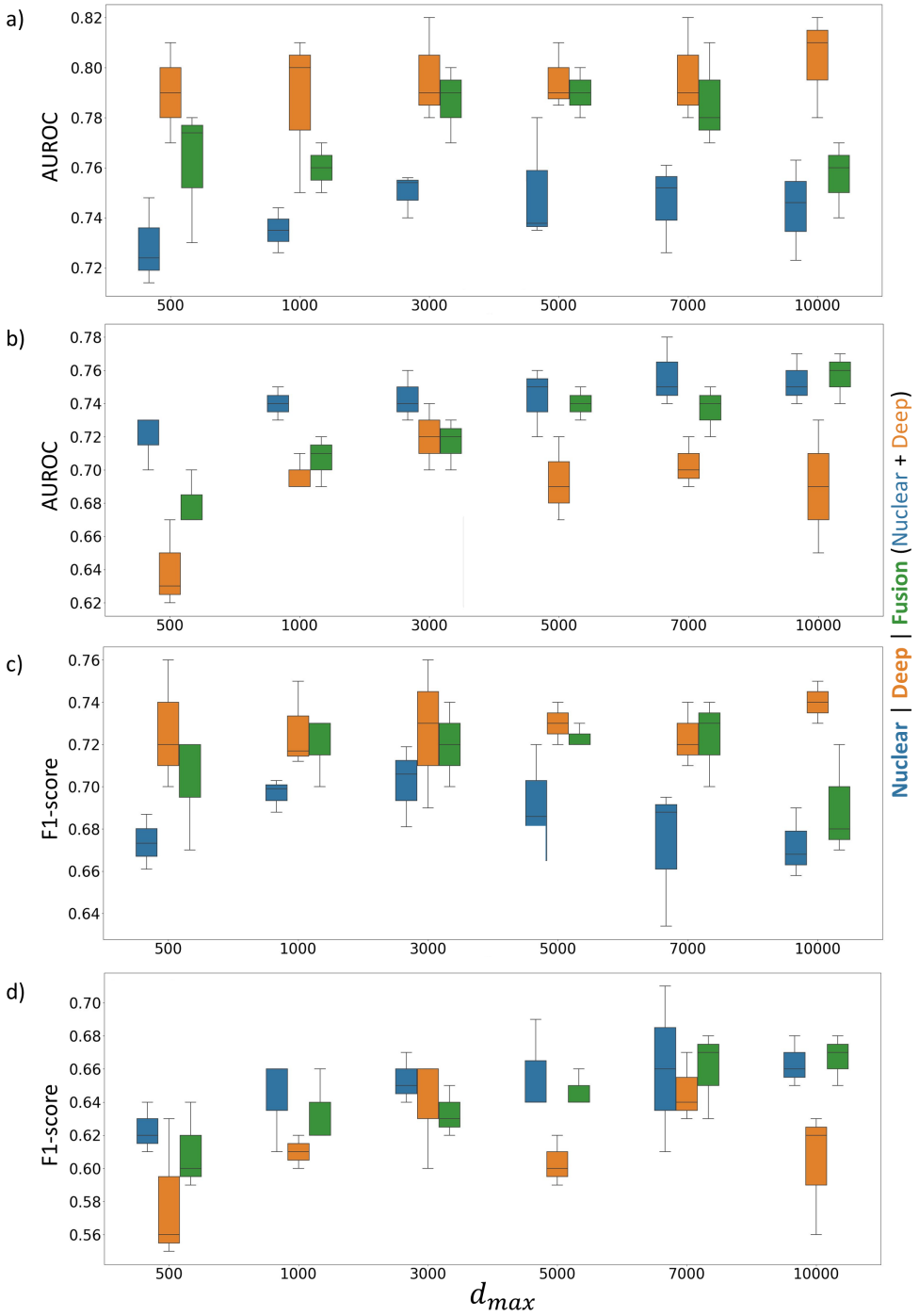


Figure 5.10: Performance of GNN with different graph features and connectivity. a) AUROC of GNN for binary grading, b) AURCO for OED malignant transformation, c) and d) show the F1-score for OED grading and transformation, respectively.

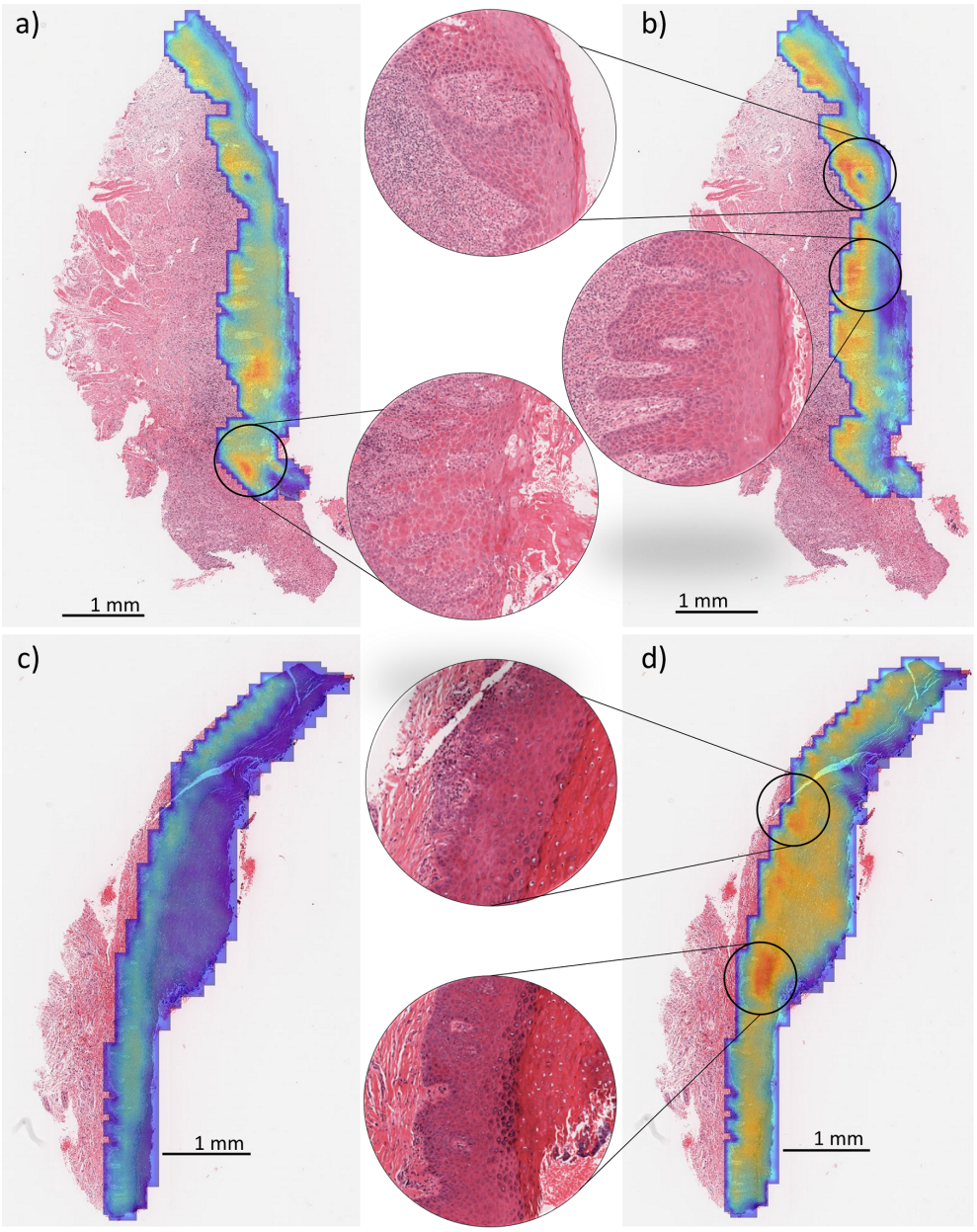


Figure 5.11: Hotspots identified by the GNN models represented as heatmaps for both OED grading and malignant prediction. a) Shows the heatmap for OED grading for a high-risk transformed case along with b) the heatmap for OED malignant prediction from GNN. It can be seen from the highlighted areas that irregular stratification of epithelium, bulbous rete ridges and peri-epithelial lymphocytes are being highlighted by the models. c) Shows the heatmap of OED grading for low-risk transformed case where the models were able to correctly identify it as low-risk transformed case. d) he heatmap for OED malignant prediction from GNN, nuclear pleomorphism can be seen from the highlighted areas, along with the start of irregular epithelium stratification.

high-grade OED, the changes are more visible in the basal and epithelial layer, and it can be seen from the significant features that hyperplasia or crowding has been picked up by the GNN. On the other hand, it can be seen that the along with the changes in basal layer density, peri-epithelial tissue cells showed significance, as it has been reported earlier in the studies [83]. Significant nuclear feature distribution for OED grading and malignant transformation can be seen in Figure 5.12. Where a rise in nuclei count for epithelial and basal layer can be seen for OED grading while rise in basal layer and tissue area nuclei count is evident in malignant transformation.

Feature	**Diagnostic**	**Prognostic**
Tissue area nuclei count	×	✓
Basal layer nuclei count	✓	✓
Epithelium layer nuclei count	✓	×
Mean epithelial layer crowdedness	✓	×
Mean basal nuclei solidity	×	✓
Mean basal layer crowdedness	×	✓

Table 5.9: Ordinary least square regression for malignant transformation with t-test significance of nuclear features with Benjamini/Hochberg [2] adjustment. Significant p-value is highlighted using ✓.

5.3.8 Survival Analysis

Table 5.10 shows the univariate analysis of the clinical, pathological, GNN diagnostic/prognostic score and nuclear features with respect to progression-free survival (PFS). Starting from the clinical attributes, namely gender and age are nonsignificant with age ($p > 0.05$, C-index = 0.59 [95%, 0.59 – 0.60]) and gender ($p > 0.05$, C-index = 0.52 [95%, 0.52 – 0.53]). The pathological features were composed of pathologists' grading, where we used binary and WHO based grades as surrogate labels for transformation. Binary grading showed that high-risk cases are more likely to transform with significance ($p = 0.004$, C-index = 0.68 [95%, 0.67 – 0.69]). Whereas for WHO based grading, we combined first (mild + moderate) against severe and then we combined (moderate + severe) against mild for PFS, and in both combinations, we got the significance ($p = 0.04$, C-index = 0.68 [95%, 0.67 – 0.68]). Our GNN model scores for each WSI were taken as GNN diagnostic and prognostic scores for predicting PFS. Where using the GNN score for predicting OED grades as a surrogate label for transformation resulted in almost the same as the original binary grading C-index with a little improvement ($p < 0.05$, C-index = 0.69 [95%, 0.68 – 0.70]). Similarly, the GNN score for predicting transformation was taken as it is for PFS, and it performed better than the binary grading with ($p < 0.05$, C-index = 0.70 [95%, 0.69 – 0.71]). Figure 5.13

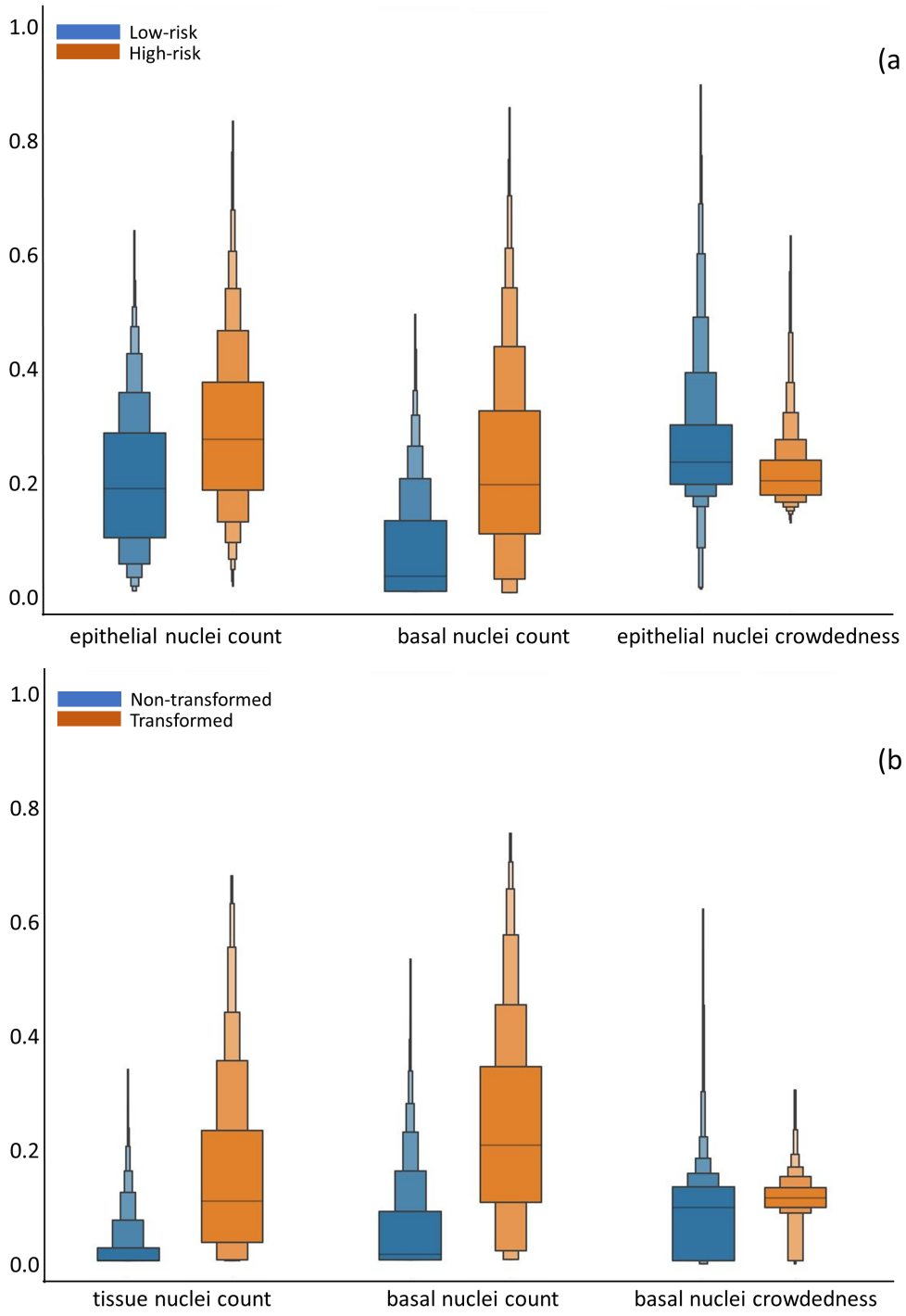


Figure 5.12: a) Boxen plot for most significant nuclear features for OED grading where it can be seen that the nuclei count in high-risk OED is higher than the low-risk. The lower value of crowdedness, the higher the density in a region, meaning the nuclei are coming closer to each other in high-risk cases. b) Boxen plot for most significant nuclear features for malignant prediction where the nuclei count in transformed cases is greater than the non-transformed ones for basal layer and tissue area.

shows the Kaplan-Meier curves plotted using cox-proportional hazard ratios with mean as a threshold between transformed and non-transformed curves. a) shows the diagnostic score and b) the prognostic score, and we can see that both scores can differentiate between the transformed and non-transformed cases with a clear separation. Next, we used the top three nuclear features, i.e., basal layer nuclei count found in the hotspots, to differentiate between transformed and non-transformed cases. To aggregate patch level features to WSI level score, they were aggregated using mean μ standard deviation σ and median m. μ basal layer nuclei count showed the highest significance with ($p < 0.05$, C-index = 0.81 [95%, 0.80 – 0.81]), while the other aggregations were competitive. Epithelial layer count based features performed comparable to or worse than the other nuclear features and GNN diagnostic and prognostic score with ($p < 0.05$, C-index = 0.70 [95%, 0.69 – 0.70]). Finally, the tissue area nuclei count performed better than the basal layer nuclei count in two of the aggregation methods while achieving ($p < 0.05$, C-index = 0.83 [95%, 0.82 – 0.84]) as the highest score with μ in tissue area nuclei in transformed and non-transformed cases. Figure 5.13 c) shows the Kaplan-Meier curves plotted using the basal layer nuclei count, and d) shows tissue area nuclei count, and we can see that both the scores can differentiate between the transformed and non-transformed cases with a clear separation. In order to find the significance in a multivariate setting, we performed the multivariate analysis using the same features. Table 5.11 shows the multivariate analysis of the clinical, pathological, GNN diagnostic/prognostic score, and significant nuclear features with respect to progression-free survival (PFS) where the combined C-index is 0.85 with confidence interval [95%, 0.84 – 0.86]. Figure 5.14 shows the multivariate forest plot showing log hazard ratio (HR) with 95% confidence interval of aforementioned features where tissue area nuclei count turned out to be the most significant PFS feature.

In this chapter, we investigated the potential of graph neural networks (GNN) for diagnostic and prognostic purposes. For the diagnostic task, we predicted the binary grading of OED, i.e., distinguishing between low-risk and high-risk. Whereas for the prognostic task, we predicted the OED malignant transformation status in the digitised oral epithelial dysplasia (OED) histology slides. We developed a weakly supervised learning framework for both tasks and trained it using the ranking loss for optimisation. We identified the most predictive areas within the epithelial and peri-epithelial tissue regions for the tasks and then compared their cellular compositions to find significant nuclear features in both distributions. Our results showed that the GNN models could predict OED grades with an AUROC of 0.81 and malignant transformation with an AUROC of 0.76, as determined by a stratified 5-fold cross-validation bootstrapped using three different random seeds. The higher performance of

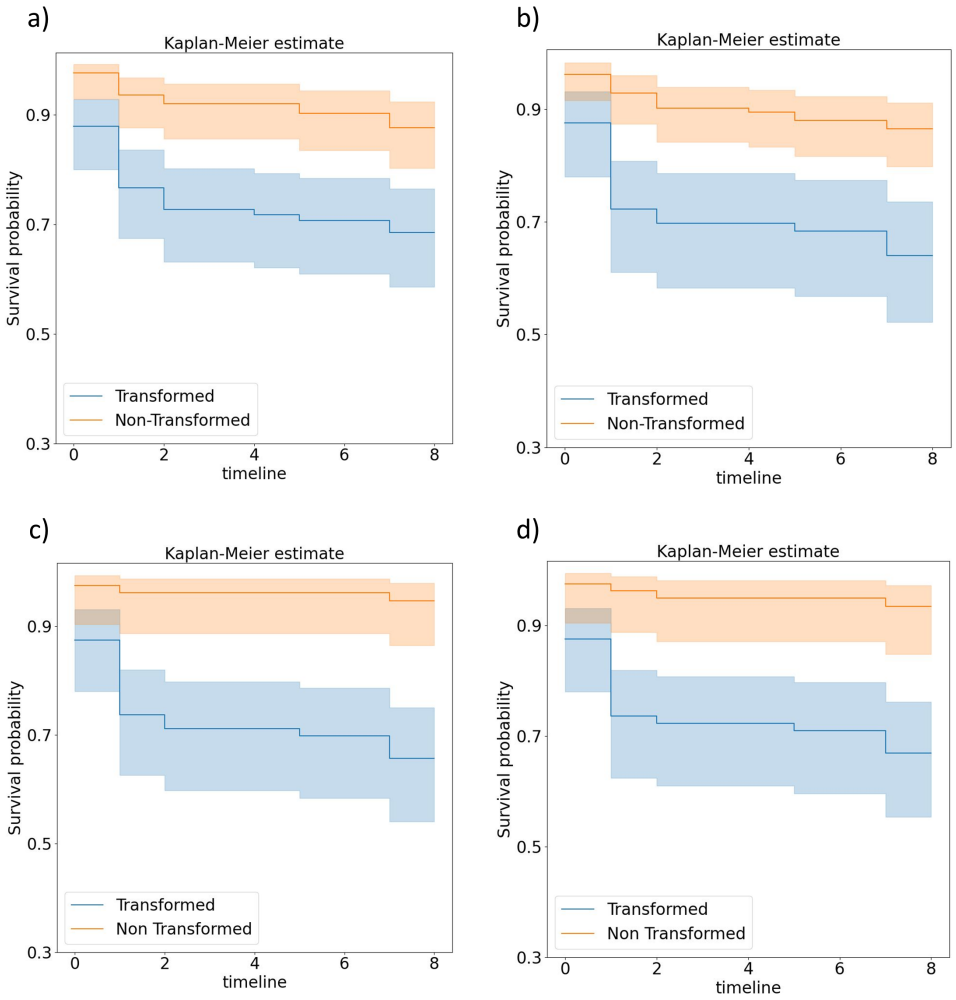


Figure 5.13: Kaplan-Meier curves plotted for progression free survival using the Cox proportional hazard ratios with mean as cut-off value. a) OED grade score, b) Malignant transformation score, d) Basal nuclei count in top 15% of transformed and non-transformed cases, d) tissue area nuclei count for top 15% patches in transformed and non-transformed cases.

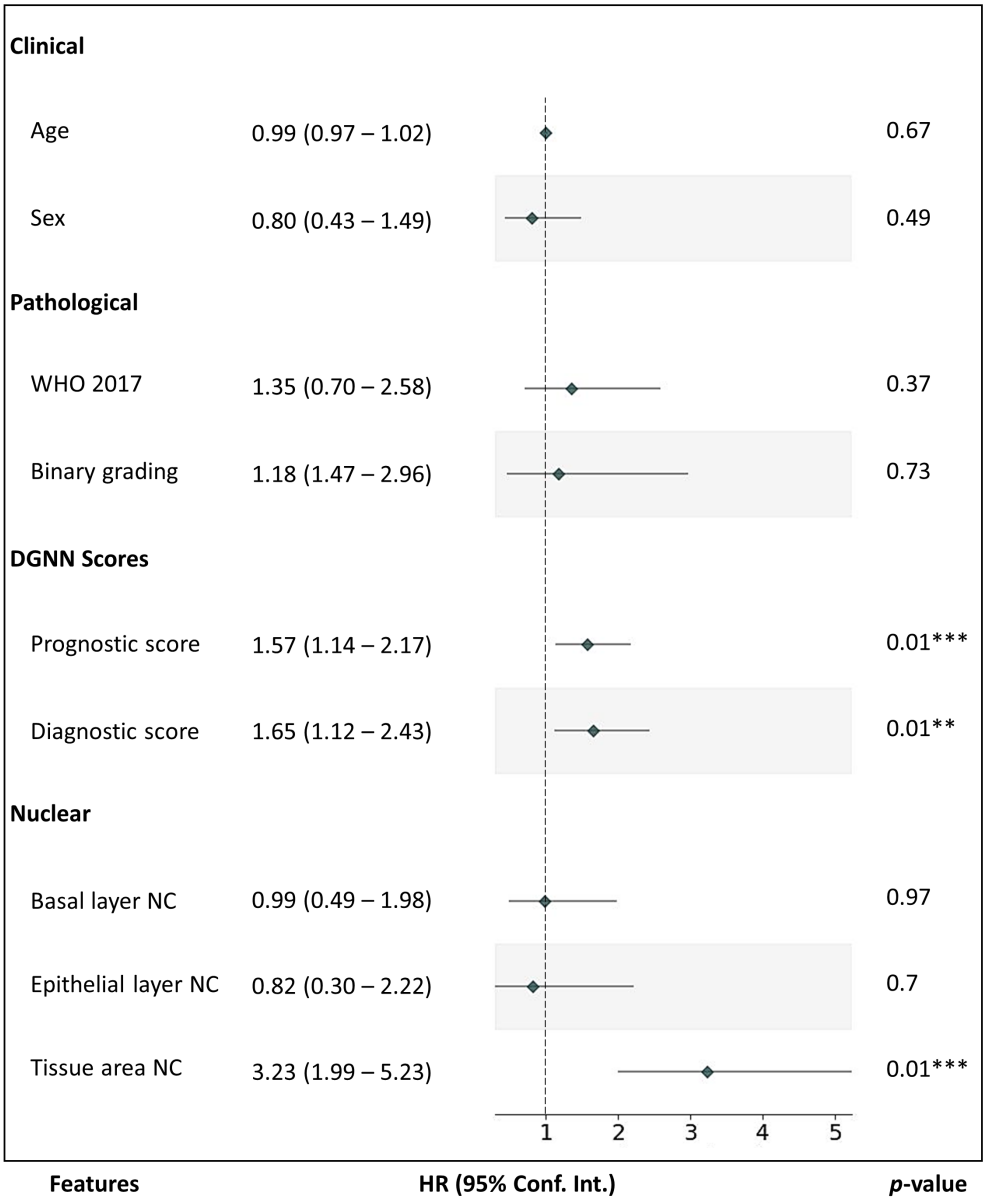


Figure 5.14: Multivariate forest plot of the log of hazard ratios for clinical, pathological, GNN scores, and top nuclear features using Cox Proportional Hazard model.

Features	Aggregation	p	C-index	Lower 95%	Upper 95%
Clinical Parameters					
Gender	-	>0.05	0.52	0.52	0.53
Age	-	>0.05	0.59	0.59	0.60
Pathological Parameters					
WHO Grading (Mild vs Mod + Severe)	-	<0.05	0.68	0.68	0.69
WHO Grading (Mild + Mod vs Severe)	-	<0.05	0.68	0.68	0.68
Binary Grading	-	<0.05	0.68	0.68	0.69
GNN scores					
Diagnostic score	-	<0.05	0.69	0.68	0.70
Prognostic score	-	<0.05	0.70	0.69	0.71
Nuclear Features					
Basal layer NC	μ	<0.05	0.81	0.80	0.82
	σ	<0.05	0.75	0.74	0.75
	m	<0.05	0.76	0.75	0.77
Epithelial layer NC	μ	<0.05	0.62	0.61	0.63
	σ	<0.05	0.70	0.69	0.70
	m	>0.05	0.59	0.58	0.60
Tissue area NC	μ	<0.05	0.83	0.82	0.84
	σ	<0.05	0.81	0.80	0.82
	m	<0.05	0.79	0.79	0.80

Table 5.10: Univariate analysis of the clinical, pathological and digital features where p is calculated using the log-rank method, and C-index is calculated using the Cox Proportional Hazard model bootstrapped 1000 times for lower and upper confidence intervals.

the GNNs can be attributed to the fact that it uses graph for learning, and the whole WSI can be efficiently represented as a graph with edges connecting the nodes, providing enough context to learn from the neighbourhood where other MIL techniques lack the spatial neighbourhood context and treat all patches either using higher attention, random or both. To the best of our knowledge, there has been very little work on oral epithelial dysplasia grading, where Sami et al. [201] proposed a CAD system for differentiation of dysplasia and carcinoma in-situ with the help of the epithelial border, i.e., bulbous shaped rete processes. However, the region of interest (ROI) was manually selected, where the pathologist selects the ROIs for further processing. Nag et al. [200] proposed a nuclear segmentation based approach to differentiate between normal epithelium and oral submucous fibrosis with the help of nuclear

Features	Aggregation	p	HR	Lower 95%	Upper 95%
C-index = 0.85, 95% CI [0.84 − 0.86]					
Clinical Parameters					
Gender	-	>0.05	0.99	0.97	1.02
Age	-	>0.05	0.80	0.43	1.49
Pathological Parameters					
WHO	-	>0.05	1.35	0.70	2.58
Binary Grading	-	>0.05	0.47	0.68	2.96
GNN scores					
Diagnostic score	-	<0.05	1.65	1.12	2.43
Prognostic score	-	<0.05	1.57	1.14	2.17
Nuclear Features					
Basal layer NC	μ	>0.05	0.99	0.49	1.98
Epithelial layer NC	σ	>0.05	0.82	0.30	2.22
Tissue area NC	μ	<0.05	3.23	1.99	5.23

Table 5.11: Multivariate analysis of the clinical, pathological, GNN scores and significant nuclear features where p is calculated using the log-rank test, and C-index is calculated using the Cox Proportional Hazard model bootstrapped 1000 times for lower and upper confidence interval.

features such as entropy, polarity and compactness where the limitation of the study facts to the use of only 100 nuclei from each group. Adel et al. [197] presented a CAD design where they used traditional feature extractors and trained SVM (Support Vector Machine) and KNN (k-Nearest Neighbours) for dysplastic and normal regions manually annotated by pathologists. Das et al. [198] presented a patch-based epithelium segmentation approach to classify keratin pearls for clinically relevant regions. They used a deep convolutional neural network (CNN) model to differentiate keratinised area vs epithelial area for segmentation and to further extract Gabor features from these areas to classify the keratin pearls. Das et al. [215] presented a patch-based deep CNN model for the classification of epithelium, muscle tissue, adipose tissue and connective tissue in oral biopsies segmenting the epithelium from other regions and artefacts for accurate ROI selection for classification purposes. Bashir et al. [50] proposed a machine learning based approach to differentiate between different dysplastic grades by exploiting irregular epithelium stratification as an important feature in the epithelial layers. Silva et al. [202] proposed a deep learning framework for nuclear segmentation and classification of OED grades in histological images from mouse oral mucosa, employing a Mask R-CNN for nuclei detection.We extracted 23 morphological and non-morphological nuclear features to train a polynomial classifier for grading. Nguyen et al.

[216] also reported the classification of epithelium vs non-epithelium regions and grading into normal, low-risk, high-risk and carcinoma using Inception v3 model and transfer learning. They also compared the performance of the AI system with human performance, with a kappa score of 0.81. Liu et al. [217] utilised DeepLab-v3 for segmenting high-risk regions from moderate and severe cases only, which were manually marked by the pathologist. Their best performing model achieved an overall accuracy and F1-Score of 93.3 percent and 90.9 percent, respectively, in a held-out test set of 44 WSIs. Regardless of the advancement and improvements in the aforementioned techniques, there was an element of manual selection of ROIs by pathologists in them, which makes them dependent on human input. Also, these techniques focused mainly on grading OED, lacking predictive factors for recurrence or malignant transformation in their studies. Whereas in our proposed work, we performed OED grading in an end-to-end manner where first epithelial layers were segmented using the coarse segmentation model and then GNN was trained for predicting the binary grade and malignant transformation status. Similarly, to the best of our knowledge, there has been little work predicting malignant transformation using AI tools from pre-cancerous lesions, i.e., in our case, OED. Mahmood et al. [204], who achieved an AUROC of 0.77 using a similar but smaller cohort and subjective assessment of nuclear features by three pathologists, showed an association between the nuclear features and malignant transformation. However, the nuclear features used correspond to OED grading, e.g., bulbous rete pegs, loss of epithelial cohesion etc., and upon adding histological grades into the mix, they observed improvements in their results. Zhang et al. [218] reported that the traditional grading systems for estimating the risk of malignant progression in oral lesions require a specially trained pathologist and suffer from poor reproducibility. In their study, an oral mucosa risk stratification (OMRS) model was developed, which outperformed the traditional three-tier system and was comparable to recent binary grading approaches. Binary classification systems have improved accuracy and shown a higher rate of malignant transformation in high-risk cases. Gan et al. [207] reported findings regarding potentially malignant epithelial lesions (PELs) using RNA sequencing of immune-infiltrating sites in cases of moderate and severe OED. The authors suggested that the absence of CD8 T-cells in the non-cytotoxic subtype and non-immune reactive subtype may contribute to moderate and severe dysplasia progression. Ellonen et al. [219] explored the frequency of transformation of oral epithelial dysplasia (OED) to oral squamous cell carcinoma (OSCC) and identified factors that influence this transformation. They concluded that the tongue and more severe grades of OED increase the risk of malignant transformation, and these patients might benefit from a more frequent follow-up to ensure early diagnosis of OSCC. Once our models were trained for

diagnostic and prognostic tasks, we investigated the cellular compositions in the top 15% of the potential malignant areas (hotspots) of transformed cases and non-transformed areas (coldspots). Our analysis found that nuclear features from the epithelial and basal layers were the most significant for diagnostic tasks, i.e., predicting OED grades. Epithelial layer nuclei count and basal layer nuclei count were found to be the most significant nuclear features differentiating the two grade distributions. During the experiments, we also identified other important features in the epithelial and basal layers, e.g., crowdedness in both layers, the solidity of cells, means major and minor axis length, etc. These nuclear features corresponded to aberrations of nuclei, such as variations in the size of nuclei captured as a variation in the minor axis of the nuclei and congestion due to the proliferation of nuclei in the epithelial and basal layers. In the same way, we found that nuclear features from the basal layer and connective tissue area were the most significant for prognostic tasks, i.e., predicting OED malignant transformation. Tissue nuclei count being the most significant feature among the transformed and non-transformed cases aligns with the finding of Bashir et al. [83] where they found peri-epithelial lymphocytes to be one of the most significant prognostic factors in their study. However, to further verify the significance of these features for both tasks, we require more multi-centric data to validate these features for their diagnostic and prognostic significance in oral pre-cancerous lesions. We also analysed different clinical, pathological, GNN scores and nuclear features for progression free survival and have found out that certain factors, such as nuclei count in the sub-layer of oral epithelium, i.e., basal layer and epithelial layer, were linked to PFS, where a higher number of nuclei count in basal layer correspond to poor survival whereas small variation nuclei count in the epithelium is linked with the improved survival. Also, apart from the epithelial layer, the peri-epithelium tissue region is also very important, as in our study, we found that the nuclei count in the adjacent connective tissue areas has more significance in predicting PFS in a multivariate setting. Apart from the nuclear features, the clinical and pathological factors such as age, gender, and pathologist's grades were nonsignificant in the presence of GNN diagnostic and prognostic scores predicted by our trained graph neural networks. Although the cohort is relatively small and unicentric, the department is a regional and national referral centre in the UK. Nonetheless, the practical application and adaptation of these methods in clinical practice require substantially large and truly multi-centric cohort data allowing more rigorous validation of the proposed algorithms.

5.4 Chapter Summary

To the best of our knowledge, this is the first study to perform diagnostic and prognostic tasks using graph neural networks and has shown that nuclear features from the epithelial and basal layers are important for OED grading. While nuclear features from the basal layer and peri-epithelium were found to be more significant for predicting malignant transformation in OED. Our proposed methodology for predicting OED grading and malignancy in an end-to-end manner has the potential to play an important role in precision medicine and personalised patient management for early prediction of malignancy risk with the potential to guide treatment decisions and risk stratification.

Chapter 6

Conclusions and Future Directions

In this chapter, we provide a summary of the methods presented in this thesis, along with their potential avenues for future exploration. In this thesis, we tackled the challenges of learning with minimal labels in histopathology images for classification, detection and segmentation. We proposed a set of computational methods using semi-supervised and weakly-supervised learning frameworks for automated analysis of H&E images. We start with partially labelled data to simultaneously detect and classify cells in DLBCL using a multi-task semi-supervised learning framework. This process includes the segmentation of nuclei and tissue regions in histology images, which is achieved using self-supervised and semi-supervised learning frameworks. Our results demonstrate that the use of consistency-based self-supervised techniques can improve performance when there is a lack of sufficient labelled data.

In the realm of weakly supervised learning, where a single label is assigned to the whole WSI, we explored MIL methods to classify the WSI for diagnostic and prognostic purposes. We proposed a pipeline for predicting malignant transformation in OED using MIL based models and developed a digital biomarker for predicting PFS in OED patients. We first segmented the epithelium into fine layers and created different graphs using deep and nuclear features for training the MIL models. A GNN was trained using ranking loss for predicting the OED grade and malignant transformation. Further, we analysed nuclear features to find a significant digital biomarker for predicting PFS in OED.

The subsequent sections provide a summary of some methods proposed in this study, highlighting the main contributions, limitations, and potential future extensions.

6.1 Self- and Semi- supervised Learning for Histology Images

In Chapters 2 and 3, we have used semi-supervised learning for classifying and segmenting cells and tissue regions in histology images using limited labelled instances. We showed that using self-supervision alongside semi-supervised learning can better cope with the challenges like generalisation, robustness and context-awareness. Although we have used self-supervision, the performance for smaller portions of the labelled data split suffered with high bias and variance compared to bigger portions. This is due to the fact that the data coming from different centres suffers highly from visual appearance (i.e., stain variation) and class imbalance (i.e., more tumour region as compared to normal tissue).

One possible extension would be to use the latest stain augmentation techniques, e.g., latent diffusion [220] to create more variation of the input images for contrastive learning. Latent diffusion and stain augmentation can also be used to balance the class representation using its near-real generative abilities. These images can serve as a basis for self-supervised learning where no additional labels would be required to train a network.

6.2 Recurrence in Oral Epithelial Dysplasia (OED)

In Chapters 4 and 5, we used multiple instance learning to predict the OED grade and malignant transformation using deep neural networks (DNN) and graph neural networks (GNN). We showed that new digital biomarkers for progression free survival (PFS) could be found using the DNN and GNN prediction maps on whole slide imaging. OED grade and malignant transformation status (i.e., in future) can help clinicians to exercise the best course of action for the patients, but it may involve the excision of OED lesions with larger margins to avoid recurrence and malignant transformation. It has been observed that in some cases where cases from low- and high- risk dysplasia only recur and do not transform into malignancy, such cases do not require resections with larger margins.

One possible future direction would be to predict the recurrence status for OED cases and explore significant digital biomarkers for predicting recurrence-free survival. Moreover, our current findings were only validated on a large internal dataset, so future work could look into the external validation of the possible extension along with the aforementioned diagnostic and prognostics digital biomarkers for clinical adaption.

6.3 GNN based Multi-Task Learning for Oral Epithelial Dysplasia analysis

In Chapter 5, we have trained single task models for diagnostic and prognostic tasks in OED using GNNs. Recently, multi-task learning (MTL) has been used to predict multiple outcomes in histology images [67], given the fact that single task models fail to scale with the addition of new tasks, e.g., predicting recurrence. Another area for improvement with the single task models is their susceptibility to generalise specifically for that task making it hard to transfer it to other tasks. On the other hand, MTL [221] can solve the aforementioned issues by utilising a single encoder which not only shares weights but can also utilise essential features across the tasks while having the flexibility to extend to new tasks.

In the future GNN based single task learning could be extended to multi-task learning for simultaneously predicting OED grade, recurrence and malignant transformation status using one model. To this end, this problem can be treated as a particular case of MTL where the same input data will have different labels associated with the single WSI commonly known as multi-label learning (MLL) [222]. Another advantage of solving it using MTL is that MLL requires annotations for all three labels, whereas, for MTL, available labels for each WSI can be utilised in each task.

6.4 Closing Remarks

In this thesis, we have proposed frameworks capable of learning from limited data and minimal labels from histology based H&E stained whole slide images. The frameworks we developed covered the classification, detection and segmentation of different nuclei and tissue regions along with the WSI level diagnostic and prognostic labels prediction. The clinical adaption and deployment of these AI based frameworks require strong validation using large-scale clinical trials spanning multiple regions across the world, including different hospitals and pathologists, covering a range of variations arising from digital scanners.

With the current momentum in AI conquering the vast majority of labour-intensive tasks with its intelligence, it is likely that CPath will be ubiquitous in digital pathology in the coming years. It will not only help assist pathologists in diagnostic, prognostic and therapeutic tasks but will also alleviate the current shortage observed in the field of histopathology.

Bibliography

[1] C. C. Compton, D. R. Byrd, J. Garcia-Aguilar, S. H. Kurtzman, A. Olawaiye, and M. K. Washington, "Hodgkin and non-hodgkin lymphomas," in *AJCC Cancer Staging Atlas*, pp. 605–617, Springer, 2012.

[2] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[3] World Health Organization, "Cancer, https://www.who.int/westernpacific/health-topics/cancer," 2023.

[4] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Global cancer observatory: Cancer today, https://gco.iarc.fr/today/home," 2022.

[5] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, *Molecular Cell Biology*. W. H. Freeman, 4th ed., 2000.

[6] D. Weller, P. Vedsted, G. Rubin, F. Walter, J. Emery, S. Scott, C. Campbell, R. S. Andersen, W. Hamilton, F. Olesen, *et al.*, "The Aarhus statement: improving design and reporting of studies on early cancer diagnosis," *British Journal of Cancer*, vol. 106, no. 7, pp. 1262–1267, 2012.

[7] L. Barnes, J. W. Eveson, D. Sidransky, and P. Reichart, *Pathology and Genetics of Head and Neck Tumours*, vol. 9. IARC, 2005.

[8] A. Wright and M. Shear, "Epithelial dysplasia immediately adjacent to oral squamous cell carcinomas," *Journal of Oral Pathology & Medicine*, vol. 14, no. 7, pp. 559–564, 1985.

[9] J. Califano, W. H. Westra, G. Meininger, R. Corio, W. M. Koch, and D. Sidransky, "Genetic progression and clonal relationship of recurrent premalignant head and neck lesions," *Clinical Cancer Research*, vol. 6, no. 2, pp. 347–352, 2000.

[10] A. K. El-Naggar, J. K. C. Chan, J. R. Grandis, T. Takata, and P. J. Slootweg, *WHO Classification of Head and Neck Tumours*. International Agency for Research on Cancer, Jan. 2017. Google-Books-ID: EDo5MQAACAAJ.

[11] L. Barnes, J. Eveson, P. Reichart, D. Sidransky, *et al.*, "World health organization classification of tumours: pathology and genetics of head and neck tumours," *WHO*, 2005.

[12] O. Kujan, R. J. Oliver, A. Khattab, S. A. Roberts, N. Thakker, and P. Sloan, "Evaluation of a new binary system of grading oral epithelial dysplasia for prediction of malignant transformation," *Oral Oncology*, vol. 42, no. 10, pp. 987–993, 2006.

[13] A. El-Naggar, J. Chan, J. Rubin Grandis, T. Takata, and P. Slootweg, "Who classification of head and neck tumours.: International agency for research on cancer; 2017," *WHO*, pp. 159–202, 2017.

[14] O. Kujan, A. Khattab, R. J. Oliver, S. A. Roberts, N. Thakker, and P. Sloan, "Why oral histopathology suffers inter-observer variability on grading oral epithelial dysplasia: an attempt to understand the sources of variation," *Oral Oncology*, vol. 43, no. 3, pp. 224–231, 2007.

[15] S. Li, K. H. Young, and L. J. Medeiros, "Diffuse large b-cell lymphoma," *Pathology*, vol. 50, no. 1, pp. 74–87, 2018.

[16] A. Smith, D. Howell, R. Patmore, A. Jack, and E. Roman, "Incidence of haematological malignancy by sub-type: a report from the haematological malignancy research network," *British journal of cancer*, vol. 105, no. 11, 2011.

[17] B. Coiffier, E. Lepage, J. Brière, R. Herbrecht, H. Tilly, R. Bouabdallah, P. Morel, E. Van Den Neste, G. Salles, P. Gaulard, *et al.*, "Chop chemotherapy plus rituximab compared with chop alone in elderly patients with diffuse large-b-cell lymphoma," *New England Journal of Medicine*, vol. 346, no. 4, pp. 235–242, 2002.

[18] A. De Jonge, T. Roosma, I. Houtenbos, W. Vasmel, K. van de Hem, J. de Boer, T. van Maanen, G. Lindauer-van der Werf, A. Beeker, G. Timmers, *et al.*, "Diffuse large b-cell lymphoma with myc gene rearrangements: Current perspective on treatment of diffuse large b-cell lymphoma with myc gene rearrangements; case series and review of the literature," *European Journal of Cancer*, vol. 55, pp. 140–146, 2016.

[19] The Royal College of Pathologists, "Histopathology, https://www.rcpath.org/discover-pathology/news/fact-sheets/histopathology.html," 2020.

[20] A. Bychkov and M. Schubert, "Constant demand, patchy supply. global pathology workforce heatmap, https://thepathologist.com/outside-the-lab/constant-demand-patchy-supply," 2023.

[21] A. N. Scholten, V. T. Smit, H. Beerman, W. L. van Putten, and C. L. Creutzberg, "Prognostic significance and interobserver variability of histologic grading systems for endometrial carcinoma," *Cancer*, vol. 100, no. 4, pp. 764–772, 2004.

[22] S. S. Cross, S. Betmouni, J. L. Burton, A. K. Dubé, K. M. Feeley, M. R. Holbrook, R. J. Landers, P. B. Lumb, and T. J. Stephenson, "What levels of agreement can be expected between histopathologists assigning cases to discrete nominal categories? a study of the diagnosis of hyperplastic and adenomatous colorectal polyps," *Modern Pathology*, vol. 13, no. 9, pp. 941–944, 2000.

[23] K. Komuta, K. Batts, J. Jessurun, D. Snover, J. Garcia-Aguilar, D. Rothenberger, and R. Madoff, "Interobserver variability in the pathological assessment of malignant colorectal polyps," *Journal of British Surgery*, vol. 91, no. 11, pp. 1479–1484, 2004.

[24] M. May, "A better lens on disease," *Scientific American*, vol. 302, no. 5, pp. 74–77, 2010.

[25] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart, "The cancer genome atlas pan-cancer analysis project," *Nature Genetics*, vol. 45, no. 10, pp. 1113–1120, 2013.

[26] A. Goode, B. Gilbert, J. Harkes, D. Jukic, and M. Satyanarayanan, "OpenSlide: A vendor-neutral software foundation for digital pathology," *Journal of Pathology Informatics*, vol. 4, no. 1, p. 27, 2013.

[27] M. Rabbani, "Jpeg2000: Image compression fundamentals, standards and practice," *Journal of Electronic Imaging*, vol. 11, no. 2, p. 286, 2002.

[28] P. Bankhead, M. B. Loughrey, J. A. Fernández, Y. Dombrowski, D. G. McArt, P. D. Dunne, S. McQuaid, R. T. Gray, L. J. Murray, H. G. Coleman, J. A. James, M. Salto-Tellez, and P. W. Hamilton, "QuPath: Open source software for digital pathology image analysis," *Scientific Reports*, vol. 7, p. 16878, Dec. 2017.

[29] "ASAP - Fluid whole-slide image viewer, https://www.computationalpathologygroup.eu/software/asap/," 2019.

[30] D. A. Gutman, M. Khalilia, S. Lee, M. Nalisnik, Z. Mullen, J. Beezley, D. R. Chittajallu, D. Manthey, and L. A. Cooper, "The digital slide archive: A software platform for management, integration, and analysis of histology for cancer research," *Cancer Research*, vol. 77, no. 21, pp. e75–e78, 2017.

[31] D. N. Louis, G. K. Gerber, J. M. Baron, L. Bry, A. S. Dighe, G. Getz, J. M. Higgins, F. C. Kuo, W. J. Lane, J. S. Michaelson, L. P. Le, C. H. Mermel, J. R. Gilbertson, and J. A. Golden, "Computational Pathology: An Emerging Definition," *Archives of Pathology & Laboratory Medicine*, vol. 138, pp. 1133–1138, 09 2014.

[32] Y. Yagi, "Color standardization and optimization in whole slide imaging," *Diagnostic Pathology*, vol. 6, pp. 1–12, 2011.

[33] W. M. Kouw and M. Loog, "A review of domain adaptation without target labels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 766–785, 2019.

[34] A. M. Khan, N. Rajpoot, D. Treanor, and D. Magee, "A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 6, pp. 1729–1738, 2014.

[35] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas, "A method for normalizing histology slides for quantitative analysis," in *2009 IEEE International Symposium on Biomedical Imaging: from Nano to Macro*, (Boston, MA, USA), pp. 1107–1110, IEEE, 2009.

[36] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A. M. Schlitter, I. Esposito, and N. Navab, "Structure-preserving color normalization and sparse stain separation for histological images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 8, pp. 1962–1971, 2016.

[37] M. T. Shaban, C. Baur, N. Navab, and S. Albarqouni, "Staingan: Stain style transfer for digital histological images," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, (Venice, Italy), pp. 953–956, 2019.

[38] H. Wu, J. H. Phan, A. K. Bhatia, C. A. Cundiff, B. M. Shehata, and M. D. Wang, "Detection of blur artifacts in histopathological whole-slide images of endomyocardial biopsies," in *2015 37th annual international Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, (Milan, Italy), pp. 727–730, IEEE, 2015.

[39] A. Janowczyk, R. Zuo, H. Gilmore, M. Feldman, and A. Madabhushi, "Histoqc: an open-source quality control tool for digital pathology slides," *JCO Clinical Cancer Informatics*, vol. 3, pp. 1–7, 2019.

[40] M. Haghighat, L. Browning, K. Sirinukunwattana, S. Malacrino, N. Khalid Alham, R. Colling, Y. Cui, E. Rakha, F. C. Hamdy, C. Verrill, *et al.*, "Automated quality assessment of large digitised histology cohorts by artificial intelligence," *Scientific Reports*, vol. 12, no. 1, p. 5002, 2022.

[41] K. Sirinukunwattana, S. E. A. Raza, Y.-W. Tsang, D. R. Snead, I. A. Cree, and N. M. Rajpoot, "Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1196–1206, 2016.

[42] S. Graham, Q. D. Vu, S. E. A. Raza, A. Azam, Y. W. Tsang, J. T. Kwak, and N. Rajpoot, "Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images," *Medical Image Analysis*, vol. 58, p. 101563, 2019.

[43] J. Gamper, N. A. Koohbanani, K. Benes, S. Graham, M. Jahanifar, S. A. Khurram, A. Azam, K. Hewitt, and N. Rajpoot, "Pannuke dataset extension, insights and baselines," *arXiv preprint arXiv:2003.10778*, 2020.

[44] M. Aubreville, N. Stathonikos, C. A. Bertram, R. Klopfleisch, N. Ter Hoeve, F. Ciompi, F. Wilm, C. Marzahl, T. A. Donovan, A. Maier, *et al.*, "Mitosis domain generalization in histopathology images—the midog challenge," *Medical Image Analysis*, vol. 84, p. 102699, 2023.

[45] M. Amgad, H. Elfandy, H. Hussein, L. A. Atteya, M. A. Elsebaie, L. S. Abo Elnasr, R. A. Sakr, H. S. Salem, A. F. Ismail, A. M. Saad, *et al.*, "Structured

crowdsourcing enables convolutional segmentation of histology images," *Bioinformatics*, vol. 35, no. 18, pp. 3461–3467, 2019.

[46] R. M. S. Bashir, M. Shaban, S. E. A. Raza, S. A. Khurram, and N. Rajpoot, "A novel framework for coarse-grained semantic segmentation of whole-slide images," in *Annual Conference on Medical Image Understanding and Analysis*, (Cambridge, UK), pp. 425–439, Springer, 2022.

[47] M. Shaban, S. A. Khurram, M. M. Fraz, N. Alsubaie, I. Masood, S. Mushtaq, M. Hassan, A. Loya, and N. M. Rajpoot, "A novel digital score for abundance of tumour infiltrating lymphocytes predicts disease free survival in oral squamous cell carcinoma," *Scientific Reports*, vol. 9, no. 1, pp. 1–13, 2019.

[48] A. Rastogi, "Changing role of histopathology in the diagnosis and management of hepatocellular carcinoma," *World Journal of Gastroenterology*, vol. 24, no. 35, p. 4000, 2018.

[49] K. Bera, K. A. Schalper, D. L. Rimm, V. Velcheti, and A. Madabhushi, "Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology," *Nature Reviews Clinical Oncology*, vol. 16, no. 11, pp. 703–715, 2019.

[50] R. S. Bashir, H. Mahmood, M. Shaban, S. E. A. Raza, M. M. Fraz, S. A. Khurram, and N. M. Rajpoot, "Automated grade classification of oral epithelial dysplasia using morphometric analysis of histology images," in *Medical Imaging 2020: Digital Pathology*, vol. 11320, (Houston, Texas, USA), pp. 245–250, SPIE, 2020.

[51] A. Shephard, N. Azarmehr, R. M. S. Bashir, S. E. A. Raza, H. Mahmood, S. A. Khurram, and N. Rajpoot, "A fully automated multi-scale pipeline for oral epithelial dysplasia grading and outcome prediction," in *Medical Imaging with Deep Learning*, (Zürich, Switzerland), 2022.

[52] J. I. Epstein, "An update of the gleason grading system," *The Journal of Urology*, vol. 183, no. 2, pp. 433–440, 2010.

[53] M. Shaban, S. E. A. Raza, M. Hassan, A. Jamshed, S. Mushtaq, A. Loya, N. Batis, J. Brooks, P. Nankivell, N. Sharma, M. Robinson, H. Mehanna, S. A. Khurram, and N. Rajpoot, "A digital score of tumour-associated stroma infiltrating lymphocytes predicts survival in head and neck squamous cell carcinoma," *The Journal of Pathology*, vol. 256, no. 2, pp. 174–185, 2022.

[54] M. Bilal, S. E. A. Raza, A. Azam, S. Graham, M. Ilyas, I. A. Cree, D. Snead, F. Minhas, and N. M. Rajpoot, "Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study," *The Lancet Digital Health*, vol. 3, no. 12, pp. e763–e772, 2021.

[55] J. N. Kather, A. T. Pearson, N. Halama, D. Jäger, J. Krause, S. H. Loosen, A. Marx, P. Boor, F. Tacke, U. P. Neumann, *et al.*, "Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer," *Nature Medicine*, vol. 25, no. 7, pp. 1054–1056, 2019.

[56] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of Research and Development*, vol. 3, no. 3, pp. 210–229, 1959.

[57] Y. Bengio *et al.*, "Learning deep architectures for ai," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[58] D. Eigen, J. Rolfe, R. Fergus, and Y. LeCun, "Understanding deep architectures using a recursive convolutional network," *arXiv preprint arXiv:1312.1847*, 2013.

[59] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[60] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[61] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[62] J. Wang, J. D. MacKenzie, R. Ramachandran, and D. Z. Chen, "A deep learning approach for semantic segmentation in histology tissue images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (Athens, Greece), pp. 176–184, Springer, 2016.

[63] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, and A. Tsirigos, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nature Medicine*, vol. 24, no. 10, pp. 1559–1567, 2018.

[64] D. Ciresan, A. Giusti, L. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[65] S. Kothari, J. H. Phan, T. H. Stokes, and M. D. Wang, "Pathology imaging informatics for quantitative analysis of whole-slide images," *Journal of the American Medical Informatics Association*, vol. 20, no. 6, pp. 1099–1108, 2013.

[66] J. Gamper, N. A. Kooohbanani, and N. Rajpoot, "Multi-task learning in histopathology for widely generalizable model," *arXiv preprint arXiv:2005.08645*, 2020.

[67] S. Graham, Q. D. Vu, M. Jahanifar, S. E. A. Raza, F. Minhas, D. Snead, and N. Rajpoot, "One model is all you need: multi-task learning enables simultaneous histology image segmentation and classification," *Medical Image Analysis*, vol. 83, p. 102685, 2023.

[68] H. Wu, Z. Wang, Y. Song, L. Yang, and J. Qin, "Cross-patch dense contrastive learning for semi-supervised segmentation of cellular nuclei in histopathologic images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (New Orleans, Louisiana, USA), pp. 11666–11675, 2022.

[69] N. Kumar, R. Verma, D. Anand, Y. Zhou, O. F. Onder, E. Tsougenis, H. Chen, P.-A. Heng, J. Li, Z. Hu, *et al.*, "A multi-organ nucleus segmentation challenge," *IEEE Transactions on Medical Imaging*, vol. 39, no. 5, pp. 1380–1391, 2019.

[70] R. M. S. Bashir, T. Qaiser, S. E. A. Raza, and N. M. Rajpoot, "Consistency regularisation in varying contexts and feature perturbations for semi-supervised semantic segmentation of histology images," *arXiv preprint arXiv:2301.13141*, 2023.

[71] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 2020.

[72] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[73] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[74] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[75] E. Gawehn, J. A. Hiss, and G. Schneider, "Deep learning in drug discovery," *Molecular Informatics*, vol. 35, no. 1, pp. 3–14, 2016.

[76] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, 2020.

[77] V. Anklin, P. Pati, G. Jaume, B. Bozorgtabar, A. Foncubierta-Rodriguez, J.-P. Thiran, M. Sibony, M. Gabrani, and O. Goksel, "Learning whole-slide segmentation from inexact and incomplete labels using tissue graphs," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, (Strasbourg, France), pp. 636–646, Springer, 2021.

[78] D. Ahmedt-Aristizabal, M. A. Armin, S. Denman, C. Fookes, and L. Petersson, "A survey on graph-based deep learning for computational histopathology," *Computerized Medical Imaging and Graphics*, vol. 95, p. 102027, 2022.

[79] W. Lu, M. Toss, M. Dawood, E. Rakha, N. Rajpoot, and F. Minhas, "Slidegraph+: whole slide image level graphs to predict her2 status in breast cancer," *Medical Image Analysis*, vol. 80, p. 102486, 2022.

[80] Y. Alon and H. Zhou, "Neuroplastic graph attention networks for nuclei segmentation in histopathology images," *arXiv preprint arXiv:2201.03669*, 2022.

[81] C. C. Mackenzie, M. Dawood, S. Graham, M. Eastwood, *et al.*, "Neural graph modelling of whole slide images for survival ranking," in *Learning on Graphs Conference*, (Virtual), pp. 48–1, PMLR, 2022.

[82] R. M. S. Bashir, T. Qaiser, S. E. A. Raza, and N. M. Rajpoot, "Hydramix-net: A deep multi-task semi-supervised learning approach for cell detection and classification," in *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, pp. 164–171, Springer, 2020.

[83] R. M. S. Bashir, A. Shephard, H. Mahmood, N. Azarmehr, S. E. A. Raza, A. Khurram, and N. Rajpoot, "A digital score of peri-epithelial lymphocytic activity predicts malignant transformation in oral epithelial dysplasia," *medRxiv*, pp. 2023–02, 2023.

[84] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]," *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542–542, 2009.

[85] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[86] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.

[87] D. C. Cireşan, U. Meier, L. M. Gambardella, and J. Schmidhuber, "Deep, big, simple neural nets for handwritten digit recognition," *Neural Computation*, vol. 22, no. 12, pp. 3207–3220, 2010.

[88] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[89] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," *Advances in Neural Information Processing Systems*, vol. 17, 2004.

[90] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey, "Symmetric cross entropy for robust learning with noisy labels," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (Seoul, Korea), pp. 322–330, 2019.

[91] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[92] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[93] T. Qaiser, M. Pugh, S. Margielewska, R. Hollows, P. Murray, and N. Rajpoot, "Digital tumor-collagen proximity signature predicts survival in diffuse large b-cell lymphoma," in *Digital Pathology: 15th European Congress, ECDP 2019, Warwick, UK, April 10–13, 2019, Proceedings 15*, (Warwick,United Kingdom), pp. 163–171, Springer, 2019.

[94] N. C. I. Surveillance Research Program, "Seer*explorer: An interactive website for seer cancer statistics [internet]. accessed at https://seer.cancer.gov/explorer/ on february 23, 2023.," 2023.

[95] B. Coiffier, E. Lepage, J. Brière, R. Herbrecht, H. Tilly, R. Bouabdallah, P. Morel, E. Van Den Neste, G. Salles, P. Gaulard, *et al.*, "Chop chemotherapy plus rituximab compared with chop alone in elderly patients with diffuse large-b-cell lymphoma," *New England Journal of Medicine*, vol. 346, no. 4, pp. 235–242, 2002.

[96] A. De Jonge, T. Roosma, I. Houtenbos, W. Vasmel, K. van de Hem, J. de Boer, T. van Maanen, G. Lindauer-van der Werf, A. Beeker, G. Timmers, *et al.*, "Diffuse large b-cell lymphoma with myc gene rearrangements: Current perspective on treatment of diffuse large b-cell lymphoma with myc gene rearrangements; case series and review of the literature," *European Journal of Cancer*, vol. 55, pp. 140–146, 2016.

[97] Z. Chen, X. Deng, Y. Ye, L. Gao, W. Zhang, W. Liu, and S. Zhao, "Novel risk stratification of de novo diffuse large b cell lymphoma based on tumour-infiltrating t lymphocytes evaluated by flow cytometry," *Annals of Hematology*, vol. 98, pp. 391–399, 2019.

[98] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[99] N. A. Koohbanani, M. Jahanifar, N. Z. Tajadin, and N. Rajpoot, "Nuclick: a deep learning framework for interactive segmentation of microscopic images," *Medical Image Analysis*, vol. 65, p. 101771, 2020.

[100] A. J. Shephard, S. Graham, S. Bashir, M. Jahanifar, H. Mahmood, A. Khurram, and N. M. Rajpoot, "Simultaneous nuclear instance and layer segmentation in oral epithelial dysplasia," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (Virtual), pp. 552–561, 2021.

[101] T. Qaiser, Y.-W. Tsang, D. Taniyama, N. Sakamoto, K. Nakane, D. Epstein, and N. Rajpoot, "Fast and accurate tumor segmentation of histology images using persistent homology and deep convolutional features," *Medical Image Analysis*, vol. 55, pp. 1–14, 2019.

[102] Q. D. Vu, C. Fong, K. von Loga, S. E. A. Raza, D. Nava Rodrigues, B. Patel, C. Peckitt, R. Begum, A. Athauda, N. Starling, *et al.*, "Digital histological markers based on routine h&e slides to predict benefit from maintenance immunotherapy in esophagogastric adenocarcinoma.," 2021.

[103] M. Jahanifar, A. Shepard, N. Zamanitajeddin, R. Bashir, M. Bilal, S. A. Khurram, F. Minhas, and N. Rajpoot, "Stain-robust mitotic figure detection for the mitosis domain generalization challenge," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (Strasbourg, France), pp. 48–52, Springer, 2021.

[104] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 555–570, 2021.

[105] Q. Da, X. Huang, Z. Li, Y. Zuo, C. Zhang, J. Liu, W. Chen, J. Li, D. Xu, Z. Hu, *et al.*, "Digestpath: a benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system," *Medical Image Analysis*, p. 102485, 2022.

[106] A. Echle, N. G. Laleh, P. Quirke, H. Grabsch, H. Muti, O. Saldanha, S. Brockmoeller, P. van den Brandt, G. Hutchins, S. Richman, *et al.*, "Artificial intelligence for detection of microsatellite instability in colorectal cancer—a multicentric analysis of a pre-screening tool for clinical application," *ESMO Open*, vol. 7, no. 2, p. 100400, 2022.

[107] M. Shaban, S. E. A. Raza, M. Hassan, A. Jamshed, S. Mushtaq, A. Loya, N. Batis, J. Brooks, P. Nankivell, N. Sharma, *et al.*, "A digital score of tumour-associated stroma infiltrating lymphocytes predicts survival in head and neck squamous cell carcinoma," *The Journal of Pathology*, vol. 256, no. 2, pp. 174–185, 2022.

[108] Y. Mao, E. T. Keller, D. H. Garfield, K. Shen, and J. Wang, "Stromal cells in tumor microenvironment and breast cancer," *Cancer and Metastasis Reviews*, vol. 32, no. 1, pp. 303–315, 2013.

[109] A. Tabesh, M. Teverovskiy, H.-Y. Pang, V. P. Kumar, D. Verbel, A. Kotsianti, and O. Saidi, "Multifeature prostate cancer diagnosis and gleason grading of histological images," *IEEE Transactions on Medical Imaging*, vol. 26, no. 10, pp. 1366–1378, 2007.

[110] J. Diamond, N. H. Anderson, P. H. Bartels, R. Montironi, and P. W. Hamilton, "The use of morphological characteristics and texture analysis in the identification of tissue composition in prostatic neoplasia," *Human Pathology*, vol. 35, no. 9, pp. 1121–1131, 2004.

[111] K. Sirinukunwattana, D. R. Snead, and N. M. Rajpoot, "A novel texture descriptor for detection of glandular structures in colon histology images," in *Medical Imaging 2015: Digital Pathology*, vol. 9420, (Orlando, Florida, United States), pp. 186–194, SPIE, 2015.

[112] D. Anoraganingrum, "Cell segmentation with median filter and mathematical morphology operation," in *Proceedings 10th International Conference on Image Analysis and Processing*, (Venice, Italy), pp. 1043–1046, IEEE, 1999.

[113] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (Munich, Germany), pp. 234–241, Springer, 2015.

[114] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *arXiv:1606.00915 [cs]*, 5 2017. arXiv: 1606.00915.

[115] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12077–12090, 2021.

[116] W. Zhang, Z. Huang, G. Luo, T. Chen, X. Wang, W. Liu, G. Yu, and C. Shen, "Topformer: Token pyramid transformer for mobile semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (New Orleans, Louisiana), pp. 12083–12093, 2022.

[117] K. Lindman, J. F. Rose, M. Lindvall, C. Lundstrom, and D. Treanor, "Annotations, ontologies, and whole slide images–development of an annotated ontology-driven whole slide image library of normal and abnormal human tissue," *Journal of Pathology Informatics*, vol. 10, no. 1, p. 22, 2019.

[118] B. E. Bejnordi, G. Litjens, M. Hermsen, N. Karssemeijer, and J. A. van der Laak, "A multi-scale superpixel classification approach to the detection of regions of interest in whole slide histopathology images," in *Medical Imaging 2015: Digital Pathology*, vol. 9420, (Orlando, Florida, United States), pp. 99–104, SPIE, 2015.

[119] M. Jahanifar, N. Z. Tajeddin, N. A. Koohbanani, and N. M. Rajpoot, "Robust interactive semantic segmentation of pathology images with minimal user input," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (Virtual), pp. 674–683, 2021.

[120] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *arXiv:1411.4038 [cs]*, 3 2015. arXiv: 1411.4038.

[121] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid sene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Honolulu, Hawaii), pp. 2881–2890, 2017.

[122] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, (Seoul, Korea), pp. 5693–5703, 2019.

[123] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[124] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (Virtual), pp. 6881–6890, 2021.

[125] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, (Atlanta, USA), p. 896, 2013.

[126] Y. Zou, Z. Zhang, H. Zhang, C.-L. Li, X. Bian, J.-B. Huang, and T. Pfister, "Pseudoseg: Designing pseudo labels for semantic segmentation," *arXiv preprint arXiv:2010.09713*, 2020.

[127] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (Virtual), pp. 2613–2622, 2021.

[128] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, and A. Courville, "Yoshua bengio generative adversarial networks," *arXiv preprint arXiv:1406.2661*, 2014.

[129] V. Badrinarayanan, A. Kendall, and R. C. SegNet, "A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, vol. 5, 2015.

[130] A. Kumar, P. Sattigeri, and T. Fletcher, "Semi-supervised learning with gans: Manifold invariance with improved inference," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[131] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang, "Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Salt Lake City, Utah.), pp. 7268–7277, 2018.

[132] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (Virtual), pp. 12674–12684, 2020.

[133] X. Lai, Z. Tian, L. Jiang, S. Liu, H. Zhao, L. Wang, and J. Jia, "Semi-supervised semantic segmentation with directional context-aware consistency," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (Virtual), pp. 1205–1214, 2021.

[134] K. Saito, D. Kim, S. Sclaroff, T. Darrell, and K. Saenko, "Semi-supervised domain adaptation via minimax entropy," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (Seoul, Korea), pp. 8050–8058, 2019.

[135] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Las Vegas, Nevada, USA), pp. 770–778, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6 2016. ISSN: 1063-6919.

[136] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[137] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv:1511.00561 [cs]*, 10 2016. arXiv: 1511.00561.

[138] P. Naylor, M. Laé, F. Reyal, and T. Walter, "Nuclei segmentation in histopathology images using deep neural networks," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, (Melbourne, VIC, Australia), pp. 933–936, IEEE, 2017.

[139] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[140] A. Azam, S. Bashir, S. Khurram, D. Snead, and N. Rajpoot, "A novel deep learning-based diagnostic algorithm for detection and segmentation of amyloid in digital whole slide images," in *VIRCHOWS ARCHIV*, vol. 477, (New York, NY, USA), pp. S214–S214, Springer, 2020.

[141] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[142] S. Graham, H. Chen, J. Gamper, Q. Dou, P.-A. Heng, D. Snead, Y. W. Tsang, and N. Rajpoot, "Mild-net: Minimal information loss dilated network for gland instance segmentation in colon histology images," *Medical Image Analysis*, vol. 52, pp. 199–211, 2019.

[143] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, "Asymmetric non-local neural networks for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (Seoul, Korea), pp. 593–602, 2019.

[144] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (Seoul, Korea), pp. 603–612, 2019.

[145] M. Fraz, S. A. Khurram, S. Graham, M. Shaban, M. Hassan, A. Loya, and N. M. Rajpoot, "Fabnet: feature attention-based network for simultaneous segmentation of microvessels and nerves in routine histology images of oral cancer," *Neural Computing and Applications*, vol. 32, no. 14, pp. 9915–9928, 2020.

[146] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[147] D. Karimi, H. Dou, and A. Gholipour, "Medical image segmentation using transformer networks," *IEEE Access*, vol. 10, pp. 29322–29332, 2022.

[148] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, no. 2, pp. 373–440, 2020.

[149] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Honolulu, Hawaii), pp. 1568–1576, 2017.

[150] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, "Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring," *arXiv preprint arXiv:1911.09785*, 2019.

[151] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.

[152] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*, (Virtual), pp. 1597–1607, PMLR, 2020.

[153] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (Virtual), pp. 9729–9738, 2020.

[154] N. A. Koohbanani, B. Unnikrishnan, S. A. Khurram, P. Krishnaswamy, and N. Rajpoot, "Self-path: Self-supervision for classification of pathology images with limited annotations," *IEEE Transactions on Medical Imaging*, vol. 40, no. 10, pp. 2845–2856, 2021.

[155] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th International Conference on Machine Learning*, (Helsinki, Finland), pp. 1096–1103, 2008.

[156] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[157] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Doha, Qatar), pp. 1532–1543, 2014.

[158] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a" siamese" time delay neural network," *Advances in Neural Information Processing Systems*, vol. 6, 1993.

[159] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[160] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[161] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma, "Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications," *arXiv preprint arXiv:1701.05517*, 2017.

[162] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[163] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[164] X. Chen, M. Ding, X. Wang, Y. Xin, S. Mo, Y. Wang, S. Han, P. Luo, G. Zeng, and J. Wang, "Context autoencoder for self-supervised representation learning," *arXiv preprint arXiv:2202.03026*, 2022.

[165] I. Misra and L. v. d. Maaten, "Self-supervised learning of pretext-invariant representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (Virtual), pp. 6707–6717, 2020.

[166] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *arXiv preprint arXiv:2005.00341*, 2020.

[167] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, (New York, NY, USA), pp. 1735–1742, IEEE, 2006.

[168] W. Liu, D. Ferstl, S. Schulter, L. Zebedin, P. Fua, and C. Leistner, "Domain adaptation for semantic segmentation via patch-wise contrastive learning," *arXiv preprint arXiv:2104.11056*, 2021.

[169] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, (Tokyo, Japan), pp. 539–546, IEEE, 2005.

[170] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Boston, Massachusetts), pp. 815–823, 2015.

[171] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[172] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[173] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21271–21284, 2020.

[174] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, (Munich, Germany), pp. 132–149, 2018.

[175] J. Xie, X. Zhan, Z. Liu, Y.-S. Ong, and C. C. Loy, "Delving into inter-image invariance for unsupervised visual representations," *International Journal of Computer Vision*, vol. 130, no. 12, pp. 2994–3013, 2022.

[176] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, (Glasgow, UK), pp. 776–794, Springer, 2020.

[177] G. French, S. Laine, T. Aila, M. Mackiewicz, and G. Finlayson, "Semi-supervised semantic segmentation needs strong, varied perturbations," *arXiv preprint arXiv:1906.01916*, 2019.

[178] Z. Ke, D. Qiu, K. Li, Q. Yan, and R. W. Lau, "Guided collaborative training for pixel-wise semi-supervised learning," in *European Conference on Computer Vision*, (Glasgow, UK), pp. 429–445, Springer, 2020.

[179] J. Li, S. Yang, X. Huang, Q. Da, X. Yang, Z. Hu, Q. Duan, C. Wang, and H. Li, "Signet ring cell detection with a semi-supervised learning framework," in *International Conference on Information Processing in Medical Imaging*, (Hong Kong, China), pp. 842–854, Springer, 2019.

[180] Z. Lai, C. Wang, Z. Hu, B. N. Dugger, S.-C. Cheung, and C.-N. Chuah, "A semi-supervised learning for segmentation of gigapixel histopathology images from brain tissues," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, (Mexico), pp. 1920–1923, IEEE, 2021.

[181] H.-T. Cheng, C.-F. Yeh, P.-C. Kuo, A. Wei, K.-C. Liu, M.-C. Ko, K.-H. Chao, Y.-C. Peng, and T.-L. Liu, "Self-similarity student for partial label histopathology image segmentation," in *European Conference on Computer Vision*, (Glasgow, UK), pp. 117–132, Springer, 2020.

[182] M. Chen, H. Xue, and D. Cai, "Domain adaptation for semantic segmentation with maximum squares loss," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (Seoul, Korea), pp. 2090–2099, 2019.

[183] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, (Miami, FL, USA,), pp. 248–255, IEEE, 2009.

[184] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9912–9924, 2020.

[185] C. Lu, D. Romo-Bucheli, X. Wang, A. Janowczyk, S. Ganesan, H. Gilmore, D. Rimm, and A. Madabhushi, "Nuclear shape and orientation features from h&e images predict survival in early-stage estrogen receptor-positive breast cancers," *Laboratory Investigation*, vol. 98, no. 11, pp. 1438–1448, 2018.

[186] J. N. Kather, L. R. Heij, H. I. Grabsch, C. Loeffler, A. Echle, H. S. Muti, J. Krause, J. M. Niehues, K. A. Sommer, P. Bankhead, *et al.*, "Pan-cancer image-based detection of clinically actionable genetic alterations," *Nature Cancer*, vol. 1, no. 8, pp. 789–799, 2020.

[187] J. A. Diao, J. K. Wang, W. F. Chui, V. Mountain, S. C. Gullapally, R. Srinivasan, R. N. Mitchell, B. Glass, S. Hoffman, S. K. Rao, *et al.*, "Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes," *Nature Communications*, vol. 12, no. 1, pp. 1–15, 2021.

[188] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[189] WHO, "Oral health, https://www.who.int/news-room/fact-sheets/detail/oral-health," 2023.

[190] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, and A. Jemal, "Global cancer statistics, 2012," *CA: A Cancer Journal for Clinicians*, vol. 65, no. 2, pp. 87–108, 2015.

[191] D. C. França, L. M. Monti, A. L. De Castro, A. M. Soubhia, L. E. Volpato, S. M. Á. de Aguiar, and M. C. Goiato, "Unusual presentation of oral squamous cell carcinoma in a young woman," *Sultan Qaboos University Medical Journal*, vol. 12, no. 2, p. 228, 2012.

[192] A. Wright and M. Shear, "Epithelial dysplasia immediately adjacent to oral squamous cell carcinomas," *Journal of Oral Pathology & Medicine*, vol. 14, no. 7, pp. 559–564, 1985.

[193] P. M. Speight, S. A. Khurram, and O. Kujan, "Oral potentially malignant disorders: risk of progression to malignancy," *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, vol. 125, no. 6, pp. 612–627, 2018.

[194] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol, *et al.*, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *JAMA*, vol. 318, no. 22, pp. 2199–2210, 2017.

[195] S. U. Akram, T. Qaiser, S. Graham, J. Kannala, J. Heikkilä, and N. Rajpoot, "Leveraging unlabeled whole-slide-images for mitosis detection," in *Computational Pathology and Ophthalmic Medical Image Analysis: First International Workshop, COMPAY 2018, and 5th International Workshop, OMIA 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16-20, 2018, Proceedings 5*, (Granada, Spain), pp. 69–77, Springer, 2018.

[196] J. Gamper, N. A. Koohbanani, K. Benet, A. Khuram, and N. Rajpoot, "Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification," in *European Congress on Digital Pathology*, (Warwick, UK), pp. 11–19, Springer, 2019.

[197] D. Adel, J. Mounir, M. El-Shafey, Y. A. Eldin, N. El Masry, A. AbdelRaouf, and I. S. Abd Elhamid, "Oral epithelial dysplasia computer aided diagnostic approach," in *2018 13th International Conference on Computer Engineering and Systems (ICCES)*, (Cairo, Egypt), pp. 313–318, IEEE, 2018.

[198] D. K. Das, S. Bose, A. K. Maiti, B. Mitra, G. Mukherjee, and P. K. Dutta, "Automatic identification of clinically relevant regions from oral tissue histological images for oral squamous cell carcinoma diagnosis," *Tissue and Cell*, vol. 53, pp. 111–119, 2018.

[199] F. Dost, K. Le Cao, P. Ford, C. Ades, and C. Farah, "Malignant transformation of oral epithelial dysplasia: a real-world evaluation of histopathologic grading," *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, vol. 117, no. 3, pp. 343–352, 2014.

[200] R. Nag, J. Chatterjee, R. R. Paul, M. Pal, and R. K. Das, "Nuclear segmentation and its quantification in h&e stained images of oral precancer to detect its malignant potentiality," in *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, (Coimbatore, India), pp. 1–6, IEEE, 2018.

[201] M. M. Sami, M. Saito, S. Muramatsu, H. Kikuchi, and T. Saku, "A computer-aided distinction method of borderline grades of oral cancer," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 93, no. 8, pp. 1544–1552, 2010.

[202] A. B. Silva, A. S. Martins, T. A. A. Tosta, L. A. Neves, J. P. S. Servato, M. S. de Araújo, P. R. de Faria, and M. Z. do Nascimento, "Computational analysis of histological images from hematoxylin and eosin-stained oral epithelial dysplasia tissue sections," *Expert Systems with Applications*, vol. 193, p. 116456, 2022.

[203] C. Gilvetti, C. Soneji, B. Bisase, and A. W. Barrett, "Recurrence and malignant transformation rates of high grade oral epithelial dysplasia over a 10 year follow up period and the influence of surgical intervention, size of excision biopsy and marginal clearance in a uk regional maxillofacial surgery unit," *Oral Oncology*, vol. 121, p. 105462, 2021.

[204] H. Mahmood, M. Bradburn, N. Rajpoot, N. M. Islam, O. Kujan, and S. A. Khurram, "Prediction of malignant transformation and recurrence of oral epithelial dysplasia using architectural and cytological feature specific prognostic models," *Modern Pathology*, vol. 35, no. 9, pp. 1151–1159, 2022.

[205] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *International Conference on Machine Learning*, (Stockholm Sweden), pp. 2127–2136, PMLR, 2018.

[206] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Honolulu, Hawaii), pp. 4700–4708, 2017.

[207] C. P. Gan, B. K. B. Lee, S. H. Lau, T. G. Kallarakkal, Z. M. Zaini, B. K. W. Lye, R. B. Zain, H. P. Sathasivam, J. P. S. Yeong, N. Savelyeva, *et al.*, "Transcriptional analysis highlights three distinct immune profiles of high-risk oral epithelial dysplasia," *Frontiers in Immunology*, vol. 13, p. 954567, 2022.

[208] S. G. Fitzpatrick, K. S. Honda, A. Sattar, and S. A. Hirsch, "Histologic lichenoid features in oral dysplasia and squamous cell carcinoma," *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, vol. 117, no. 4, pp. 511–520, 2014.

[209] E. A. Georgakopoulou, M. D. Achtari, M. Achtaris, P. G. Foukas, and A. Kotsinas, "Oral lichen planus as a preneoplastic inflammatory model," *Journal of Biomedicine and Biotechnology*, vol. 2012, 2012.

[210] I. Chami, S. Abu-El-Haija, B. Perozzi, C. Rac, and K. Murphy, "Machine learning on graphs: A model and comprehensive taxonomy," *Journal of Machine Learning Research*, vol. 23, no. 89, pp. 1–64, 2022.

[211] L. P. Chew, "Constrained delaunay triangulations," in *Proceedings of the third Annual Symposium on Computational Geometry*, (New York, NY, USA), pp. 215–222, 1987.

[212] M. Dawood, M. Eastwood, M. Jahanifar, L. Young, A. Ben-Hur, K. Branson, L. Jones, N. Rajpoot, and F. u. A. A. Minhas, "Data-driven modelling of gene expression states in breast cancer and their prediction from routine whole slide images," *bioRxiv*, pp. 2023–04, 2023.

[213] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[214] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[215] N. Das, E. Hussain, and L. B. Mahanta, "Automated classification of cells into multiple classes in epithelial tissue of oral squamous cell carcinoma using transfer learning and convolutional neural network," *Neural Networks*, vol. 128, pp. 47–60, 2020.

[216] K. Sakamoto, T. Ikeda, *et al.*, "Deep-learning application for identifying histological features of epithelial dysplasia of tongue," *Journal of Oral and Maxillofacial Surgery, Medicine, and Pathology*, vol. 34, no. 4, pp. 514–522, 2022.

[217] Y. Liu, E. Bilodeau, B. Pollack, and K. Batmanghelich, "Automated detection of premalignant oral lesions on whole slide images using convolutional neural networks," *Oral Oncology*, vol. 134, p. 106109, 2022.

[218] X. Zhang, F. O. Gleber-Netto, S. Wang, R. R. Martins-Chaves, R. S. Gomez, N. Vigneswaran, A. Sarkar, W. N. William Jr, V. Papadimitrakopoulou, M. Williams, *et al.*, "Deep learning-based pathology image analysis predicts cancer progression risk in patients with oral leukoplakia," *Cancer Medicine*, 2023.

[219] R. Ellonen, A. Suominen, J. Kelppe, J. Willberg, J. Rautava, and H. Laine, "Histopathological findings of oral epithelial dysplasias and their relation to malignant transformation," *Cancer Treatment and Research Communications*, vol. 34, p. 100664, 2023.

[220] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (New Orleans, Louisiana, USA), pp. 10684–10695, 2022.

[221] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, pp. 41–75, 1997.

[222] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, 2013.