WARWICK
THE UNIVERSITY OF WARWICK

A Thesis Submitted for the Degree of PhD at the University of Warwick

Permanent WRAP URL:

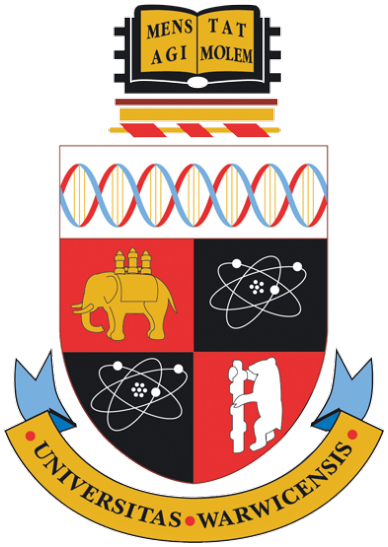
http://wrap.warwick.ac.uk/159706

Copyright and reuse:
This thesis is made available online and is protected by original copyright.
Please scroll down to view the document itself.
Please refer to the repository record for this item for information to help you to cite it.
Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

warwick.ac.uk/lib-publications

WARWICK
THE UNIVERSITY OF WARWICK



Spatial Context in Computational Pathology

by

Muhammad Shaban

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Computer Science

December 2020

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 would like to thank Prof. Nasir M. Rajpoot, my PhD supervisor, for his constant support over the years. I am genuinely grateful to him for his supervision, especially for the one to one supervisory meetings, which helped me to polish my academic and professional skills. His invaluable comments and suggestions have contributed a lot to the success of my research projects. Besides my supervisor, I would like to thank my PhD advisors, Dr Victor Sanchez and Dr.Till Bretschneider, for their constructive feedback in annual review meetings. I am grateful to my external and internal examiners, Dr. Constantino Carlos and Dr. Abhir Bhalerao for their valuable time and insightful comments on my thesis.

I would also like to thank all the faculty members and research staff in Tissue Image Analytics (TIA) Lab at the University of Warwick for their constructive criticism on my research work during lab presentations which helped me to improve both my research work and presentation skills. I personally thank Prof. David Epstein for all our on and off discussions and especially his moral support in the ProBe project. I am also grateful to Dr Shan E Ahmed Raza for his keen interest in my research work and insightful comments on my thesis and paper write-up. A special thanks to Dr Moazam Fraz for his brotherly guidance in general and help in paper write-ups in particular during his stay at TIA Lab.

I am also grateful to all my clinical collaborators for curating datasets, marking annotations and explaining the histology of different cancers which pave the way for the success of my research projects. I would like to give my special thanks to Dr Syed Ali Khurram for his patience on my non-stop

data annotation requests on short notices. I would also like to thank Dr Ayesha Azam, Dr Hanya Mahmood, and Dr Yee Wah Tsang for their active collaboration during my PhD.

I would like to thank all of the current and previous TIA Lab members for their help and support during my PhD. I want to mention a few of the many reasons for my special thank to some of the lab members: Mike, Talha, and Navid for accompanying me in sports activities; Najah, Simon, Navid, Ruqayya, and Saad for their active collaboration; Talha and Ruqayya for their moral support over the years; Jevgenij for his critical questions that often made me think hard; John for helping me in setup related issues; Navid, Rawan, and especially Ruqayya for proofreading my write-ups; Talha, Hammam and Saad for accompanying me in the prayer hall; Hammam for countless pick and drops to attend get-togethers; Hammam, Rawan, and Saad for taking my help in their projects.

Last but not least, I would like to thank my family for their moral support not just for PhD but for my whole educational journey. I am grateful to my mother for her continued prayers, my sisters for their unconditional love, and my brothers for their financial support and confidence in my abilities. A special thanks to my nephews and nieces, who helped me to cope with homesickness through countless hours of video calls and entertained me with their naive questions regarding my surroundings. I would like to thank all those who supported me during my PhD. More importantly, I want to thank my wife and in-laws for their trust and unconditional support.

I dedicate this thesis to a man who envisioned my PhD years ago but left me halfway through the PhD. He was my motivation; he was my **father**.

Sponsorship and Grants

I would like to acknowledge the support by Engineering and Physical Sciences Research Council (EPSRC) grant 1829583, and the partial financial support from the Department of Computer Science, University of Warwick.

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.

Muhammad Shaban
15th December 2020

Publications

First-Authored Publications

Journal Articles

- **M. Shaban**, R. Awan, M. M. Fraz, A. Azam, Y. Tsang, D. Snead, and N. M. Rajpoot, "Context-aware convolutional neural network for grading of colorectal cancer histology images," IEEE Transactions on Medical Imaging, vol. 39, no. 7, pp. 2395–2405, 2020

- **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

Co-Authored Publications

Journal Articles

- H. Mahmood, **M. Shaban**, B. I. Indave, A. R. Santos-Silva, N. M. Rajpoot, S. A. Khurram "Use of Artificial Intelligence in Diagnosis of Head and Neck Precancerous and Cancerous Lesions: A Systematic Review" Oral Oncology, 110, (2020): 104885.

- S. E. A. Raza, L. Cheung, **M. Shaban**, S. Graham, D. Epstein, S. Pelengaris, M. Khan, and N. M. Rajpoot, "Micro-net: A unified model for segmentation of various objects in microscopy images," Medical image analysis, vol. 52, pp. 160–173, 2019

- M. Fraz, S. Khurram, S. Graham, **M. Shaban**, M. Hassan, A. Loya, and N. Rajpoot, "Fabnet: feature attention-based network for simultaneous segmentation of microvessels and nerves in routine histology images of oral cancer," Neural Computing and Applications, pp. 1–14, 2019

- Q. D. Vu, S. Graham, T. Kurc, M. N. N. To, **M. Shaban**, T. Qaiser, N. A. Koohbanani, S. A. Khurram, J. Kalpathy-Cramer, T. Zhao, et al., "Methods for segmentation and classification of digital microscopy tissue images," Frontiers in bioengineering and biotechnology, vol. 7, p. 53, 2019

Conference and Workshop Papers

- N. Alsubaie, **M. Shaban**, D. Snead, A. Khurram, and N. Rajpoot, "A multi-resolution deep learning framework for lung adenocarcinoma growth pattern classification," in Annual Conference on Medical Image Understanding and Analysis, pp. 3–11, Springer, 2018

- S. Graham, **M. Shaban**, T. Qaiser, S. A. Khurram, and N. Rajpoot, "Classification of lung cancer histology images using patch-level summary statistics," in Medical Imaging 2018: Digital Pathology, vol. 10581,p. 1058119, International Society for Optics and Photonics, 2018

- M. Fraz, **M. Shaban**, S. Graham, S. A. Khurram, and N. M. Rajpoot, "Uncertainty driven pooling network for microvessel segmentation in routine histology images," in Computational pathology and ophthalmic medical image analysis, pp. 156–164, Springer, 2018

- R. Awan, N. A. Koohbanani, **M. Shaban**, A. Lisowska, and N. Rajpoot, "Context-aware learning using transferable features for classification of breast cancer histology images," in International Conference Image Analysis and Recognition, pp. 788–795, Springer, 2018

- 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, p. 1132011, International Society for Optics and Photonics, 2020

- N. A. Koohbanani, T. Qaisar, **M. Shaban**, J. Gamper, and N. Rajpoot,"Significance of hyperparameter optimization for metastasis detection in breast histology images," in Computational Pathology and Ophthalmic Medical Image Analysis, pp. 139–147, Springer, 2018

- Y. Zhou, S. Graham, N. Alemi Koohbanani, **M. Shaban**, P. A. Heng, and N. Rajpoot, "CGC-Net: Cell graph convolutional network for grading of colorectal cancer histology images," in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019

Abstract

In recent years, computational pathology has emerged as a discipline representing big-data based approaches for the diagnosis and prognosis of cancer patients using different sources of data, mainly digitised histology images and clinical information. A plethora of computational methods have been developed for fast and reproducible diagnosis and prognosis of cancer, lately dominated by deep learning based methods. However, current deep learning methods do not incorporate the whole spatial landscape of histology images due to limited computational and memory resources. In this thesis, I develop deep learning based methods which incorporate the broader spatial context of histology images for cancer diagnosis and prognosis.

I propose a novel framework to incorporate large contextual information inheritably available in histology images by a context-aware neural network. The proposed framework first encodes the local representation of an input image into low dimensional features then aggregates the features by considering their spatial organization to make a final prediction. The framework is designed for a set of histology problems which requires both high-resolution appearance of tissue along with large contextual information such as colorectal grading, and growth pattern classification. I have also proposed two novel objective measures for the quantification of tumour microenvironment of head and neck squamous cell carcinoma (HNSCC) patients for their better stratification and prognostication. The first measure quantifies the tumour infiltrating lymphocytes abundance (TILAb) whereas the second one is for the quantification of tumour-associated stroma infiltrating lymphocytes to tumour-associated stroma ratio (TASIL-Ratio). Both TILAb and TASIL-Ratio based scores show prognostic significance similar to manual scores but with the added advantages of a more rapid and objective quantification.

Abbreviations

AUC : Area Under the Curve

BAM : Best Alignment Metric

CB : Context Block

CI : Confidence Interval

C-Index : Concordance Index

CNNs : Convolutional Neural Networks

CRA : Colorectal Adenocarcinoma

CRF : Conditional Random Field

CSNet : Coarse Segmentation Network

DFS : Disease-free Survival

DSS : Disease-specific Survival

H&E : Haemotoxylin and Eosin

H&N : Head and Neck

HNSCC : Head and Neck Squamous Cell Carcinoma

HPV : Human Papillomavirus

HR : Hazard Ratio

L-Percentage : Lymphocyte Percentage

LR : Logistic Regression

LR-CNN : Local Representation Convolutional Neural Network

LSTM : Long Short Term Memory

LTAS-Col : Colocalisation of Lymphocyte and Tumour-associated Stroma

LTAS-Ratio : Lymphocyte to Tumour-associated Stroma Ratio

LT-Col : Colocalisation of Lymphocyte and Tumour

LT-Ratio : Lymphocyte to Tumour Ratio

MH : Morisita-Horn

Non-ROI : Non Regions of Interest

OPSCC : Oropharyngeal Squamous Cell Carcinoma

OSCC : Oral Squamous Cell Carcinoma

RA-CNN : Representation Aggregation Convolutional Neural Network

RNN : Recurrent Neural Network

SC : Skip Connections

SD : Shannon Diversity

SKMCH&RC : Shaukat Khanum Memorial Cancer Hospital and Research Centre

SVM : Support Vector Machine

TASILAb : Tumour-associated Stroma Infiltrating Lymphocytes Abundance

TASIL-Ratio : Tumour-associated Stroma Infiltrating Lymphocytes to tumour-associated stroma Ratio

TAS-Percentage : Tumour-associated Stroma Percentage

TAST-Col : Colocalisation of Tumour-associated Stroma and Tumour

TAST-Ratio : Tumour-associated Stroma to Tumour Ratio

TILAb : Tumour Infiltrating Lymphocytes Abundance

TIL-Ratio : Tumour Infiltrating Lymphocytes to tumour Ratio

TILs : Tumour Infiltrating Lymphocytes

TME : Tumour Microenvironment

TNM : Tumour, Node and Metastasis

T-Percentage : Tumour Percentage

TRC : Tissue Region Classifier

WSIs : Whole Slide Images

Chapter 1

Introduction

1.1 Cancer

Cancer is the common name of a set of diseases that cause abnormal growth of a human body's cells and their spread into neighbouring tissues. If this spread is not controlled in time, then it can result in death [3] . Cancer can start almost anywhere in the human body, even in blood and may involve any type of cells. A tumour is a cancer of solid tissue which can be benign or malignant [3]. A benign tumour is noncancerous, and it usually stays at the place of origin and does not spread to the neighbouring tissues. However, a malignant tumour can invade into other solid tissues such as organs, muscles or bones. The tumour invasion to distant tissues is known as metastasis. A timely diagnosis of malignant tumour and the right treatment plan is a key factor in disease-free and long term survival of a cancer patient [4].

Cancer can be diagnosed in different ways which include physical examination, laboratory tests, imaging tests, biopsies, and resections [5] . However, a biopsy and resection are the most reliable and accurate ways for cancer diagnosis. Tissue sample obtained by a biopsy or resection is sliced to be placed on glass slides, stained to enhance the contrast of different tissue regions, and analysed under a microscope by an expert pathologist for cancer diagnosis. Apart from the binary diagnosis, a pathologist also determines the aggressiveness/grade of cancer in case of cancerous tissue. After diagnosis, treatment comes as the next step, such as removing cancerous regions, reducing the chance of cancer recurrence, and improving the patient's life quality. Selection of the right type of treatment depends on many factors, such as the site of origin, histological types, and grade of cancer.

1.2 Cancer Types

Cancers are mainly categorised in two different ways, site of origin and histological type. The site of origin based categorisation depends on the primary location of the tumour where it first appears, such as oral, breast, lung and colon. The histological type based categorisation lead to hundreds of different cancer types due to number of different cell types, tissues, and organs. However, these cancer types are grouped into six main categories according to the International Classification of Diseases for Oncology, Third Edition [6]. Cancerous tumours of epithelial origin fall under the carcinoma category. Carcinoma accounts for $80 - 90\%$ of all the cancer cases [7]. Adenocarcinoma and squamous cell carcinoma are the two main subtypes of carcinoma. Sarcoma is the second category which encompasses the tumours of supportive and connective tissues such as bones, muscle, and fat. Myeloma and Leukemia categories deal with bone marrow tumours of plasma cells and white blood cells, respectively. Malignancies of the lymphatic system come under the Lymphoma categories. Finally, all those tumours that lie under two or more categories are grouped under the mixed type category such as adenosquamous carcinoma, teratocarcinoma, and carcinosarcoma. Slightly more detailed histology of colorectal adenocarcinoma (CRA) and head & neck squamous cell carcinoma (HNSCC) is presented in the following sections as digital histological profiling of these two types of cancers is studied in this thesis.

1.2.1 Colorectal Adenocarcinoma

Colorectal cancer, also known as bowel cancer, is the second deadliest and fourth most common cancer in the United Kingdom [8]. CRA is the most dominant type of colorectal cancer, which accounts for more than 90% of colorectal cases [9]. It originates from epithelial cells in the lining of the colon or rectum. The aggressiveness or growth of CRA is mainly determined by its grade and/or stage. Early diagnosis and appropriate treatment are vital for the long-term survival of CRA patients. However, CRA detection at a higher grade and stage is one of the major causes of most CRA related deaths [10] as high-grade cancer cells tend to grow and spread more quickly than low-grade cancer cells.

The CRA grade is mainly based on cell aberrance and the morphology of glandular structure in the lining of the bowel whereas CRA stage is defined by the size of the tumour at the primary site and spread of the tumour to other sites. In a normal case, all glands appear as either round or elliptical depending on the section of tissue (see Figure 1.1 a-b). The American Joint Committee on Cancer [11] described a four-tier grading system for CRA. First, the well-differentiated grade is when all the glands are more normal like

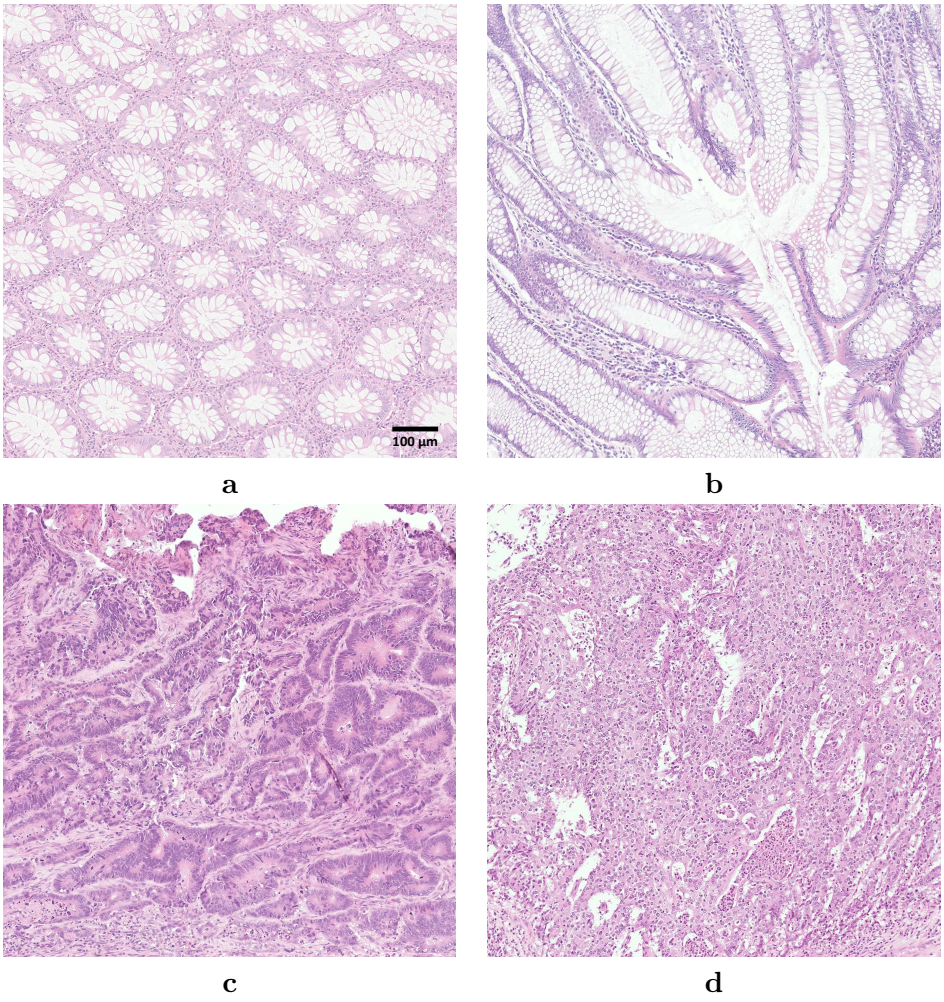


Figure 1.1: (**a, b**) Images of normal colorectal tissue. (**c-d**) Images of low and high grade colorectal adenocarcinoma tissues.

morphological structure, and cell appearance is also close to normal. Second, the moderately differentiated grade consists of glands with moderate aberrance in glandular structure and cell appearance, but more than 50% glands preserve their glandular morphology. Third, the poorly differentiated grade is assigned to those cases where more than 50% of glands have loosed their structure and cells become abnormal. Finally, the undifferentiated grade is when glands appear as a sheet of abnormal cells instead of their typical glandular structure.

The task of cancer grading is subjective in nature and thus suffer from inter- and intra-observer variability. Some studies [12, 13] suggested a two-tier grading system to reduce the subjectivity. They merged the well and moderately differentiated grades into one grade (see Figure 1.1a-b) and named it as *low* grade, whereas *high* grade consists of poorly and undifferentiated grades (see Figure 1.1c-d). The tumour, node and metastasis (TNM) staging system [11] is another way to measure the severity of CRA. Each parameter in

TNM staging is further divided into sub-stages. The T sub-stage of tumour represents its size; N sub-stage describes the number of lymph nodes invaded by tumour and the M sub-stage indicates the metastatic spread of CRA to other organs. Grading requires biopsy or resection samples only from the primary site as compared to TNM staging, which requires samples from lymph nodes and distant organs as well as the primary site.

Although both grading and TNM staging have shown prognostic signification for CRA [9, 14, 15], the issue of subjectivity is still there. An objective assessment of CRA may lead to better prognostic analysis. Moreover, the whole process of CRA grading and staging is tedious and labour intensive, which needs to be assisted by an automated method. Therefore, an objective computer-aided method can address the subjectivity issue with objective grading and staging with the added advantage of rapid assessment of CRA cases.

1.2.2 Head & Neck Squamous Cell Carcinoma

HNSCC comprises of the tumours of oral cavity, pharynx, or larynx (Figure 1.2). It is the eighth most common cancer in the United Kingdom [16] and accounts for 2% of all cancer-related deaths. Oral and oropharyngeal subtypes, which are mainly studied in this thesis, are briefly described in the following sections.

Oral Squamous Cell Carcinoma

Oral squamous cell carcinoma (OSCC) is the most common malignancy of the head and neck (H&N) region [17] in both males (42%) and females (46%). The OSCC prevalence is almost twice as common in males and three times more in females than the next most common cancer, which is laryngeal squamous cell carcinoma (26% in males, 13% in females) [18]. OSCC is associated with invasion and destruction of local tissues and maxillofacial bones with significant associated morbidity. In addition to early recurrence, frequent lymph nodes metastasis and extranodal extension [19] are further challenges in the management of OSCC patients. The high morbidity and mortality rates in OSCC patients [20, 21] highlight the need for an objective and quantitative analysis of any potential prognostic markers to help identify tumours which may respond poorly to therapy [22].

Unlike CRA, grading of OSCC is mainly based on the amount of keratin within the tumour [23] where low grade and high grade correspond to more than 20% and less than 20% of keratin, respectively [24]. However, the OSCC TNM based staging system is similar to CRA. The TNM staging has been used for the treatment planning of OSCC, but the prognostic significance of some

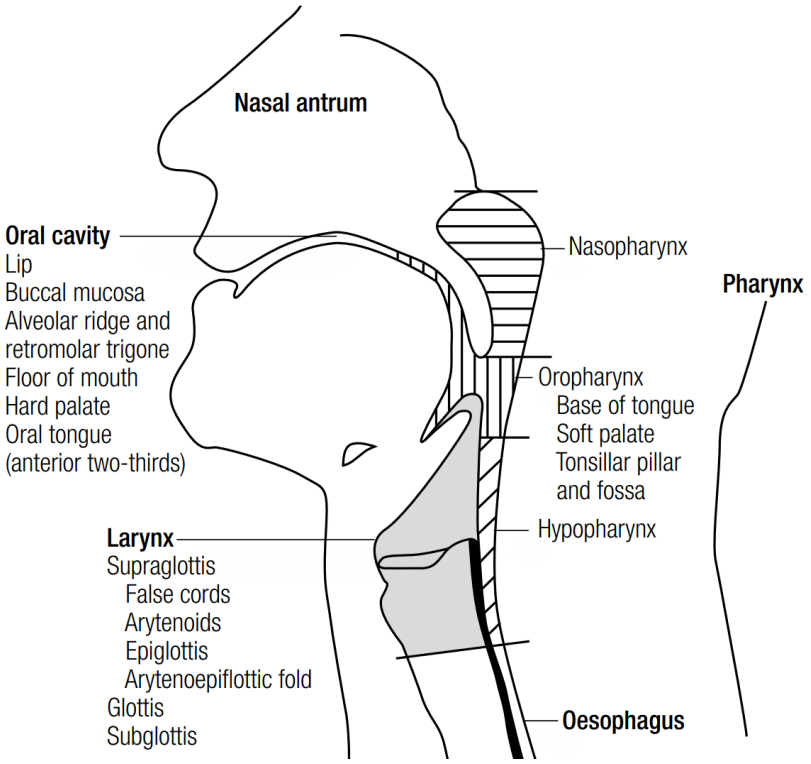


Figure 1.2: Anatomical sites and subsites of the H&N cancer [1].

stages is not consistent [25]. The pattern of tumour invasion is another system for histological classification of OSCC [26] and the patterns are determined based on the four features; the lymphocytic infiltration, pattern of invasion, keratinisation, and nuclear pleomorphism [27, 28]. Some of these features have shown prognostic significance for OSCC such lymphocyte infiltration in tumour [29, 30] and pattern of invasion [31, 32]. Lymphocytes are the white blood cells and part of human immune system. Tumour Infiltrating Lymphocytes (TILs) are the lymphocytes which are moved from the blood into the tumour regions. TILs are normally quantified into four groups: absent, low, moderate and high. Their quantification is conducted by eyeballing over haemotoxylin and eosin (H&E) stained histology tissue by an expert pathologist (Figure 1.3). However, the quantification process is also subjective and prone to inter- and intra-observer variability, just like cancer grading. Therefore, a computer-based automated method is required to eliminate the subjectivity in the quantification of lymphocytic infiltration.

Oropharyngeal Squamous Cell Carcinoma

The incidence rate of oropharyngeal squamous cell carcinoma (OPSCC) is increasing in many developed countries [33, 34] such as the United Kingdom where the number of OPSCC cases has doubled between 1990-2006 and 2006-

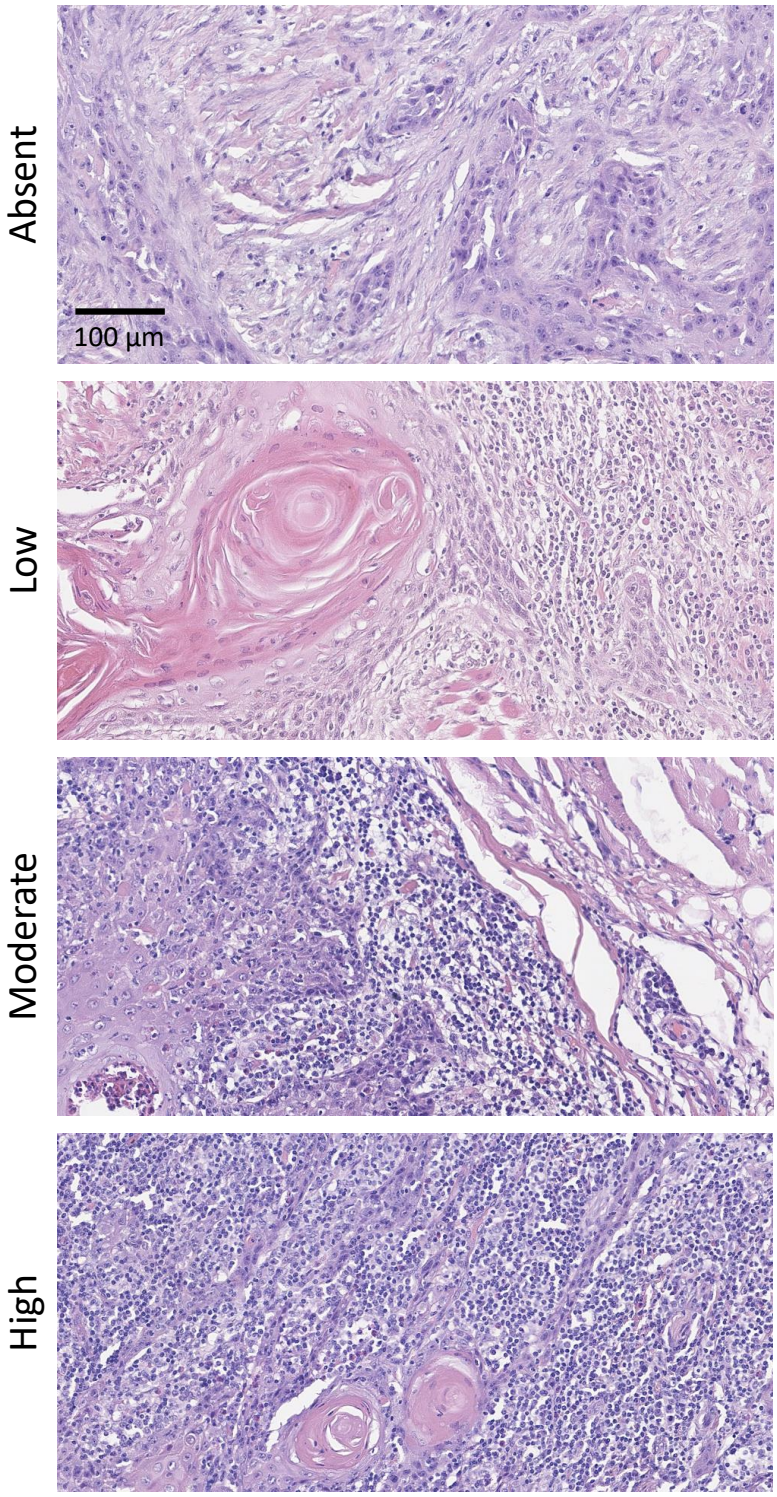


Figure 1.3: Illustration of the degrees of tumour infiltrating lymphocytes in oral squamous cell carcinoma. Each image represents a part of tumour region where the number of lymphocytes, the round shaped purple dots, in each image represent the degree of infiltration in tumour regions.

2010. Human papillomavirus (HPV) infection is one of the causes of this rapid rise, and it accounts for 70% of OPSCC cases. HPV status of OPSCC cases is a strong prognostic indicator for survival analysis. However, the HPV test requires immunohistochemistry stained slides which are not used in routine histology diagnosis. TILs have also been studied for the prognosis of OPSCC [35, 36]; however, the quantification process for TILs is same as in OSCC, which suffers from the aforementioned issues. Therefore, a computer-based automated method for TIL quantification is required for OPSCC as well.

1.3 Computational Pathology

The digitisation of histology slides [37] has led to widespread adaptation of whole slide images (WSIs) in digital pathology. A stained tissue mounted on a glass slide is digitally scanned at different resolutions up to 40× to produce WSI (Figure 1.4). This digitisation opens up new avenues of research for computer vision, machine learning and deep learning communities to develop computational methods to quantify and improve cancer treatment procedures. A plethora of computational methods [38–41] have been developed for fast and reproducible diagnosis and prognosis of different types of cancers such as head and neck, colorectal, breast, and lung cancer. Early works on histology image analysis are mainly based on traditional machine learning methods where the research community has proposed numerous handcrafted feature-based methods [39] for cancer diagnosis and prognosis. However, deep learning based methods have been used more frequently in recent years due to the technological developments in computer hardware and availability of large histology dataset [41].

Deep learning methods have significantly improved the state-of-the-art in many natural images based computer vision problems such as visual object detection and recognition [42–45] and scene labelling [46]. Multi-gigapixel WSIs are also amenable to the application of deep learning methods for analysis due to the sheer size of pixel data present in them. WSIs can be readily absorbed by data-hungry deep learning methods to tackle computational pathology problems. However, the processing of WSIs as a whole through a deep learning method is still not possible due to limited computational and memory resources. Most of the deep learning methods rely on a patch-based approach where WSIs are chopped into small manageable patches for processing. The state-of-the-art deep learning based methods for natural image classification and segmentation are quite generic. They may apply to multiple cancer types with slight domain adaptation by model retraining using WSIs of specific cancer types.

Convolutional neural networks (CNNs) are one of the most common deep learning networks for computer vision problems. CNNs have been used for

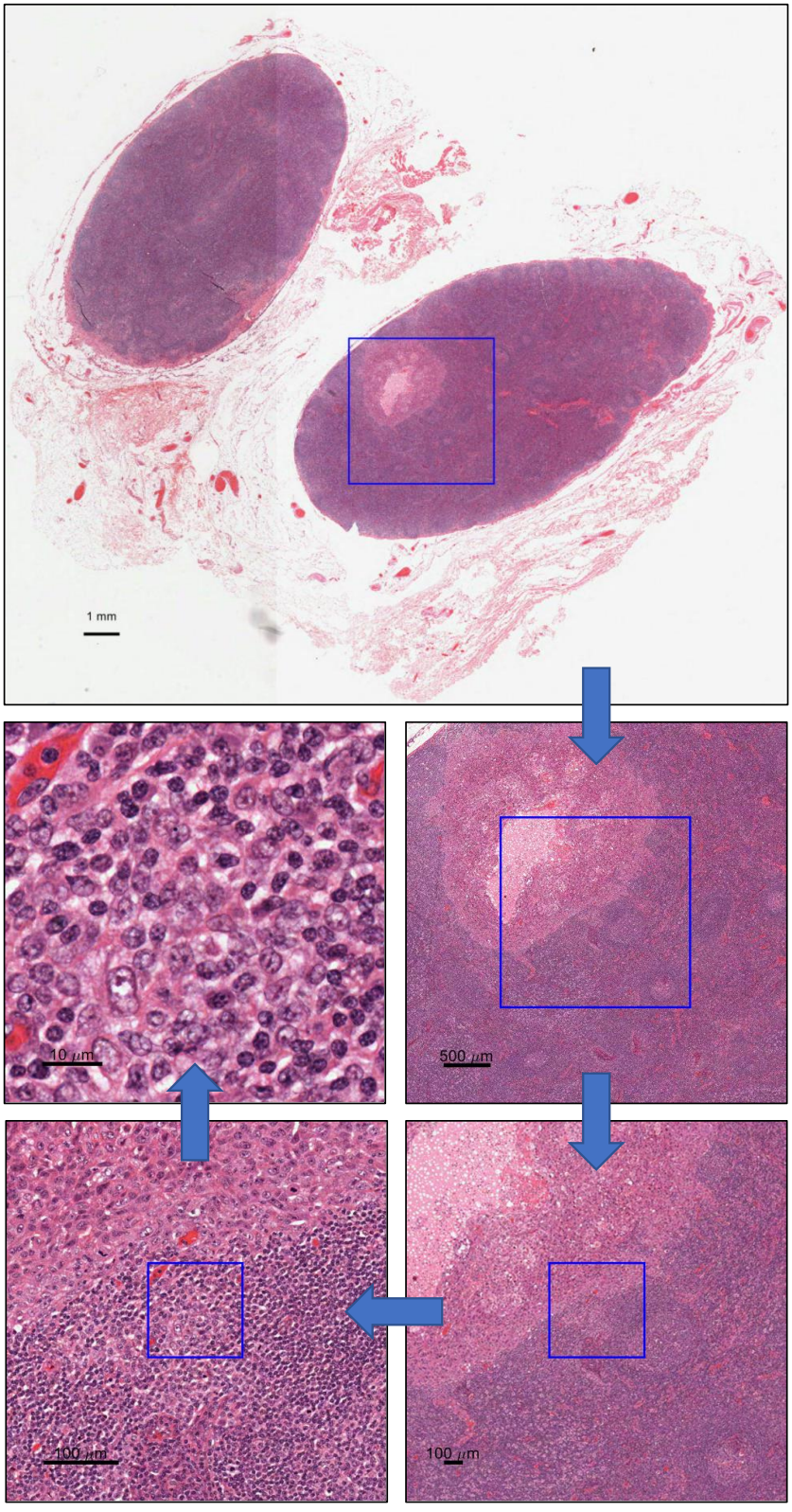


Figure 1.4: Illustration of a multi-gigapixel WSI at different resolutions. Top image presents the low resolution view of a whole slide image whereas each subsequent image represents a higher resolution view of the rectangular region in the previous image.

classification and segmentation of different histology primitives in a WSI. The most common use cases are cell detection and segmentation [47–49], metastasis detection [50–52], gland segmentation [53], cancer grading [54–56] and cancer classification into sub-types [57]. Moreover, CNNs have been used as an intermediate step for high-level histology image analysis tasks such as morphometric analysis [58], mutation prediction [59] and patient survival analysis [60].

1.4 Spatial Analysis

WSIs represent the whole landscape of histology tissues which consists of spatial contextual information of the cellular organisation, gland structure, and tissue architecture. The local context of a cell is important for cell classification, whereas cancer grading requires a broader spatial context as the whole glandular structure defines the CRA cancer grade. Contextual information about overall tissue architecture helps in understanding the tumour microenvironment (TME). Analysis of histology images with limited spatial context may lead to incorrect diagnosis and prognosis in some cases.

1.4.1 Context-aware Analysis

Standard CNN based methods are not capable of capturing the spatial context required for all types of histology image analysis tasks such as cancer grading. These methods can only process moderately sized images due to memory constraints which limits the amount of context captured by each image. Although these methods are trained on tens of thousands of patches extracted from several WSIs, the spatial relationship between these patches is not incorporated during training and inference. Therefore, there is a need for context-aware deep learning methods for the analysis of histology images (see Figure 1.5).

1.4.2 Tumour Microenvironment Analysis

The TME consists of many different cell types (e.g. immune cells and stromal cells) with different biological roles and their unique relationships with cancer cells. Immune cells are part of body defence system where as stromal cells are connective tissue cell. Genomic based analysis of the TME has resulted in several prognostic biomarkers for different cancer types [61–63]. However, information about spatial relationships between tumour and immune cells is not available in genomics based analyses. Histology images have this missing spatial relationship information which may help to understand the TME better. Statistical methods can be employed to quantify spatial patterns in histology

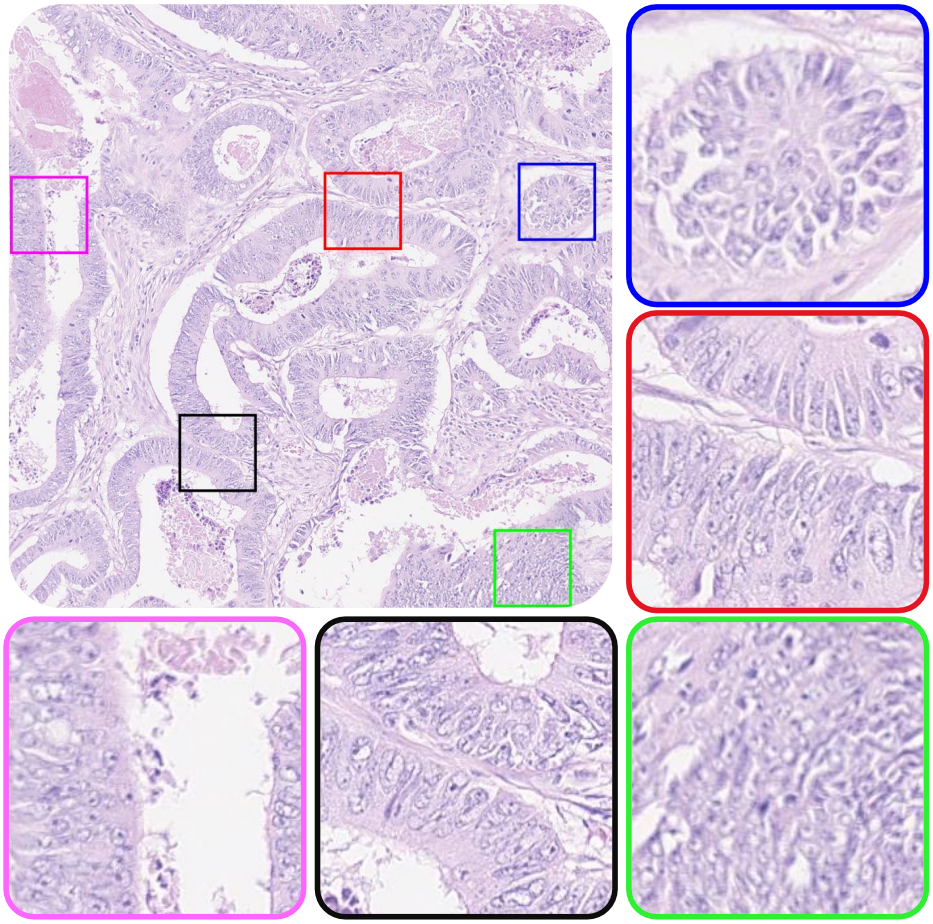


Figure 1.5: Illustration of the importance of spatial context. Top left image presents a colorectal tissue at lower magnification where tube like structures represent glandular morphology which is key for colorectal adenocarcinoma grading. However, the high resolution view of five differently colored regions do not present holistic view of glandular morphology. Colors are used to indicate the corresponding location of each region in the image.

images such as TILs which can be quantified based on the spatial co-occurrence of tumour and lymphocytes. However, there could be some other spatial patterns that need to be explored to extend our understanding of TME and develop better and reproducible digital biomarkers for cancer prognosis.

1.5 Aim of the Thesis

The overarching aim of the thesis to develop novel methods for diagnosis and prognosis of cancer using the spatial contextual information available in cancer histology images. Although the methods proposed in the thesis are generic and can apply to many different cancer types, I only consider CRA and HNSCC for evaluation. I propose different methods to incorporate spatial contextual information for better diagnosis and prognosis of cancer. CNN based novel methods are developed to capture contextual information to improve cancer grading and tissue segmentation. However, a combination of CNN and statistical methods is used to summarise the spatial landscape of histology images for the development of novel digital biomarkers and enhanced cancer prognosis.

1.5.1 Main Contributions

- I propose a novel framework for context-aware learning from large high-resolution input images for CRA grading. I report the results of comprehensive experiments (with 100+ network models) and comparisons to demonstrate the superiority of context-aware learning over traditional patch-based methods. (Chapter 3).

- I propose a novel scoring method for TIL abundance, termed as the TILAb score, to quantify the extent of spatial lymphocytic infiltration in the tumour region which is a combination of lymphocyte to tumour ratio and their statistical colocalisation in a WSI. The reproducibility and objectivity of the TILAb score are investigated by evaluating the prognostic significance of TILAb score for disease-free survival of OSCC patients (Chapter 4).

- I propose a novel coarse segmentation method for segmentation of clinically significant regions which includes segmentation of tumour, tumour-associated stroma, and lymphocytes in a WSI. The proposed method also addresses the issues of limited context and noisy prediction, which occurred in patch-based segmentation. It does not require pixel-level ground truth and can learn from partially annotated images as well (Chapter 5).

- I profile the TME of HNSCC by 13 different quantification methods and explore their prognostic significance for stratification of HNSCC patients into low and high-risk groups. I show that tumour-associated stroma infiltrating lymphocyte based scores have the prognostic ability to be used as novel digital biomarkers for disease-specific survival of HNSCC patients (Chapter 5).

1.6 Thesis Organisation

Chapter 2: Background

A review of existing literature on computational pathology in general and spatial analysis, in particular, is presented in chapter 2. I summarise the existing problem-specific methods for automated grading and HNSCC prognosis. I also review existing methods for context-aware learning and tissue segmentation, and discuss the limitations of existing methods and how these methods are different from the proposed methods.

Chapter 3: Context-Aware Convolutional Neural Network

Digital histology images are amenable to the application of CNNs for analysis due to the sheer size of pixel data present in them. CNNs are generally used for representation learning from small image patches (e.g. 224×224) extracted from digital histology images due to computational and memory constraints. However, standard CNNs based methods do not incorporate high-resolution contextual information in histology images. I propose a novel way to incorporate a broader context by a context-aware neural network based on images with a dimension of 1792×1792 pixels. The proposed framework first encodes the local representation of a histology image into low dimensional features then aggregates the features by considering their spatial organisation to make a final prediction. I evaluate the proposed method on two colorectal cancer datasets for the task of cancer grading and show that our method outperforms traditional patch-based approaches, problem-specific methods, and existing context-based methods. I also present a comprehensive analysis of different variants of the proposed method.

Chapter 4: Spatial Quantification of Tumour Infiltrating Lymphocytes Abundance

The abundance of TILs is a key prognostic indicator in a range of cancers with emerging evidence of its role in OSCC progression and treatment response. However, current methods for TIL quantification are subjective and prone

to variability in interpretation. An automated method for quantification of TIL abundance has the potential to facilitate better stratification and prognostication of oral cancer patients. Chapter 4 presents a novel method for objective quantification of TIL abundance in OSCC histology images. The proposed tumour infiltrating lymphocytes abundance (TILAb) score is calculated by first segmenting the WSIs into underlying tissue types (tumour and lymphocytes) and then quantifying the spatial colocalisation of lymphocytes and tumour regions in a novel fashion. The TILAb score is motivated by the biological definition of TILs as tumour infiltrating lymphocytes, with the added advantages of objective and reproducible quantification. I investigate the prognostic significance of TILAb score on digitised WSIs of H&E stained slides of OSCC patients. I show that the TILAb score is a strong prognostic indicator ($p = 0.0006$) of disease-free survival on our OSCC test cohort. The automated TILAb score has a significantly higher prognostic value than the manual TIL score ($p = 0.0024$).

Chapter 5: Coarse Segmentation for Profiling of Tumour Microenvironment

I propose a new framework for the quantification of three most significant components (tumour, tumour associated stroma and lymphocytes) of TME in HNSCC. A novel coarse segmentation method is proposed to overcome the issue of limited context and noisy predictions for segmentation of TME components in HNSCC WSIs. Unlike patch-based segmentation methods, the proposed method predicts a label for each 32×32 region in a patch of size 256×256, which generates 64 times denser prediction map than a standard patch classifier. The coarse segmentation of HNSCC WSIs through the proposed method is then used for quantification of different spatial patterns of the tumour, tumour-associated stroma, and lymphocytes. I show that our proposed quantification method for the quantification of tumour-associated stroma infiltrating lymphocytes to tumour-associated stroma ratio (TASIL-Ratio) carries prognostic significance (p-value=0.002) for better disease-specific survival of HNSCC patients. The TASIL-Ratio score remains prognostic indicator for disease-specific and disease-free survival of OSCC and OPSCC. I also compared the predictive ability of TASIL-Ratio based survival model with existing quantification methods through concordance index measure where TASIL-Ratio achieved the highest concordance score as compared to its counterparts. The TASIL-Ratio also shows a positive correlation with molecular estimates of CD8 T cells which kill the cancerous cells in the human body.

Chapter 6: Conclusions

A summary of the proposed methods and some potential future directions for each of the proposed methods are presented in the last chapter.

Chapter 2

Background

Histology images represent the high-resolution view of histology landscape which serve as the gold standard for cancer diagnosis and helps to understand tumour microenvironment (TME) for better patient prognosis. Machine learning based methods have been the first choice for the development of histology imaging based computational methods for cancer diagnosis and prognosis. Recently, deep learning, a branch of machine learning, based methods have become the new standard for histology image analysis. In this chapter, I present a brief review of state-of-the-art deep learning methods which are developed for histology image analysis. I mainly focus on the methods which are technically or clinically relevant to the novel methods proposed in the thesis. In technical methods, I consider context-aware learning, automated tissue segmentation, and TME profiling related approaches, whereas, in clinically related works, I review the literature on automated patient prognosis and cancer grading.

2.1 Context-Aware Learning

In literature, various approaches have been presented to incorporate the contextual information for the classification of histology images. Most common approaches for context-aware learning include image downsampling, use of multi-resolution images, and use of two-step methods. A summary of these approaches is presented in the following sections.

2.1.1 Image Downsampling

Histology images are large images (usually aroung 200000×100000 pixels) which do not fit in the memory of a graphic processing unit. Therefore, image downsampling is the most straightforward approach to fit these images into memory to capture the context from these large histology images. Several studies [64–66] have followed this approach as it is also a common practice in natural image classification. However, this approach is only suitable for

a limited amount of downsampling, e.g. 1/2 or 1/4. Downsampling with a larger factor will result in loss of cell-level features which may result in poor performance [67].

2.1.2 Multi-resolution Images

The use of multi-resolution images for context-aware learning is another obvious approach. Inherently, histology images have multiple resolutions where each resolution is the downsampled version of the highest resolution with a certain downsampling factor (i.e. 1/2, 1/4, 1/8, or 1/16). Many studies [51, 67–69] leverage the multi-resolution nature of histology images and used multi-resolution based classifiers to capture contextual information. Alsubaie *et al.* [68] used multi-resolution input images based convolutional neural networks (CNNs) for the classification of growth patterns in lung cancer. They used 224×224 patches at $20\times$ and $10\times$ where $10\times$ patches contains 4 times larger context as compared to $20\times$ patches but at 1/4 times lower resolution. The use of both resolutions results in better classification performance as compared to single resolution based CNN classifiers. However, these multi-resolution approaches only consider a small part of an image at high resolution and the remaining part at low resolutions to make a prediction. Therefore, these approaches lack the contextual information of cellular architecture of the whole image at high resolution.

2.1.3 Traditional Classifiers with CNN

Some studies [57, 70–72] have used traditional methods to incorporate larger context from patch-based feature representation of histology images. Awan *et al.* [57] presented a context-aware network for breast cancer classification. They used standard support vector machine (SVM) to learn the context from CNN based features of patches extracted from a high-resolution image. This method is only capable of capturing a limited context due to the use of a SVM, which works well on low dimensional feature vectors. Wang *et al.* [70] used an adaptive patch selection approach to aggregate the CNN based patch features to generate a fixed-length feature vector. They used the random forest for the classification of lung carcinoma whole slide images (WSIs). Li *et al.* [72] used a CNN based feature extractor followed by a conditional random field (CRF) for context learning from image patches of size 672×672 in end-to-end trainable manner.

2.1.4 Stacked Networks

Stacked or two-tiered networks have been used for context-aware learning in histology images. Bejnordi *et al.* [73] used a stacked network for breast tissue

classification. They trained their network in two steps. In the first step, they used a small patch size, and in the second step, they fixed the weights of half of the network to feed a larger patch for training the remaining half of the network. They managed to train a network with the largest patch size of $1,024 \times 1,024$ pixels with a small batch size of 10 patches. Recently, some works [74, 75] have used larger high-resolution patches to improve the segmentation of histology images. Agarawalla *et al.* [74] and Kong *et al.* [75] used a 2D Long Short-Term Memory instead of CRF to improve tumor segmentation. Sirinukunwattana *et al.* [76] presented a systematic comparison of different context-aware methods to highlight the importance of context-aware learning.

2.2 Automated Cancer Grading

In literature, many automated methods have been developed for objective grading of the prostate, and colorectal cancer [54, 56, 77–80]. Most of the existing works used traditional machine learning methods [77–80], simple CNN based method [54] or a combination of both [56].

Rathore *et al.* [80] develop a three-step method for the grading of colorectal adenocarcinoma (CRA) images. First, they segment the glandular region in each image using K-means clustering. Second, they use the cellular morphology, spatial architectural patterns of glands, and texture features to train three different SVM classifiers. Finally, an SVM based ensemble classifier is trained using the probabilities of intermediate classifiers to predict the grade of an input image.

Arvaniti *et al.* [54] used CNN based classifier for Gleason grading of prostate cancer in tissue microarrays. They experimented with multiple standard classifiers (i.e. Resnet [43], VGG-16 [42], Inception-V3 [44], DenseNet-121 [81] and MobileNet [82]) to find best classifier. MobileNet turns out as the best classifier which achieves best kappa score, which measure inter-rater reliability, when compared to ground truth grades marked by two different pathologists.

Awan *et al.* [56] presented a method for two-tier CRA grading based on the extent of deviation of the gland from its normal shape (circular/elliptical). They developed a novel best alignment metric (BAM) for this purpose. The BAM measures the aberrance in shape of each gland relative to the typical shape of a normal gland. As a preprocessing step, CNN based gland segmentation was performed, followed by the calculation of BAM for each gland. For every image, average BAM was considered as a feature along with two more features inspired by BAM values. Finally, an SVM classifier was trained using these feature for CRA grading.

The method proposed in this thesis for CRA grading differs from these existing methods in two ways. First, it does not depend on the intermediate step

of gland segmentation, making it independent of segmentation inaccuracies. Second, the proposed method is entirely based on a deep neural network which makes this framework independent of cancer type. Therefore, the proposed framework can be used for other context-based histology image analysis problems.

2.3 Tissue Segmentation

Segmentation of WSIs into histological primitives such as nuclei, cells, glands, and other tissue components is the basic building block for many histology image analysis problems such as morphometric analysis, cancer grading, and cancer classification. The pixel-level segmentation is the most accurate way to delineate the contour of histological primitives. However, the patch-based segmentation method is commonly used in histology image analysis as it requires less computation and memory resources as compared to a deep learning based pixel segmentation method.

2.3.1 Pixel based Segmentation

The pixel-based segmentation is mostly used for nuclei, cells, and glands segmentation which can be then used for morphometric analysis of histology images. Several methods have been developed for pixel-based segmentation under different nuclei [83], and gland [84] segmentation challenges organised by different research groups working on histology images.

Zhou *et al.* [85] presented a contour-aware CNN for the nuclei segmentation. The network consists of two decoder modules, one for nuclei segmentation and other for contour segmentation. A bi-directional hierarchical feature aggregation strategy is then used for final prediction. Their method outperformed all its counterpart methods for the task of nuclei segmentation in MoNuSeg challenge organised by Kumar *et al.* [83]. Graham *et al.* [86] presented a multi-head CNN based method for simultaneous segmentation and classification of nuclei. They used horizontal and vertical distance maps to separate clustered nuclei for accurate instance segmentation. A separate decoder is used to predict the label of each segmented nucleus. Their method has shown superior performance than many existing pixel-based segmentation methods [87–89] including Zhou *et al.* method [85].

Chen *et al.* [90] developed a generic CNN based segmentation network which can be used for the segmentation of both nuclei and glands. Their method simultaneously predicts an object and its contour, which is then integrated to get more precise object segmentation. Their method stood first in the gland segmentation challenge organised by Sirinukunwattana *et al.* [84]. Graham

et al. [53] presented CNN based gland segmentation methods which consist of multiple residual and dilated residual units and an atrous spatial pyramid pooling unit. The network also contains two decoders for gland and contour segmentation. Their method outperformed other gland segmentation methods, including Chen *et al.* approach on the dataset used in the gland segmentation challenge [84].

Although deep learning based pixel segmentation methods delineate object of interests with high precision, these methods are computationally expensive and take a significant amount of time, more than an hour, to process a WSI [91]. Moreover, precise annotation of a large number of objects is required for the robust training of the segmentation methods, which is a tedious and error-prone task. Segmentation methods are mainly required for morphometric based downstream analysis [58]. In contrast, spatial profiling of TME based cancer prognosis can be conducted using patch-based methods since the spatial location of TME constituent is required instead of precise segmentation of each object of interest.

2.3.2 Patch based Segmentation

The most common use cases of patch-based tissue segmentation include cancer detection and classification of cancer subtypes. In patch-based segmentation, histology images are divided into many smaller patches, usually 224×224 pixels at $20\times$, for classification into the required number of classes. The image-level labels are then predicted by aggregation of the patch level predictions.

Most participants of Camelyon16 challenge [50] organized by Bejnordi *et al.* used patch based tissue segmentation approach for the detection of lymph node breast metastasis. Wang *et al.* [92], winner of the challenge, used GoogLeNet [93] for patch based segmentation of WSIs followed by handcrafted feature based random forest classifier. Similarly, different patch based tissue segmentation methods have been used in breast cancer sub-type classification challenge [94] by different teams [64–66].

Recently, patch-based segmentation methods have been used for TME analysis [95–97]. Geessink *et al.* [95] is used VGG16 [42] for patch-based segmentation of tumour and stromal regions to calculate tumour-stroma ratio which is then used for survival analysis of rectal cancer patients. Kather *et al.* [96] segment the colorectal cancer histology WSIs into 9 classes using a patch-based segmentation method (VGG19) [42]. Then, they evaluate the prognostic significance of the abundance of each class (tissue type). Saltz *et al.* [97] used a CNN for patch-based segmentation of lymphocyte in the 13 different types of cancers for the spatial profiling of lymphocytes patterns to explore their prognostic significance in different cancers.

In patch-based segmentation methods, the size of a patch and its resolution becomes important for segmentation precision, which then impacts the TME analysis. Use of large patch at a lower resolution will results in noisy prediction, whereas smaller patch size lacks the contextual information for correct classification methods. The CNN based coarse segmentation method proposed in this thesis addresses both segmentation precision and limited context issues.

2.4 Tumour Microenvironment

The TME consists of many different cell types, with different biological roles and their unique relationships with cancer cells as shown in figure 2.1. The main categories of cells in a TME are tumour cells, immune cells and stromal cells.

Immune cells are the cell which are responsible to fight against the tumour. Major types of immune cells are lymphocytes, neutrophils, and monocytes/macrophages. Each type have specific function. Lymphocytes are white blood cells and can be categories in three types B-cell, T-cell, and NK-cell. B-cells develop in bone marrow and their main function is to produce antibodies against the foreign substances in the body. T-cells complete their development in thymus and their main responsibility is to attack the cells infected with viruses. NK cells are natural kill cells and they kill cells infected by virus by injecting a killer potion of chemicals in them. The half of the white blood cells are Neutrophils which mainly kill bacteria by ingesting it. Monocytes make up 5 to 10 percent of white blood cells. Monocytes change their shape and size when enter a tissue and become macrophages which are essential for killing fungi and bacteria.

Stromal cells are connective tissue cells of an organ and support the function of that organ. Fibroblast are the most common type of stromal cells. The normal fibroblasts are vital in tissue repair in wound healing as they aid in the production of extracellular matrix's components such as collagens, fibres, glycosaminoglycans and glycoproteins [98]. However, tumour associated fibroblast is the key component of the TME [99] and known to promote angiogenesis, supporting the formation of tumours and thus proliferation of cancer cell and metastasis [100, 101].

The main categories of cells in TME can be analysed through routinely used hematoxylin and eosin stained slides. The hematoxylin stains cell nuclei as blue, and eosin stains the extracellular matrix and cytoplasm as pink. Immunostaining is used to analyse the sub categories of lymphocytes. Immunohistochemistry (IHC) is the most common approach which selectively identify antigens in tissue cells. It based on the principle of antibodies binding specifically to antigens in biological tissues. B-cells are usually detected by

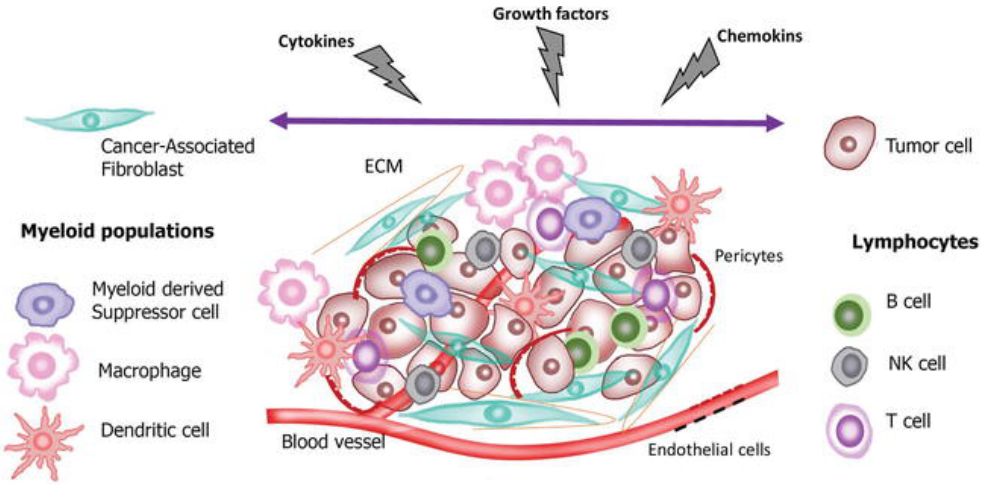


Figure 2.1: Synthetic illustration of tumour microenvironment which shows different types of cell present in a tumour microenvironment. Figure source: [2]

CD19 (cluster of differentiation 19) or CD20 antigen based IHC stains whereas T-cells are indetified by CD8 and CD4 glycoproteins.

2.5 Tumour Microenvironment Profiling

TME profiling refer to the in depth analysis of the TME which then use to patient treatment planning. Currently, most of the work on TME profiling is conducted using molecular profiling and omics data (e.g. genomics, proteomics, metabolomics, and glycomics) which do not contain spatial information. However, the spatial analysis may help to understand the intrinsic architecture of TME in several ways. Therefore, computational methods are required for the spatial profiling of TME in histology images which contains spatial contextual information of whole histology landscape. In literature, TME profiling using histology images has been conducted in a limited number of studies for a different type of cancers. However, to the best of our knowledge, there is no existing work on the automated spatial profiling of TME of HNSCC. A summary of the existing TME profiling method is presented in the following sections.

2.5.1 Quantification of Lymphocytes

Immune cells (mainly, the lymphocytes) are very important for cancer suppression. Interaction between cancer cells and lymphocytes can be investigated based on the quantification of lymphocyte count, ratio, hotspots, and colocalisation with a tumour cells.

Yinyin Yuan [102] investigates the role of lymphocyte infiltration in the

TME of triple-negative breast cancer. A cell classifier is used to detect and classify each cell in a WSI as a cancer cell or lymphocyte using 100 different morphological features [103]. The lymphocytes are then categorised into three classes (intra-tumour, adjacent-tumour, and distal-tumour) based on their spatial location to neighbouring tumour cells. The ratio of each lymphocytic category to the tumour is calculated for survival analysis. Only intra-tumoural lymphocytes based ratio showed prognostic significance for disease-free survival where a higher ratio is associated for more prolonged disease-free survival as compared to a lower ratio.

Nawaz *et al.* [60] presented a statistical methodology to explore the clinical significance of spatial heterogeneity of cancer and immune cells and their relationships in estrogen receptor negative breast cancer. First, cell detection and cell classification were done using a morphological feature (size and circularity of nucleus) based automated method [103]. Then they applied the Getis-Ord [104] method, a hot spot analysis method, on the detected cells to get statistically significant spatial hotspots. Three hotspot maps were generated for tumour, immune and both types of cells. They found that the colocalised hotspots of the tumour and immune cells correlate with better prognosis of estrogen receptor negative breast cancer.

Maley *et al.* [105] explored the role of colocalization of tumour and lymphocytes in TME of breast cancer. They employed a statistical measure of colocalisation [106] for the quantification of spatial co-occurrence of tumour and lymphocytes. They found that the higher colocalisation score is associated with good disease-specific survival. Saltz *et al.* [97] have studied the correlation between the spatial organisation of tumour infiltrating lymphocytes (TILs) and molecular characteristics in histology images of 13 different type of cancers. First, they used CNN based classifier for detection of lymphocytes in tissue regions, which they referred to as TILs. Then, they used an affinity propagation algorithm to identify/quantify the local spatial patterns in the detected lymphocytes for survival analysis. They found that some spatial patterns of lymphocytes are associated with better patient prognosis for a few cancer types.

2.5.2 Quantification of Tumour-associated Stroma

Tumour-associated stroma also plays a vital role in tumour development. Unlike lymphocytes, stroma can change its natural behaviour during malignancy and can start promoting cancer growth instead of suppressing it [107], which leads to reduced survival of cancer patients. Some studies have explored the clinical significance of stroma in TME through automated methods.

Lan *et al.* [108] study TME of ovarian cancer by quantifying the ratio of

lymphocyte, and stromal cells in WSIs. Cells are detected by handcrafted feature-based cell detection and classification method [103]. They found that high stromal cell ratio and low lymphocyte ratio are associated with poor overall survival. Moreover, the high stromal ratio also remains prognostic for disease-free survival of ovarian cancer patients.

Yuan *et al.* [103] explored the prognostic significance of spatial patterns of different types of cells in breast cancer. First, they used a morphological feature based cell detection and classification method for the localisation of the tumour, stroma, and lymphocytes. Then, they employed Ripley's K statistics [109] for the quantification of spatial patterns of each type of cells where high K-score represents a dense cluster of cells, and low K-score denotes scattered distribution of cells of a particular class. They found that high K-score of stromal cells is associated with better survival of estrogen receptor negative breast cancer. However, the same K-score statistics did not show any prognostic significance for estrogen progesterone negative breast cancer.

Failmezger *et al.* [110] presented a graph-based method for the spatial profiling of TME of melanoma. First, they detected tumour, stroma and lymphocytes in WSIs. Then they constructed the graph over the cell locations where nodes are connected based on their spatial distance from each other. Finally, two quantification scores, stromal clusters and stroma barrier, were calculated for prognostic analysis. Both scores were associated with poor overall survival of melanoma patients.

2.5.3 Multi-class Cell Quantification

Most of the aforementioned works only considered one type of non-tumour cell and explored its role to the tumour in TME of different cancer types. However, some works [111, 112] have explored the role of multiple types of cells in TME, simultaneously.

Heindl *et al.* [111] quantify the heterogeneity of different types of cells (tumour, stroma, and lymphocytes) and their relative abundance in the TME of metastatic lesions of ovarian cancer. The cell heterogeneity is calculated by Shannon entropy [113] where higher heterogeneity represents the similar cell abundance of different types and lower heterogeneity denotes the dominance of one type of cells in TME. They found that the higher cell heterogeneity is associated with poor overall and disease-free survival of ovarian cancer patients with metastatic.

Sirinukunwattana *et al.* [112] predicted the distant metastasis by quantifying the cell-cell connections in colorectal cancer WSIs. First, they used CNN based cell detection and classification method to identify four different types of cells. Then, the Delaunay triangulation based graph is constructed based

on the cell locations to find a cell-cell connection. Histogram based features are then used to quantify the occurrences of each type of cell-cell connections. The resultant features are then used for the prediction of distant metastasis using logistic regression.

2.6 Evaluation Measures

I use number of evaluation measures to evaluate the performance of my proposed methods and to compare it with existing methods. Evaluation measures related to classification performance and statistical analysis are described in following sections.

2.6.1 Classification Performance Analysis

I use accuracy, sensitivity, specificity, precision, recall, f1-score, AUC, and Rank-sum for classification related tasks. Accuracy measure the percentage of correctly classified samples in a test dataset. Sensitivity measure the rate of true positives whereas specificity is the rate of true negatives. Precision and recall measures the proportion of true positives in total positive detections and total positive samples in dataset, respectively. The f1-score is defined as

$$f1 = \frac{2 \times (prec \times rec)}{prec + rec}, \tag{2.1}$$

where $prec$ and rec represents precision and recall, respectively. The AUC measure is used to measure the performance across all possible threshold values in a binary classification task. AUC stands for area under the receiver operating characteristic (ROC) curve. It is based on ROC curve which is plotted using precision and recall values at different thresholds.

The Rank-sum measure rank the performances of different methods with respect to the best performing method. The best performing method get first rank and then the methods which lies within 97.5% and 95% of the best performing method get second and third rank, respectively. All other methods get the 4th rank.

2.6.2 Statistical Analysis

The statistical analysis is performed for disease-specific and disease-free survival in order to demonstrate the prognostic significance of the proposed methods in this thesis. The statistical analysis methods used in this thesis are described below.

Kaplan-Meier [70] curve is used to check the probability of a survival event at a certain time interval. It helps in visualizing the separation between two

groups (low risk and high risk) of patients. Cox proportional-hazards model [40] is used to investigate the association between the survival time of patients and one or more predictor variables. Log-rank [114] and Wald [115] test are used to assess the statistical significance (p-value) of the survival distributions of two different patient groups. Along with p-value, hazard ratio (HR) is also used to present the hazard rate (e.g. chances of death or recurrence) to in one patient group with respect to other group.

The concordance index (C-index), developed by Harrell *et al.* [116] is used to compare the predictive ability of patient risk survival models. The index value will be if a patient with the higher risk score have a shorter time-to-event and vice-versa. Spearman correlation coefficient is used to analyse the correlation between to two variables. Its range is from -1 to 1 where negative values represent negative correlation and positive values represent positive correlation. Zero correlation indicates that both variable are independent.

Chapter 3

Context-Aware Convolutional Neural Network

3.1 Introduction

Convolutional neural networks (CNNs) have been widely used to achieve state-of-the-art results for different histology image analysis tasks such as nuclei detection and classification [47–49], metastasis detection [50–52], tumor segmentation [117] and cancer grading [54–56]. Each task requires a different amount of contextual information; for instance, cell classification needs only high-resolution cell appearance along with little neighbouring tissue. In contrast, tumour detection and segmentation rely on a larger context covering multiple cells simultaneously. However, cancer grading requires both high-resolution cell information and a broader view of the spatial organization of cells. Most existing CNN based methods applied to histology images follow a patch-based approach to train different models which tend to ignore contextual information due to memory constraints. Although these models are often trained on a large number of image patches extracted from histology images, often spatial relationships between neighbouring patches are ignored. Due to the lack of necessary contextual information, the inference is independent of underlying tissue architecture, and it is performed based on the limited context captured by individual patches. This approach works well for problems where contextual information is relatively less important for prediction. However, contextual information becomes vital in problems where diagnostic decisions are made based on underlying tissue architecture, such as cancer grading.

In this chapter, I consider colorectal adenocarcinoma (CRA) grading to demonstrate the significance of context-aware CNN in cancer histology image analysis. CRA is the fourth most common cause of cancer-related deaths worldwide [118]. Pathologists determine the grade of CRA by collective analysis of individual cancer cells' abnormality and their spatial organization as a

distorted glandular structure in the histology image. Several studies have adopted a two-tiered grading system to reduce the inter-observer variability [15, 119] by merging the well and moderately differentiated glands into a low-grade tumour and classifying tissue with poorly and undifferentiated glands as a high-grade tumour. I opted for the same two tier CRA grading approach for automated grading.

A CNN based method for CRA grading requires an input image with large contextual information to capture cell organization for accurate grading. I propose a novel framework for context-aware grading of histology images. The proposed framework first learns the local representation by a CNN (LR-CNN) and then aggregates the contextual information through a representation aggregation CNN (RA-CNN), as shown in Figure 3.1. The proposed framework takes a large size image (1792×1792) as an input unlike the usual input image size (224×224) of standard patch classifiers. The input image is then divided into small patches (224×224) in sliding window fashion with no-overlap. The LR-CNN takes the patches as input and converts them into low-dimensional feature vectors where the length of feature vectors depends on the choice of LR-CNN network. These feature vectors are arranged in the form of a feature-cube using the same spatial arrangement in which the corresponding patches were extracted. The feature-cube is then fed into the RA-CNN to make predictions based on both low-resolution feature representation and spatial context. The proposed context-aware framework is flexible enough to incorporate any state-of-the-art image classifier as LR-CNN for local representation learning with the RA-CNN. I present detailed results and show that our proposed framework achieves superior performance over traditional patch-based approaches and existing context-aware methods. Moreover, the proposed framework also outperforms the methods designed specifically for CRA grading using handcrafted features based on gland architecture. Our main contributions in this work are as follows:

- I propose a novel framework for context-aware learning from large high-resolution input images.
- The proposed framework is highly flexible since it can leverage any state-of-the-art network design for local representation learning.
- I explore different context-aware learning and training strategies to examine the framework's ability to learn the contextual information.
- I report the results of comprehensive experiments (with 100+ network models) and comparisons to demonstrate the superiority of the proposed context-aware learning framework over traditional patch-based methods and existing context-aware learning methods.

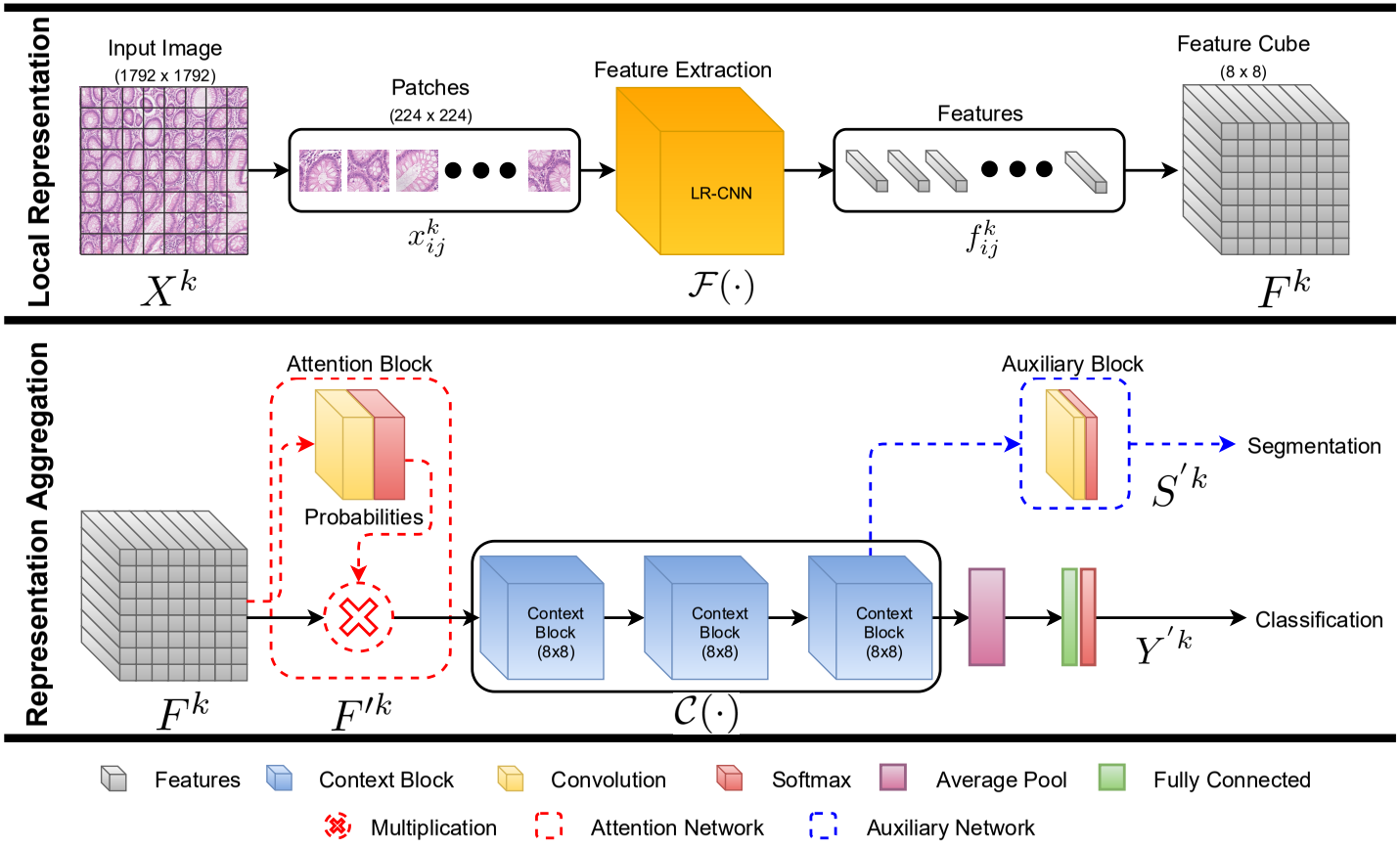


Figure 3.1: Flow diagram of the proposed context-aware framework for CRA grading.

3.2 Method

The proposed framework for context-aware grading consists of two stacked CNNs, as shown in Figure 3.1. The first network, LR-CNN, converts a high-resolution input image into low dimensional feature-cube through patch-based feature extraction. The second network, RA-CNN, aggregates the learned representation in order to learn the spatial context from the feature-cube to make a prediction. I leverage the power of traditional patch classifiers to learn local representation from individual patches. However, I explore different network architectures for context blocks in RA-CNN for context-aware learning. Moreover, different training strategies are explored to find the optimal configurations of the context-aware grading framework. The following section explains each building block of the proposed framework in detail. The notations used to describe each building block are summarized in Table 3.1.

3.2.1 Network Input

The input to our framework is an image (X^k) from a dataset, $D = \{X^k, Y^k, S^k; k = 1, \dots K\}$, of large high resolution images which consists of K images with corresponding labels $Y^k \in \{1, \dots, C\}$ for classification into C classes and patch level segmentation masks $S^k \in \{1, \dots, C\}$ for multi-task learning. Each image is divided into $M \times N$ patches of same size where x^k_{ij} and y^k_{ij} represent the ij^{th} patch of k^{th} image and its corresponding label, respectively. I used a patch dataset, $d = \{(x^k_{ij}, y^k_{ij}), \mid x^k_{ij} \in X^k, y^k_{ij} \in Y^k\}$, which consists of patches and their corresponding labels for pre-training of LR-CNN.

3.2.2 Local Representation Learning

The first part of the proposed framework encodes an input image X^k into a feature-cube F^k. All the input images are processed through the LR-CNN in a patch-based manner. The proposed framework is flexible enough to use any state-of-the-art image classifier as LR-CNN such as ResNet50 [43], MobileNet [82], Inception-v3 [44], or Xception [120]. This flexibility also enables it to use pre-trained weights in case of a limited dataset. Moreover, it is possible to train the LR-CNN independently before plugging it into the proposed framework, enabling it to learn meaningful representation [121] which leads to early convergence of the context-aware learning part of the framework.

3.2.3 Feature Pooling

The spatial dimensions of the output feature f^k_{ij} of a patch x^k_{ij} may vary depending on the input patch dimensions and the network architecture of LR-CNN. A global feature pooling layer is employed to get a one-dimensional

Table 3.1: Enumeration of symbols used in this chapter.

Symbol	Description	Symbol	Description
D	Image dataset	X^k	k^{th} image
K	Number of images	Y^k	Label of k^{th} image
C	Number of Classes	S^k	Mask of k^{th} image
$\mathbf{X}$	Set of all images	$\mathbf{Y}$	Labels of $\mathbf{X}$
$\mathbf{S}$	Masks of $\mathbf{X}$	d	Patch dataset
M	Patches in an image row	i	$1, \ldots, M$
N	Patches in an image column	j	$1, \ldots, N$
x_{ij}^k	ij^{th} patch of X^k	y_{ij}^k	Label of x_{ij}^k patch
$\mathcal{F}(\cdot)$	Feature extractor	f_{ij}^k	Features of x_{ij}^k
L_f	Fully connected layer	L_p^g	Global pooling layer
$L_c^{a\times a}$	$a \times a$ convolution layer	L_s	Softmax layer
$\rightarrow$	Transition between layers	$\bullet$	Preceding layer's output
$\otimes$	Hadamard product	$\oplus$	Feature Concatenation
$\mathcal{B}(\cdot)$	Context-block	$\mathcal{C}(\cdot)$	Context-Net
$\mathbf{F}$	Feature of $\mathbf{X}$	$\mathbf{F}'$	Weighted Feature of $\mathbf{X}$
$\mathbf{Y}'$	Predicted labels of $\mathbf{X}$	Y'^k	Predicted label of X^k
$\mathbf{S}'$	Predicted Masks of $\mathbf{X}$	S'^k	Predicted Mask of X^k
W^k	k^{th} image weight	θ	Learnable Parameters
$\mathcal{L}_{cls}$	Classification cost function	$\mathcal{L}_{wgt}$	Weighted cost function
$\mathcal{L}_{seg}$	Segmentation cost function	$\mathcal{L}_{joint}$	Joint cost function

feature vector for all variations of the proposed framework. Both global average-pooling and global max-pooling strategies are explored. In global average-pooling, values of each feature map are averaged across spatial dimensions whereas only maximum value is considered in global max-pooling from each feature map. After global pooling, features of all patches are rearranged in the same spatial order ($M \times N$) as extracted patches to construct the feature-cube F^k for context-aware learning. The depth of the feature-cube, again, depends on the choice of LR-CNN architecture. Let $\mathbf{F}$ be the output of our LR-CNN which is defined as,

$$\mathbf{F} = \mathcal{F}(\mathbf{X}, \theta_{\mathcal{F}}) \rightarrow L_p^g(\bullet) \tag{3.1}$$

where $\mathcal{F}$ represents the fully convolutional part of the LR-CNN and acts as a feature extractor whereas $\mathbf{X}$ is the batch of images and $\mathbf{F}$ is the local feature representation of $\mathbf{X}$ after pooling L_p^g, which could be a global average-pooling or global max-pooling layer. The operator ($\rightarrow$) provides the output of the preceding layer to the following layer, and the operator ($\bullet$) represents the output of the preceding layer.

3.2.4 Feature Attention

The input of the proposed framework has large spatial dimensions; therefore, there may be some regions of the input image that may be more significant

than others for the prediction of the image label. To exploit these significant regions, I introduce an attention block which gives more weight to the features of the significant regions and less weight to the features of insignificant regions. The architecture of the attention block is illustrated by red dotted lines in Figure 3.1. This attention block takes feature-cube as input and learns the weights for each value in the feature-cube. Hadamard product (element-wise product) is then taken between the weights and input feature-cube to increase the impact of more important regions of an image in label prediction. The weighted feature-cube $\mathbf{F}'$ is defined as:

$$\mathbf{F}' = L_c^{1\times1}(\mathbf{F}, \theta_c) \rightarrow L_s(\bullet) \otimes \mathbf{F}, \tag{3.2}$$

where $L_c^{1\times1}$ and θ_c represent the 1×1 convolution layer and its parameters, respectively. L_s denotes the softmax layer, and the operator $\otimes$ is used to represent Hadamard product.

3.2.5 Context Blocks

Since the LR-CNN is used to encode the patch-based image representation into a feature-cube, the main aim of the context block (CB) is to learn the spatial context within the feature cube. The CB learns the relation between the features of the image patches considering their spatial location. I propose three different CB architectures, each with different complexity and capability to capture the context information. First CB, $\mathcal{B}_1(\cdot)$, is comprised of a 3×3 convolution layer followed by ReLU activation and batch normalization. Second CB, $\mathcal{B}_2(\cdot)$, uses residual block [43] based architecture with two different filter sizes. It consists of three convolution layers each followed by a batch normalization and ReLU activation. The first and last layers are with 1×1 convolution filter to squeeze and expand the feature depth. The output feature-maps of the last layer are concatenated with the input features-maps which makes its final output. The $\mathcal{B}_2(\cdot)$ is defined as:

$$\mathcal{B}_2(\mathbf{F}', \theta_{\mathcal{B}_2}) = [L_c^{1\times1}(\mathbf{F}', \theta_{\mathcal{B}_2^1}) \rightarrow L_c^{3\times3}(\bullet, \theta_{\mathcal{B}_2^2}) \rightarrow L_c^{1\times1}(\bullet, \theta_{\mathcal{B}_2^3})] \oplus \mathbf{F}', \tag{3.3}$$

where $L_c^{1\times1}$ and $L_c^{3\times3}$ denote the convolution layers with 1×1 and 3×3 filter sizes; $\theta_{\mathcal{B}_2^1}$, $\theta_{\mathcal{B}_2^2}$, and $\theta_{\mathcal{B}_2^3}$ are the parameters of different convolution layers and $\theta_{\mathcal{B}_2}$ represents parameter of the whole context block for brevity. The operator $\oplus$ represents the concatenation of feature-maps.

Unlike the previous two context blocks, our third CB processes the input feature-maps in parallel with different filter sizes to capture context from varying receptive fields. Similar to the blocks in [44], it consists of multiple 1×1 and 3×3 convolution layers each followed by a batch normalization and

ReLU activation. A 3×3 average pooling layer $L_p^{3\times3}$ is also used to average the local context information. The CB, $\mathcal{B}_3$, is defined as:

$$\begin{aligned}\mathcal{B}_3(\mathbf{F}', \theta_{\mathcal{B}_3}) = &[L_c^{1\times1}(\mathbf{F}', \theta_{\mathcal{B}_3^1}) \rightarrow L_c^{3\times3}(\bullet, \theta_{\mathcal{B}_3^2}) \rightarrow L_c^{3\times3}(\bullet, \theta_{\mathcal{B}_3^3})] \\ &\oplus [L_c^{1\times1}(\mathbf{F}', \theta_{\mathcal{B}_3^4})] \\ &\oplus [L_c^{1\times1}(\mathbf{F}', \theta_{\mathcal{B}_3^5}) \rightarrow L_c^{3\times3}(\bullet, \theta_{\mathcal{B}_3^6})] \\ &\oplus [L_p^{3\times3}(\mathbf{F}') \rightarrow L_c^{1\times1}(\bullet, \theta_{\mathcal{B}_3^7})],\end{aligned} \tag{3.4}$$

where $\theta_{\mathcal{B}_3^1}$ to $\theta_{\mathcal{B}_3^7}$ are the parameters of different convolution layers and $\theta_{\mathcal{B}_3}$ represents parameter of the whole context block for the sake of notational simplicity.

3.2.6 Representation Aggregation

A cascaded set of three context blocks ($\mathcal{C}(\cdot)$) of the same type ($\mathcal{B}_1$,$\mathcal{B}_2$, or $\mathcal{B}_3$) is used in RA-CNN. The output of $\mathcal{C}(\cdot)$ is followed by a global average pooling layer, a fully connected layer, and a softmax layer to make the final prediction in the required number of classes. The final prediction $\mathbf{Y}'$ from the features of input images $\mathbf{X}$ is computed as:

$$\mathbf{Y}' = \mathcal{C}(\mathbf{F}', \theta_{\mathcal{C}}) \rightarrow L_p^g(\bullet) \rightarrow L_f(\bullet, \theta_{f'}) \rightarrow L_s(\bullet), \tag{3.5}$$

where $\theta_{\mathcal{C}}$ and $\theta_{f'}$ represent the parameters of all context blocks and the fully connected layer in RA-CNN, respectively. The architecture of the RA-CNN is illustrated by black lines in Figure 3.1.

3.2.7 Auxiliary Block

The proposed framework is designed for the classification of large input images. Therefore, the label of an input image may depend on a set of different primitive structures (such as glands, nerves, or vessels) and their spatial organization. I proposed an auxiliary block to exploit these primitive structures. The architecture of the auxiliary block is highlighted by blue dotted lines in Figure 3.1. This auxiliary block acts as a patch based segmentation of the primitive structures in an input image (k) and outputs a patch based segmentation mask (S'^k). The segmentation masks ($\mathbf{S}'$) of input images $\mathbf{X}$ from their features $\mathbf{F}'$ is defined as:

$$\mathbf{S}' = \mathcal{C}(\mathbf{F}', \theta_{\mathcal{C}}) \rightarrow L_c^{1\times1}(\bullet, \theta_{c'}) \rightarrow L_s(\bullet), \tag{3.6}$$

where $L_c^{1\times1}$ is a convolution layer with $\theta_{c'}$ parameters. The addition of auxiliary block enables the proposed framework to learn in a multi-task setting [122–125] where both tasks share the same base network which helps to overcome the

issues of representation bias and overfitting. The loss function for one task acts as a regularizer for the other tasks.

3.2.8 Loss Functions

The proposed framework without auxiliary block is trained only with categorical cross-entropy loss based cost function $\mathcal{L}_{cls}(\cdot)$ which is defined as:

$$\mathcal{L}_{cls}(\mathbf{Y}, \mathbf{Y}') = -\frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C} Y_c^k \log_2(Y_c^{'k}), \tag{3.7}$$

where Y_c^k and $Y_c^{'k}$ are the ground truth and predicted probabilities of k^{th} image for c^{th} class, respectively. The weights of the proposed framework with auxiliary block are optimized based on a joint loss($\mathcal{L}_{joint}$) which consist of $\mathcal{L}_{cls}$ and segmentation-map based loss function ($\mathcal{L}_{seg}$). Both $\mathcal{L}_{seg}$ and $\mathcal{L}_{joint}$ are defined as:

$$\mathcal{L}_{seg}(\mathbf{S}, \mathbf{S}') = -\frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C} S_c^k \log_2(S_c^{'k}), \tag{3.8}$$

$$\begin{aligned}\mathcal{L}_{joint}(\mathbf{Y}, \mathbf{Y}', \mathbf{S}, \mathbf{S}') =& \alpha \times \mathcal{L}_{cls}(\mathbf{Y}, \mathbf{Y}') + \\ & (1-\alpha) \times \mathcal{L}_{seg}(\mathbf{S}, \mathbf{S}'),\end{aligned} \tag{3.9}$$

where α is a hyper-parameter which defines the contribution of both loss functions in the final loss. The loss functions are minimized with RMSprop optimizer [126].

3.2.9 Training Strategies

I trained the proposed framework in four different ways for the sake of completeness in experimentation. First, the proposed framework is trained without attention block and by minimizing the $\mathcal{L}_{cls}(\cdot)$ loss only. Solid black line blocks in Figure 3.1 represent this configuration. Second, the same configuration as first but trained with a sample-based weighted loss function, $\mathcal{L}_{wgt}(\cdot)$, which give more weight to the input patches with relatively less region of interest (glandular region) as compared to the background. The weight of an input patch and $\mathcal{L}_{wgt}(\cdot)$ are defined as follow,

$$W^k = \begin{cases} \frac{1}{R_{roi}^k}, & \text{if } R_{roi}^k > \alpha \\ \frac{1}{\alpha}, & \text{otherwise} \end{cases} \tag{3.10}$$

$$\mathcal{L}_{wgt}(\mathbf{Y}, \mathbf{Y}') = -\frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C} W^k Y_c^k \log_2(Y_c^{'k}), \tag{3.11}$$

Table 3.2: Distribution of visual fields of different classes for both dataset.

Dataset	Normal	Low Grade	High Grade	Total
CRA [56]	71	33	35	139
Extended CRA	120	120	60	300

where R^k_{roi} and W^k represent the ratio of the region of interest and the weight of the k^{th} image. The α is the ratio threshold, selected empirically as 0.10, sets the upper limit of an image weight. Third, multi-task learning based training with the help of an auxiliary block by using joint classification and segmentation loss, $\mathcal{L}_{joint}$. Last, training using the same joint loss but with attention-based feature-cube to amplify the contribution of more important features in the feature-cube. The network configuration of this strategy is represented by both solid and dotted lines blocks in Figure 3.1. I termed these strategies as *standard*, *weighted*, *auxiliary*, and *attention*, respectively.

3.3 Datasets & Performance Measures

I evaluate our proposed framework on two CRA datasets using multiple evaluation metrics. A detailed explanation of both datasets and evaluation metrics is presented in the following subsections.

3.3.1 Datasets

The proposed framework is evaluated on two CRA datasets in order to demonstrate its context-aware grading capabilities. The first CRA dataset was used by Awan *et al.* [56] for the same task of CRA grading. It is comprised of visual fields extracted from 38 haemotoxylin and eosin (H&E) stained whole slide images (WSIs) of CRA cases based on the two-tier grading system [15, 119]. The CRA dataset consists of 139 visual fields with an average size of 4548×7520 pixels obtained at $20\times$ magnification. These visual fields are classified into three different classes (normal, low grade, and high grade) based on the organization of the glands in the visual fields by expert pathologists. I extend this dataset with more visual fields extracted from another set of 68 H&E stained WSIs using the same criteria. Our extended CRA (Extended CRA) dataset consists of 300 visual fields with an average size of 5000×7300 pixels. A detailed distribution of the visual fields of different grades is presented in Table 3.2 for both datasets.

I follow 3-fold cross-validation for a fair comparison of the proposed method with the method presented by Awan *et. al.* [56]. All visual fields extracted from one case only lies in one fold, and I use one fold for training, one for validation (hyper-parameter tuning) and one for the testing to do strong

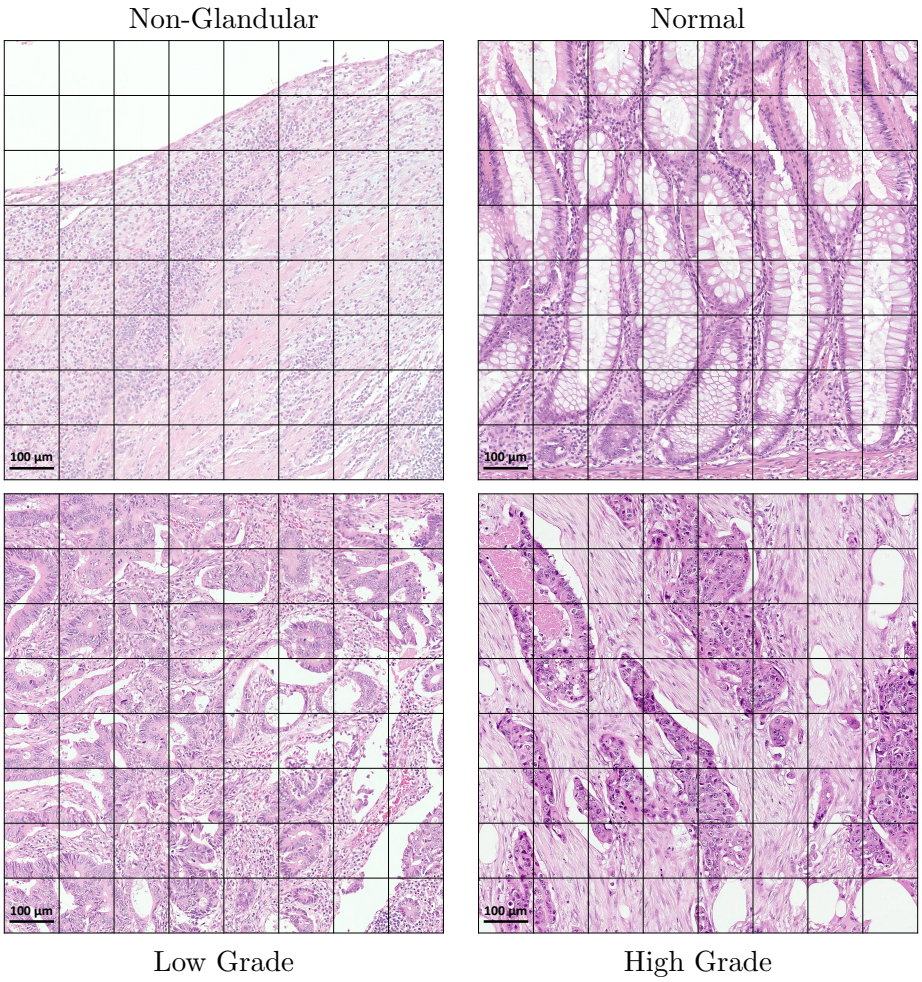


Figure 3.2: Exemplar patches of size 1792×1792 pixels used for the training of the proposed method. Each box of the overlaid grid shows the 224×224 patch used for the training of patch classifiers.

cross-validation on extended CRA dataset. Patches of two different sizes 224×224 and 1792×1792 pixels are extracted for the training of traditional patch classifiers and our proposed framework, respectively (Figure 3.2). A background class is introduced to handle the patches with no or little glandular regions, and background patches are only extracted from normal visual fields. However, patches of all other classes are extracted from glandular regions of the visual fields of respective classes. I extracted 30 000 patches for patch classification and 6000 overlapping patches for context-aware classification for each class in each fold, using random rotation and flipping based augmentation for both datasets.

3.3.2 Performance Measures

I used two metrics, the accuracy and Rank-sum measure, for performance evaluation. The average accuracy refers to the percentage of visual fields

classified correctly. In contrast, weighted accuracy is the sum of the accuracy of each class weighted by the number of samples in that class. The Rank-sum measure rank the performances of different methods with respect to the best performing method. The best performing method get first rank and then the methods which lies within 97.5% and 95% of the best performing method get second and third rank, respectively. All other methods get the 4th rank.Rank-sum based evaluation metric is used to summarize the accuracy of different models in order to compare models trained with different context-blocks and LR-CNNs. Different colours are used to represent different rank for better illustrative visualization, as shown in Tables 3.4 and 3.5. The orange colour indicates the best performing method, whereas the green and blue colours indicate that the results are within 97.5% and 95% of the best performing method, respectively. The rank for these colors are: orange = 1, green = 2, blue = 3, and no colour = 4. The lowest rank-sum represents the best performance.

3.4 Experimental Results

The results of the different variants of the proposed framework are presented to show the superior performance of these variants over simple patch-based methods. These variants include the use of four different state-of-the-art classifiers for local representation learning in LR-CNN; spatial dimensionality reduction through global average-pooling and global max-pooling; the usage of three different context-blocks in RA-CNN; and four different training strategies. By employing different combinations of variations mentioned above, I trained around 100 models in total for each fold on the CRA dataset. The details of experimental evaluation are given in following subsections.

3.4.1 Experimental Setup

The CRA visual fields are divided into patches of size 1792×1792, and the label of each patch is predicted using the proposed framework with a stride of 224×224. The use of small stride can significantly increase the processing time of a visual field due to redundant processing of overlapping regions. I process each visual field with LR-CNN to get representation features of a whole visual field. Afterwards, RA-CNN has applied in a sliding window manner on the features of the visual field to aggregates local representation for context-aware predictions. Through this approach, I process a visual field with a 64 times bigger context as compared to standard patch classifier with only 10% additional processing time. The overall label of a visual field is derived from counting the most predicted class (majority voting), excluding background class in a visual field. Both accuracy and Rank-sum based evaluation measures

Table 3.3: Accuracy based comparison of four patch classifiers.

Network	Fold-1	Fold-2	Fold-3	Mean	Standard Deviation
ResNet50	93.48	93.62	89.13	92.08	2.08
MobileNet	93.48	**95.74**	89.13	**92.78**	2.74
Inception-v3	**95.65**	91.49	86.96	91.37	3.55
Xception	93.48	91.49	**91.30**	92.09	**0.98**

represent the performance at visual fields level.

3.4.2 LR-CNN based Classifiers

Four different LR-CNNs are trained using ResNet50 [43], Inception-v3 [44], MobileNet [82], and Xception [120] with patch size of 224×224 to get the baseline patch-based classification results. The ResNet-50 [43] and Inception network are the winner of Image-Net [127] challenge in 2015 and 2016, respectively. MobileNet is a lightweight network with just 3 million parameters, whereas the Xception network uses separable convolutions which results in a significant reduction in computational complexity. The performance of these classifiers for CRA grading is reported in Table 3.3. Although the performance of all classifiers is comparable, MobileNet shows superior performance with the highest mean accuracy. On the other hand, Xception classifier shows consistent performance across three folds, with the lowest standard deviation.

3.4.3 RA-CNN based Context-Aware Learning

I experimented with three context-blocks, $\mathcal{B}_1$, $\mathcal{B}_2$, and $\mathcal{B}_3$, to train three different variations of RA-CNN, which I termed as RA-CNN 1, RA-CNN 2, and RA-CNN 3. These three RA-CNN classifiers are trained separately with all four LR-CNNs, as explained in section 3.2.6, hence giving 12 different combinations of the context-aware network. The results in table 3.4 shows that context-aware networks achieve superior performance as compare to standard patch-based classifiers. The RA-CNN 3 achieves the best Rank-sum (lowest) which shows its robustness across different representation learning networks. The other two context-aware networks also show comparable performance by remaining in the 97.5% of the best performer.

3.4.4 Local Representation Robustness

I also conducted different experiments to analyze the robustness of local representation learned by different LR-CNNs. These LR-CNNs are used in combination with different RA-CNNs for context learning along with different feature pooling strategies. Each LR-CNN is used to training three RA-CNNs with both global average, and global max pooled feature-cubes. The table

Table 3.4: Rank-sum based comparison of three different context-aware networks with standard patch classifiers. The orange, green, and blue represent the rank 1, 2 and 3, respectively.

LR-CNN (Avg)	Baseline	RA-CNN 1	RA-CNN 2	RA-CNN 3
ResNet50	92.08±2.08	94.25±2.70	92.08±2.08	93.51±3.10
MobileNet	92.78±2.74	93.52±3.55	93.52±1.78	94.25±2.70
Inception-v3	91.37±3.55	94.23±3.71	94.96±2.72	95.68±1.78
Xception	92.09±0.98	94.96±2.72	94.96±2.72	95.68±3.55
Rank-sum	10	7	8	**5**

Table 3.5: Robustness analysis of feature extractors across different methods. The orange, green, and blue represents the rank 1, 2 and 3, respectively.

Methods	ResNet50 (%)	MobileNet(%)	Inception-v3(%)	Xception(%)
RA-CNN 1 (Avg)	94.25±2.70	93.52±3.55	94.23±3.71	94.96±2.72
RA-CNN 1 (Max)	93.52±1.87	93.51±3.10	94.23±2.07	93.54±3.03
RA-CNN 2 (Avg)	92.08±2.08	93.52±1.78	94.96±2.72	94.96±2.72
RA-CNN 2 (Max)	95.68±3.55	93.52±3.55	92.80±2.72	93.54±3.03
RA-CNN 3 (Avg)	93.51±3.10	94.25±2.70	95.68±1.78	95.68±3.55
RA-CNN 3 (Max)	94.23±2.07	92.82±2.01	94.25±2.70	94.96±2.72
Rank-sum	12	12	10	**8**

3.5 compares the results using Rank-sum based measure. It can be observed that the Xception model turns out as the most robust feature extractor in LR-CNNs with the best rank-sum score of 8. The Inception-v3 model shows comparable results to the best performer as its network design has significant overlap with Xception architecture.

3.4.5 Training Strategies

I experimented with four different context related training strategies (*Standard*, *Weighted*, *Auxiliary* and *Attention*) to explore their impact on overall performance. Details of each training strategy are given in Section 3.2.9. Table 3.6 shows the comparison of these training strategies for Xception based LR-CNN. Each entry in the table contains the average accuracy across three RA-CNNs for particular feature pooling (shown in rows) and the training strategies (in columns). Attention based training shows the superior results for max-pooled features, whereas standard training strategy achieves comparable performance for average-pooled features. However, auxiliary loss based training remains robust for both pooling types and achieves the best overall accuracy. More importantly, most of the model shows superior performance than the baseline LR-CNN classifier as shown in Figures 3.3, 3.4, 3.5, and 3.6. These figures present the graphical illustration of 96 experiments using different LR-CNNs. The results obtained with different combinations of feature pooling type, the context blocks in RA-CNN and the training strategies used for the experiments

Table 3.6: Comparison for different training strategies based on average accuracy across three RA-CNNs with Xception based features.

Feature (Pooling)	Standard	Weighted	Auxiliary	Attention
Xception (Max)	94.01	94.49	94.73	**95.21**
Xception (Avg)	**95.20**	94.72	94.72	94.00
Mean	94.61	94.60	**94.72**	94.61

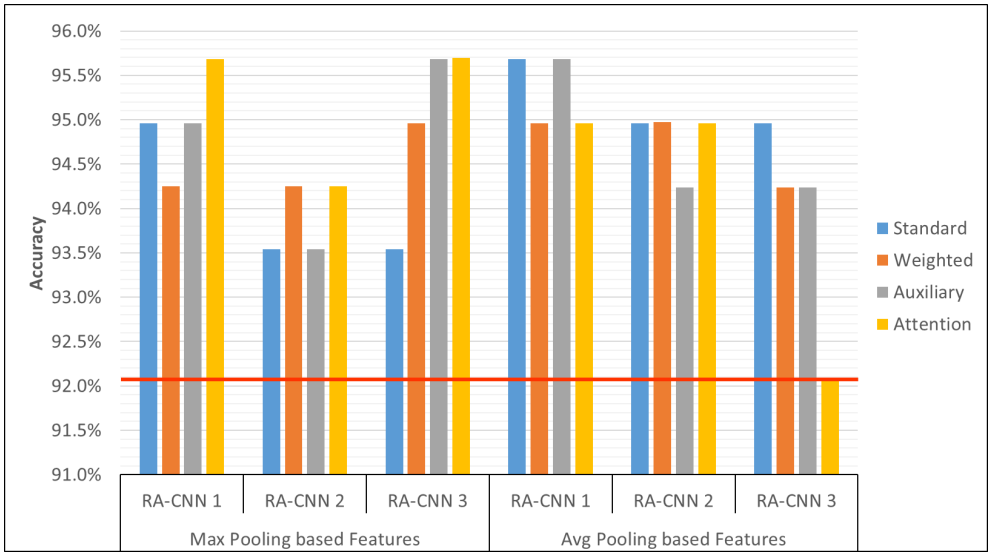


Figure 3.3: Results of 24 experiments using the best performing local representation features (Xception). Legend represents the different training strategies, whereas different bars represent the results for three context-aware networks with max and average pooling based features. Red line indicates the baseline accuracy of patch based Xception classifier.

are illustrated in bar-chart format for better visual comparison. The accuracy obtained by LR-CNN is considered as the baseline and represented by a horizontal red line in each figure for comparative analysis.

3.4.6 Result Summary

The gist of the detailed experimentation and comparisons is that bigger contextual information helps in better automated grading of CRA and the proposed approach demonstrated the ability to capture broader context. In practice, Xception based LR-CNN is the most robust feature extractor for context-aware learning and RA-CNN 3 showed robustness to most of the feature extraction methods. Attention based training strategy is suitable for both RA-CNN 1 and RA-CNN 3 with max-pooling features. Last but not least, almost all proposed variations of context-aware framework perform better than the baseline patch-based classifiers.

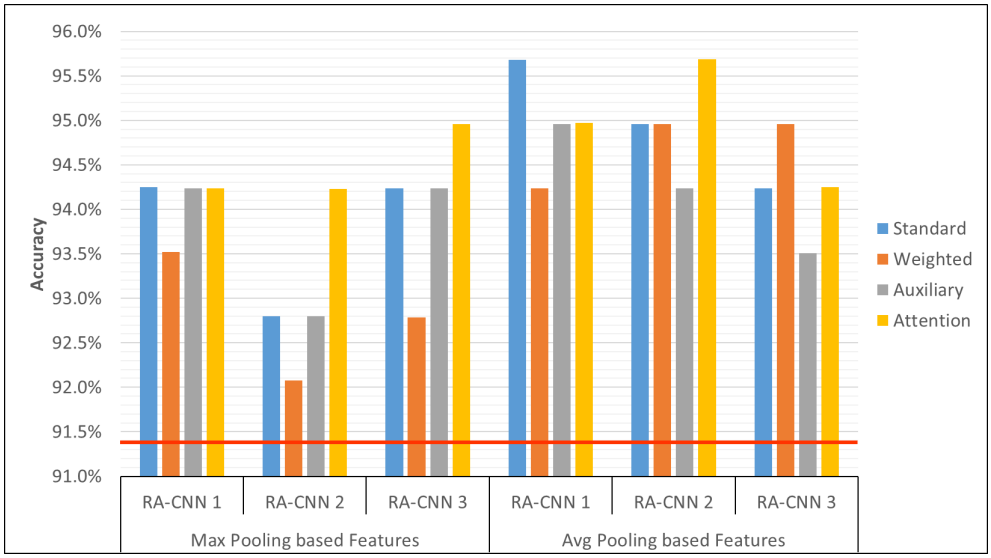


Figure 3.4: Results of 24 experiments using Inception-v3 based local representation features. Results show that all variations of the proposed context-aware method achieved superior performance compared to the Inception-v3 based patch classifier performance (denoted by horizontal red line).

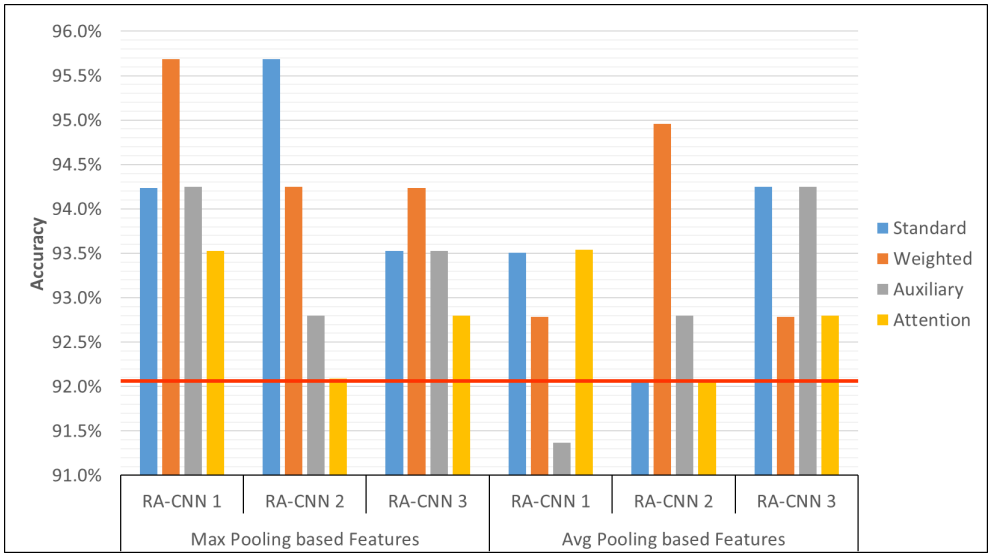


Figure 3.5: Results of 24 experiments using ResNet50 based local representation features. Results show that most of the variations of the proposed context-aware method achieved superior performance compared to the ResNet50 based patch classifier performance (denoted by horizontal red line).

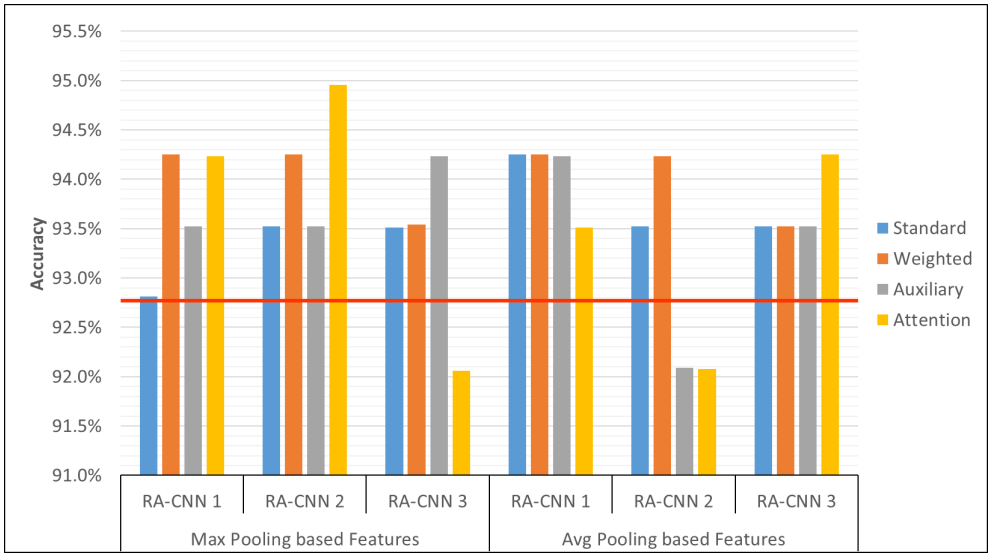


Figure 3.6: Results of 24 experiments using MobileNet based local representation features. Results show that MobileNet based patch classifier achieved reasonably high performance (denoted by horizontal red line) but some of the variations of the proposed context-aware method with MobileNet features does not performed as good as the simple patch classifier.

3.5 Comparative Results

The results of the best performing context-aware method are compared with state-of-the-art approaches on both CRA and Extended CRA datasets. These approaches are categorized into problem-specific methods, traditional patch-based classifiers, and context-aware methods. The brief description of these approaches and comparative analysis is presented in the following subsections.

3.5.1 Problem Specific Methods

Awan *et al.* [56] presented a two-step problem-specific method for CRA grading. They experimented with best alignment metric based two feature sets which I refer to as BAM-1 and BAM-2. BAM-1 comprises of average BAM and BAM entropy while BAM-2 comprises of an additional feature known as regularity index. They evaluated their method using only average accuracy based measures for both binary and 3-class grading. I reported the results presented by the author in their paper [56] on CRA dataset to avoid any retraining bias and compared using average accuracy based measure for a fair comparison. Their method achieved good accuracy for binary grading, normal vs cancer; however, it lacks the robustness required for multi-class grading of CRA visual fields whereas the proposed method achieved superior performance on both tasks (see Table 3.7).

Table 3.7: Average Accuracy based grading comparison of the proposed context-aware method with state-of-the-art methods on CRA Dataset.

ID	Methods	Binary (%)	Three-class (%)
1	BAM - 1 [56]	95.70±2.10	87.79±2.32
2	BAM - 2 [56]	97.12±1.27	90.66±2.45
3	ResNet50 [43]	98.57±1.01	92.08±2.08
4	MobileNet [82]	97.83±1.77	92.78±2.74
5	Inception-v3 [44]	98.57±1.01	91.37±3.55
6	Xception [120]	98.58±2.01	92.09±0.98
7	CNN-SVM [128]	96.44±3.61	92.12±3.57
8	CNN-LR [128]	98.58±2.01	93.52±0.07
9	CNN-LSTM [76]	96.44±3.61	89.96±3.54
10	Proposed	**99.28±1.25**	**95.70±3.04**

Table 3.8: Accuracy based grading comparison of the proposed context-aware method with state-of-the-art methods on the Extended CRA Dataset.

ID	Methods	Patch Size	Binary Classification		3-Class Classification	
			Average (%)	Weighted (%)	Average (%)	Weighted (%)
1	ResNet50 [43]	224x224	95.67±2.05	95.69±1.53	86.33±0.94	80.56±1.04
2	MobileNet [82]	224x224	95.33±2.49	95.42±2.23	84.33±3.30	77.78±4.83
3	Inception-v3 [44]	224x224	93.67±1.89	94.31±1.57	84.67±1.70	81.11±1.97
4	Xception [120]	224x224	96.67±2.05	96.80±1.71	86.33±0.94	81.39±1.71
5	Xception [120]	112x112	92.00±3.27	92.22±2.64	81.33±3.40	74.72±4.53
6	Xception [120]	448x448	97.00±2.83	97.08±2.36	**86.67±0.94**	80.42±1.25
7	CNN-SVM [128]	224x224	96.00±0.82	96.39±0.86	82.00±1.63	76.67±2.97
8	CNN-LR [128]	224x224	96.33±1.70	96.39±1.37	**86.67±1.25**	82.50±0.68
9	CNN-LSTM [76]	1792x1792	95.33±2.87	94.17±3.58	82.33±2.62	83.89±2.08
10	Proposed	1792x1792	**97.67±0.94**	**97.64±0.79**	**86.67±1.70**	**84.17±2.36**

3.5.2 Patch-based Classifiers

The results for four standard patch classifiers on both datasets are presented in Tables 3.7 and 3.8. There is a slight difference in the ranking of these classifiers on both datasets. However, Xception classifier remains consistent in terms of low variance in performance on both datasets. I further experimented with different patch sizes using Xception classifier on the Extended CRA dataset. The results show that the significant change in the patch size without any modification in the network architecture leads to a decrease in the performance as can be seen in Table 3.8 for Xception network. The performance of all the patch based classifiers is below the performance of the proposed method.

3.5.3 Context-Aware Methods

The decision fusion based methods [128, 129] can be loosely considered as context-aware methods if used to predict the visual field labels through the aggregation of patch predictions. I compared our method with the two approaches used by Hou *et al.*[128] on the Extended CRA dataset. They used

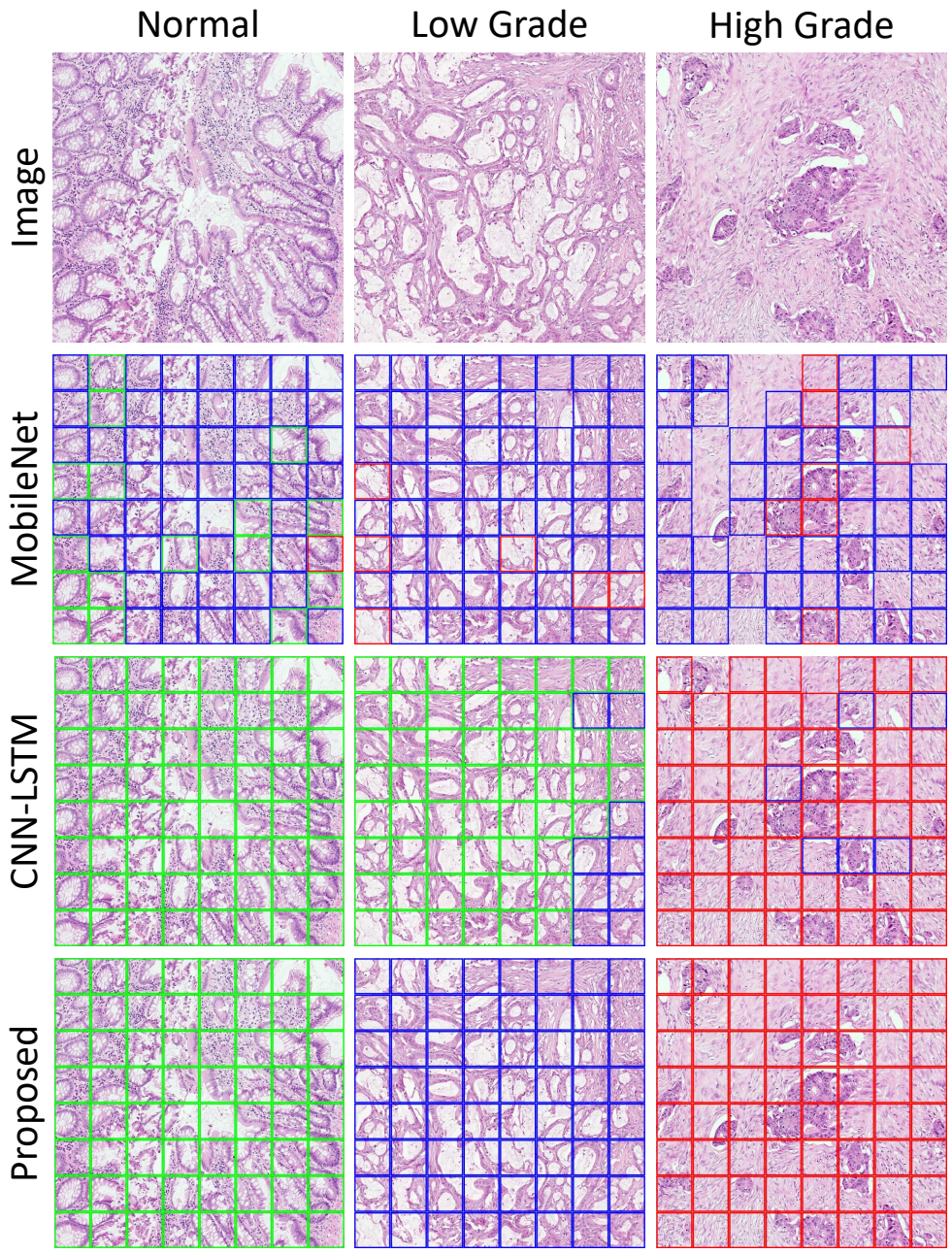


Figure 3.7: Visual results on CRA grading dataset are shown for patch classifier (MobileNet), existing context (CNN-LSTM), and the proposed method on an image of size 1792×1792. The stride size for context networks is equal to the size of patch (224×224) used for patch classifier. Green, blue and red colours of overlaid rectangular boxes show the normal, low and high-grade predictions respectively, whereas empty box areas represent non-glandular/background regions. See text for result analysis.

support vector machine (SVM) with radial basis function kernel (CNN-SVM) and logistic regression (CNN-LR) for decision fusion from the class histogram of patch probabilities. I used the best performing patch classifiers for each dataset (MobileNet for CRA, Xception for Extended CRA) to get the patch probabilities. The CNN-LR shows some performance improvements over the best performing patch classifiers, but this performance is still below the proposed method on both datasets (see Table 3.7 and 3.8). The CNN-SVM method does not perform as good as the simple majority voting based patch classifier. A similar performance pattern can be observed in the Hoe *et al.* paper [128] for the task of Glioma classification. I believe that the major difference between these simple decision fusion and context-aware methods is the ability to adjust the prediction of a patch using its neighbourhood information. The decision fusion based methods only use predicted patch probabilities whereas as context-aware methods have access to the features of neighbouring patches.

I also compared our method with a long short-term memory (LSTM) based context-aware method (CNN-LSTM) proposed in a systemic study on context-aware learning by Sirinukunwattana *et al.* [76] using prostate and breast cancer datasets. They used LSTM to capture the context from CNN features of four downsampled versions (1×, 2×, 4×, and 8×) of the input patch. The code is publicly available by the authors of the paper [76] and I use that code to retrain the method on both datasets for a fair comparison. Our best performing context-aware method outperformed the CNN-LSTM method on both datasets (see Table 3.7 and 3.8). This performance improvement could be attributed to the proposed method's ability to use high-resolution input patch without any downsampling for context learning, unlike CNN-LSTM. Moreover, I used a relatively more powerful CNN network (e.g. Xception) for LR-CNN for feature extraction whereas Sirinukunwattana *et al.* opted for a lightweight network for feature extraction to make their network end-to-end trainable.

3.5.4 The Proposed Method

The different variants of the proposed method have shown comparable performance, but I consider our best performing context-aware configuration for comparative analysis. The best performance is achieved by RA-CNN 1 trained with attention based training strategy on max pooled features. It shows 3.61% and 2.78% better performance as compared to simple patch classifiers on both CRA (Table 3.7) and Extended CRA (Table 3.8) datasets, respectively. I also investigated the performance based on the patch-based segmentation using RA-CNN 1 trained with auxiliary training strategy on the Extended CRA dataset. Although it achieves the weighted accuracy of 87.50%, it has a high variance of 5.14% across three folds of the Extended CRA dataset. Therefore,

I did not consider it as our benchmark for comparative analysis in Table 3.8.

3.5.5 Visual Comparison

The visual comparison of best performing patch classifier (MobileNet), Sirinukunwattana *et al.* (CNN-LSTM) and the proposed method on three different images with normal, low and high grades are shown in Figure 3.7. Patch classifier's prediction is quite irregular for any given image due to the lack of contextual information. The predictions of CNN-LSTM are relatively smooth, but it predicts the wrong label for the low-grade image, which might be due to the use of low-resolution images for context learning. However, the proposed method predictions are smooth and consistent with the ground truth labels.

3.6 Summary

In this chapter, I present a novel context-aware deep neural network for CRA grading, which is able to incorporate 64 times larger context than standard CNN based patch classifiers. The proposed network is well-suited for the CRA grading task, which relies on recognizing abnormalities in glandular structures. These clinically significant structures vary in size and shape that cannot be captured efficiently with standard patch classifiers due to computational and memory constraints. The proposed context-aware network is comprised of two stacked CNNs. The first LR-CNN is used for learning the local representation of the histology image. The learned local representation is then aggregated, considering its spatial pattern by RA-CNN. The proposed context-aware model is evaluated on two datasets for CRA grading. A comprehensive analysis of different variations of the proposed model is presented and compared with existing approaches in the same evaluation setting. The qualitative and quantitative results demonstrate that our method outperformed the patch-based classification methodologies, the problem-specific techniques, and existing context-based methods. This approached is suitable for cancer analysis which requires large contextual information in the histology images. This includes Gleason grading in prostate cancer and tumour growth pattern classification in lung cancer.

Chapter 4

Spatial Quantification of Tumour Infiltrating Lymphocytes Abundance

4.1 Introduction

Tumour infiltrating lymphocytes (TILs) have been analysed in a wide range of cancers with strong evidence demonstrating their prognostic value as a supplement to the tumour, node, and metastasis (TNM) staging [27, 28, 31]. TILs mainly comprise T lymphocytes which migrate from the blood into the tumour as part of the body's immune 'fight-back- response'. However, it is crucial to analyse these cells in the correct context. A large number of lymphocytes can be present in inflamed and cancerous tissues and, therefore, it is vital to develop methods to specifically analyse lymphocytes infiltrating the tumour as these are the ones that are likely to be of prognostic significance [29, 30, 130]. These areas can be referred to as the TIL regions where both tumour cells and lymphocytes are co-localised (as illustrated in Figure 4.1). Numerous studies have reported the correlation of TIL density with improved disease-specific survival and longer disease-free survival [131, 132]. It has been shown that the quantification of spatial patterns of TILs in the tumour regions can have a prognostic value significantly supplementing or even superseding the TNM staging in certain settings [97, 133]. However, the currently used method of visual TIL quantification is subjective with inter- and intra-observer variability and lack of diagnostic reproducibility [134]. Therefore, it is imperative to develop an automated method for objective spatial quantification of TIL to address these challenges.

In this chapter, I present a novel framework for spatial quantification of TILs and explore its prognostic significance for disease-free survival of oral squamous cell carcinoma (OSCC) patients. The proposed framework, as illus-

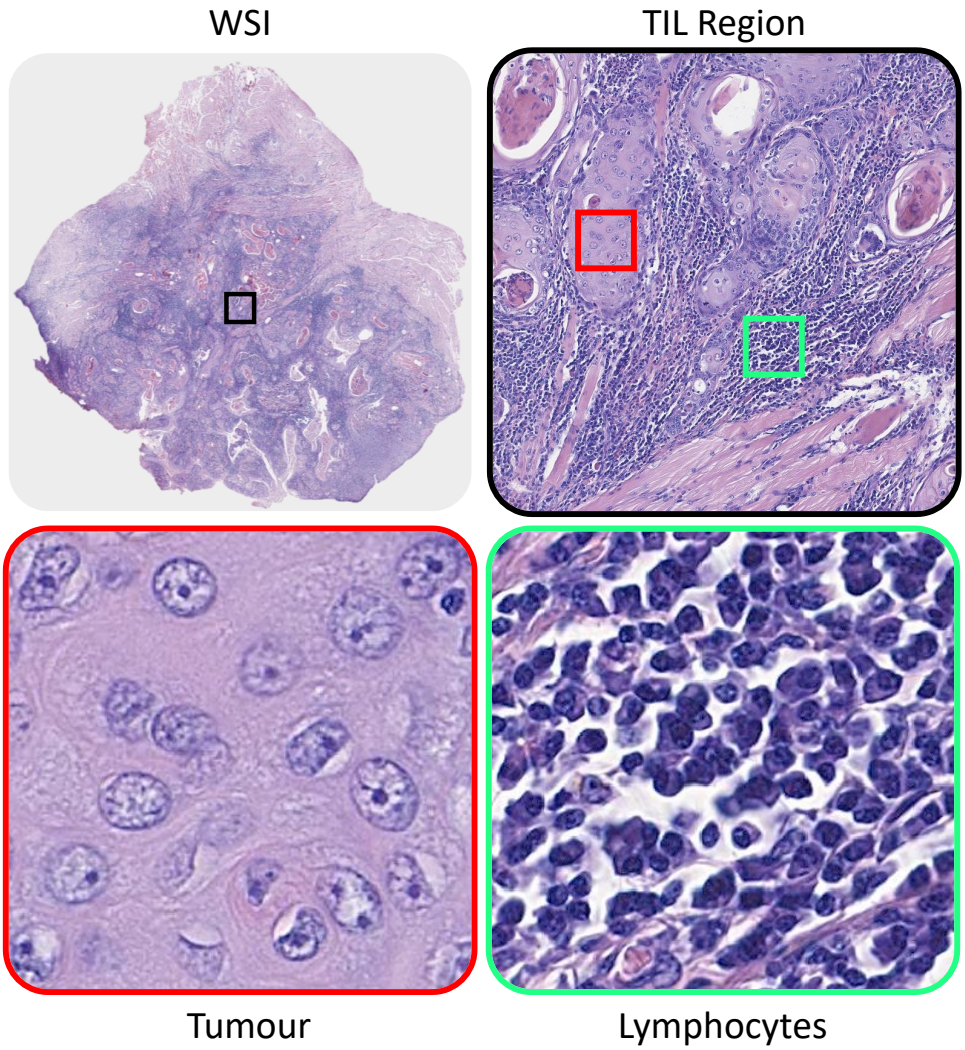


Figure 4.1: An example image of tumour infiltrating lymphocytes region (black) in a whole slide image. High resolution view of tumour (red) and lymphocyte (green) regions are shown in the bottom row.

trated in Figure 4.2, comprises of three main components: WSI segmentation into biologically significant tissue phenotypes, identification and quantification of TILs, and their prognostic analysis. First, WSI segmentation into biologically significant tissue phenotypes is modelled as a patch-based tissue classification problem. Different tissue regions such as tumour and lymphocytes in a WSI are classified using a CNN. Second, a tumour-lymphocyte colocalisation based binary classifier is developed using statistical colocalisation measures for detecting the presence or absence of TILs in OSCC tissue slides. The extent of spatial lymphocytic infiltration, which I term as the tumour infiltrating lymphocytes abundance (TILAb) score, in the tumour region is quantified by a combination of lymphocyte to tumour ratio and their statistical colocalisation. Finally, the prognostic significance of the TILAb score for disease-free survival is investigated by employing univariate and multivariate analysis. To the best of our knowledge, there is no existing method for automated spatial quantification of TIL abundance from digitised WSIs for OSCC patients survival analysis. I show that the TILAb score is a strong prognostic indicator of disease-free survival in OSCC patients in agreement with previous findings based on manual TIL quantification [135]. Our main contributions in this work are as follows:

- A methodology for the segmentation of biologically significant regions in OSCC tissue is presented, which includes segmentation of tumour areas and lymphocytes in a WSI.
- I propose a novel scoring of TIL abundance, termed as the TILAb score, to quantify the extent of spatial lymphocytic infiltration in the tumour region which is a combination of lymphocyte to tumour ratio and their statistical colocalisation in a WSI.
- The reproducibility and objectivity of the TILAb score are investigated in two different ways: First, by analysing the consistency between statistical colocalisation based TIL detection and a pathologist's detection. Second, by evaluating the prognostic significance of TILAb score for disease-free survival of OSCC patients.

4.2 Methods

WSIs are multi-gigapixel images and cannot be used directly for image analysis tasks, particularly training a deep learning based classifier. Therefore, I divide the WSIs into patches for processing. A deep learning based classifier (see next section) is applied on the patches to identify whether the patch contains tumour, lymphocytes or other histological primitives. However, the regions where the lymphocytes are infiltrating the tumour may not be confined within a patch.

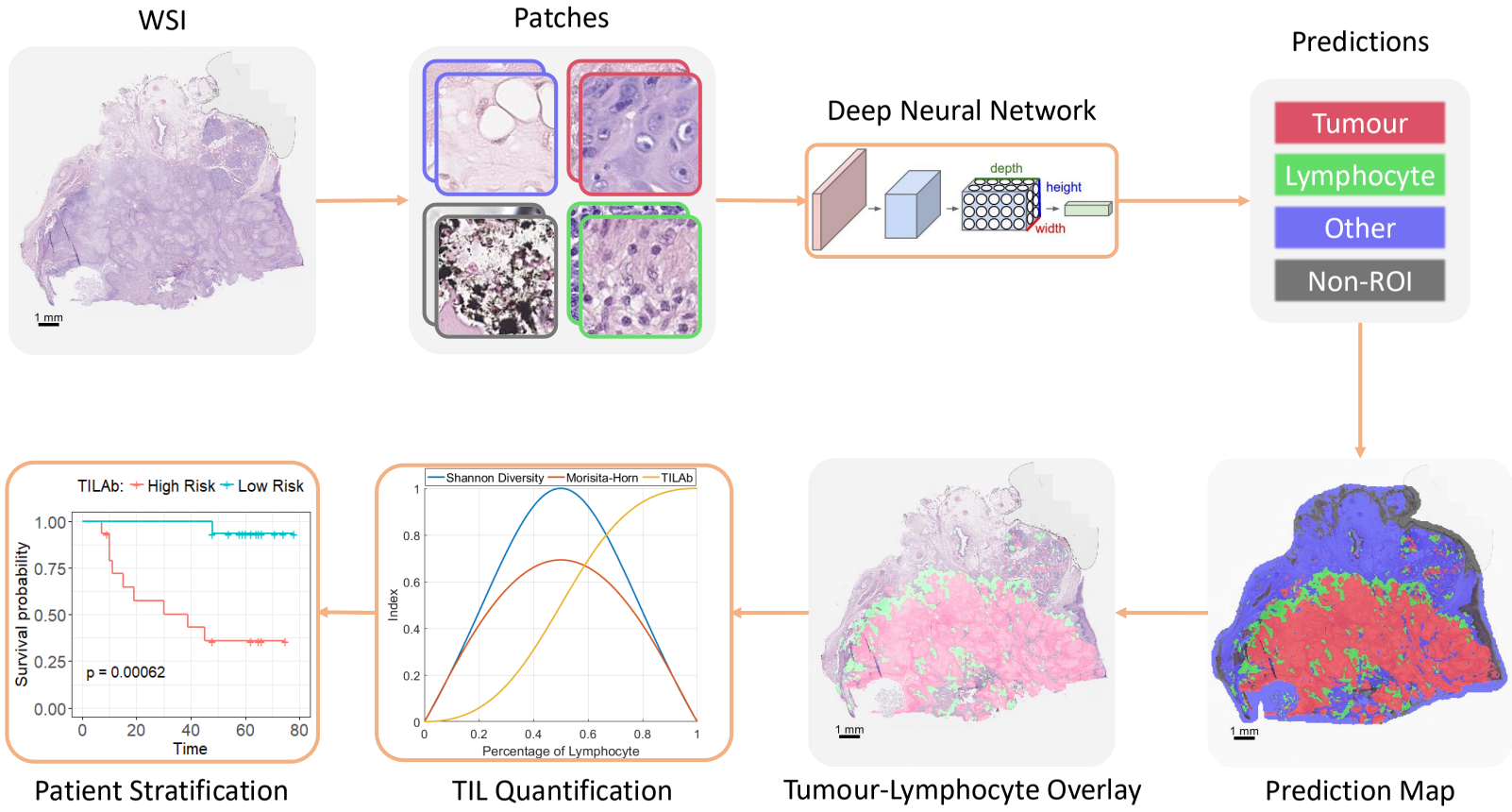


Figure 4.2: Flow diagram of the proposed framework.

Besides, there is considerable variation in the size of TIL regions, making the quantification of TILs a non-trivial task. I identify TILs by calculating spatial colocalisation of tumour and lymphocyte patches. The colocalisation measure is further incorporated into the computation of the proposed score of spatial lymphocytic infiltration, i.e. the TILAb score.

4.2.1 Tissue Region Classifier

A tissue section in a WSI contains many different types of cells and regions, such as tumour cells, lymphocytes, and other regions (i.e. fibroblasts, endothelial cells, blood vessels, muscle, fat and red-blood cells). A WSI may also contain slide preparation and scanning artefacts, such as tissue folding and blurring, which need to be ignored. Therefore, I classify OSCC tissue sections into biologically significant regions. Tumour and lymphocyte rich regions are important for the detection and quantification of TILs. Precise classification of all other regions is necessary to discriminate between TILs and regular lymphocytes that do not lie within the vicinity of tumour regions. The fourth and final class of regions consists of scanning and tissue artefacts, which are labelled as non-region of interest (Non-ROIs). I opted for patch-based tissue region classification instead of pixel-based classification. Figure 4.3 shows three exemplar patches of size 128×128 pixels at $20\times$ of each class.

Deep learning models have significantly improved the state-of-the-art in many natural image-based problems such as visual object detection and recognition [127, 136] and scene labelling [46]. Most popular deep learning networks for the classification task are ResNet [43], DenseNet [81], Inception [44], Xception [120] and MobileNet [82]. Each network shows competitive results on one of the largest image classification datasets, ImageNet [127]. I train these networks for tissue classification task to get a strong baseline model. I extracted $400,000$ patches for training and $100,000$ patches for validation from WSIs. Both training and validation datasets have equal numbers of patches for each class. During training, I leverage online data augmentation with a random rotation of 0, 90, 180 or 270 and random flipping. I select the best model of each classifier after training it for at least $125,000$ optimisation steps with RMSProp optimiser. During testing, the patch classifier takes non-overlapping patches from a WSI and outputs probabilities of all classes for each patch, resulting in a probability map at the WSI level. The probability maps are converted into prediction maps by selecting the class with the highest probability for each patch (Figure 4.2). The prediction map at the WSI level is eventually used for TIL identification and computation of the TILAb score.

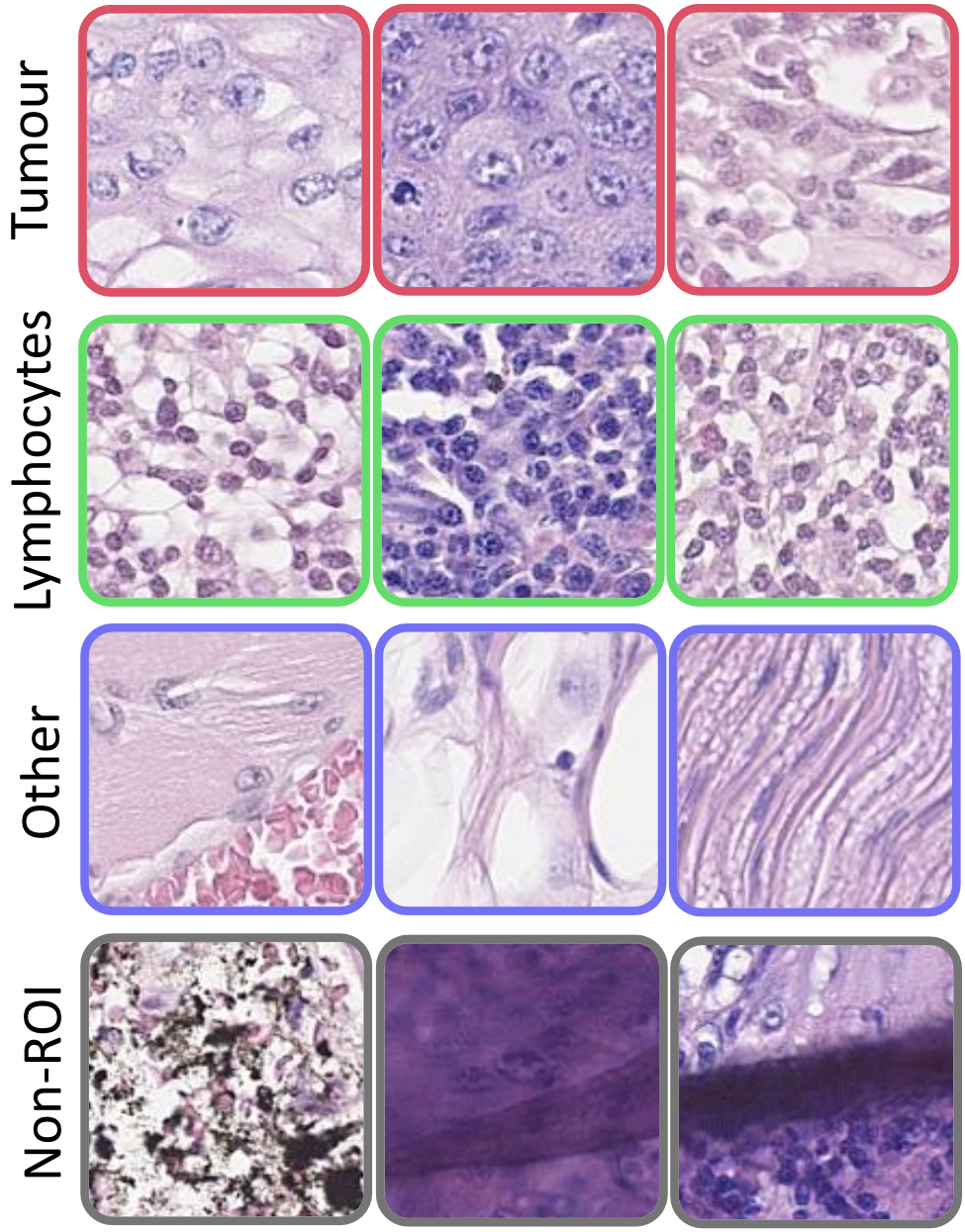


Figure 4.3: Exemplar patches of tumour, lymphocyte, other, and Non-ROI (artefacts) classes.

4.2.2 TIL Detection and Quantification

The TIL regions can easily be detected just by localising the lymphocyte in the vicinity of tumour regions. However, objective quantification of TILs is a non-trivial task as it depends on the meticulous aggregation of the colocalisation of a tumour and lymphocytic regions. In ecology, colocalisation of different species is used to understand their community structure [137, 138]. Tumour and lymphocytes in a WSI could be considered as two different interacting species in the histological landscape. Therefore, I investigate the utility of colocalisation based methods, used in ecology domain, for objective TIL quantification in a WSI. For this purpose, a WSI is divided into $m \times n$ grid of equally sized cells, such that grid-cell size is greater than the size of input patch for the region classifier. The colocalisation score M in terms of the Morisita-Horn [106] index is then defined as follows,

$$M = \frac{2\sum_{i=1}^{m}\sum_{j=1}^{n}(p_{ij}^{l} \times p_{ij}^{t})}{\sum_{i=1}^{m}\sum_{j=1}^{n}(p_{ij}^{l})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n}(p_{ij}^{t})^2}, \tag{4.1}$$

where p_{ij}^{l} and p_{ij}^{t} represent the percentage of lymphocytic and tumour patches (regions) in the $(i,j)^{th}$ grid-cell, respectively (Figure 4.4). Each grid-cell represents the spatial colocalization of tumour and lymphocytes, whereas M is the overall colocalization score. If a grid-cell does not contain any tumour and lymphocytic region, then it would not contribute towards the overall colocalization score. However, if a grid-cell only contains one type of region, either tumour or lymphocyte, then it only contributes to the denominator of the equation thus results in a relatively small colocalization score. If all the grid-cells contain only a unique type of regions, then the colocalization score becomes zero. The overall colocalization score ranges from 0 to 1, and the score is maximum when each of the grid-cell has the same number of tumour and lymphocyte regions, as shown in Figure 4.4.

I also consider the Shannon diversity index [113] to quantify the colocalization of tumour and lymphocytic regions in a WSI. It computes the diversity of two classes in a given region that is also aligned with the definition of TILs. For $m \times n$ grid of equal cell sizes in a WSI, the Shannon diversity index, S, is defined as follow,

$$S = \frac{-\sum_{i=1}^{m}\sum_{j=1}^{n}(p_{ij}^{l} \times \ln p_{ij}^{l} + p_{ij}^{t} \times \ln p_{ij}^{t})}{m \times n}, \tag{4.2}$$

where p_{ij}^{l} and p_{ij}^{t} represent the percentage of lymphocytic and tumour regions in the $(i,j)^{th}$ grid-cell. The colocalization computed using Shannon diversity index is relatively smaller in magnitude as compared to the Morisita-Horn index with the maximum value of 0.7 at the maximum colocalization point of

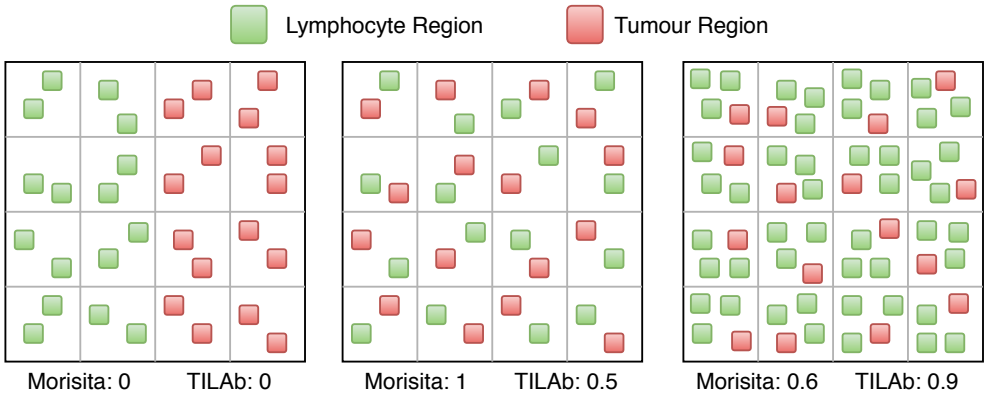


Figure 4.4: The illustration of tumour and lymphocyte colocalization patterns in synthetic images with 4×4 grid size. (Left) The highly segregated appearance of tumour and lymphocytic regions. (Center) Fully co-localized regions. (Right) Lymphocyte rich colocalization.

tumour and lymphocytic regions (Figure 4.5).

4.2.3 TIL Abundance Score

The Morisita-Horn and Shannon diversity are two objective and efficient measures for quantification of colocalisation. However, these methods give equal importance to all the constituent classes (or species), which consequently results in a symmetric colocalisation score (as can be seen in Figure 4.5). For instance, 20% lymphocytes and 80% tumour in a region will give the same colocalisation score as another region consisting of 80% lymphocytes and 20% tumour. However, the lymphocyte proliferation in the tumour is considered to be a good prognostic indicator for patient survival. Therefore, the symmetric nature of these measures is not ideal for obtaining an objective TIL quantification score of prognostic importance. I proposed the TILAb score, T, which is a combination of the lymphocyte to tumour ratio and their colocalisation, as defined below,

$$T = \begin{cases} \frac{C}{2} \times \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(p_{ij}^{l})}{\sum_{i=1}^{m}\sum_{j=1}^{n}(p_{ij}^{t})}, & \text{if } \sum_{i=1}^{m}\sum_{j=1}^{n}(p_{ij}^{t}) > 0 \\ 1, & \text{otherwise} \end{cases} \tag{4.3}$$

where C is a colocalisation measure and M or S can be used as a colocalisation measure. The right half of the above equation shows the lymphocytes to tumour ratio in WSI. I normalise the range of TILAb score between 0 to 1 by dividing it by a factor of 2. The proposed TILAb score objectively quantifies the TILs, and its formulation is generic enough to work with both Morisita-Horn and Shannon diversity indices. Figure 4.5 shows the distribution of TILAb score using Morisita-Horn index based colocalisation at different percentages

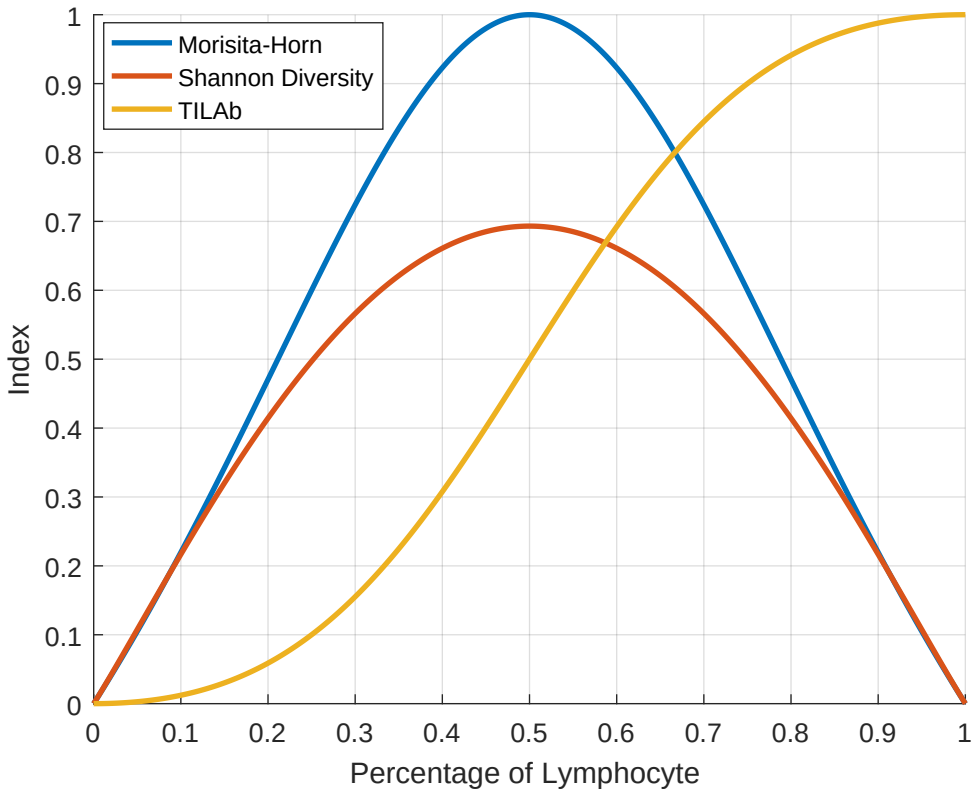


Figure 4.5: Plots of Morisita-Horn, Shannon diversity and TILAb indices for different percentages of lymphocyte in a grid-cell. At each point, the percentage of tumour is equal to 1 − percentage of lymphocyte. TILAb is calculated using Morisita-Horn index based colocalisation.

of lymphocytes. It can be seen that both Morisita-Horn and Shannon diversity indices have relatively small values even with high lymphocytic percentage, whereas the TILAb score increases with the increase in lymphocytic infiltration (Figure 4.6). Moreover, TILAb score remains the same for different tumour and lymphocyte density with the same ratio (Figure 4.7).

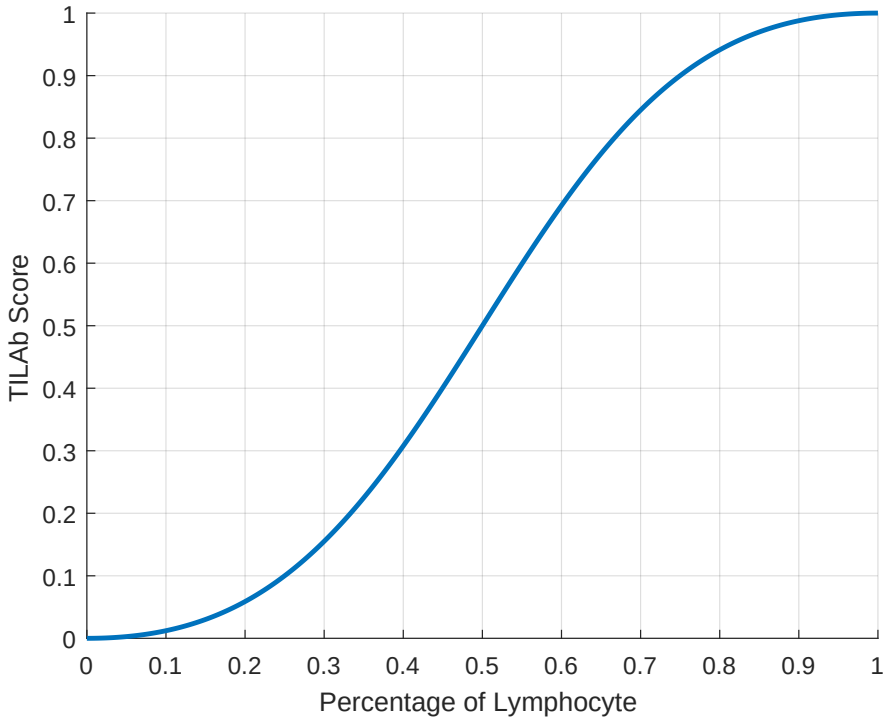
4.2.4 Statistical Analysis

The TILAb score based statistical analysis is performed for disease-free survival in order to demonstrate its prognostic significance as an independent biomarker. Kaplan-Meier [70] and Cox proportional-hazards model [40] are used for survival and hazard analyses, respectively. To stratify patients into high-risk (short-term survival) and low-risk (long-term survival) groups, I find an optimal cut-point on the TILAb score value from the modelling subset where the statistical significance of the difference in disease-free survival between the two groups is the largest. Log-rank test based p-value is used to assess the statistical significance of the survival models where $p < 0.05$ is considered significant. For multivariate analysis, Cox proportional-hazards model is used, which simultaneously evaluates the effect of several factors on survival. I report the hazard ratio along with lower and upper 95% confidence interval. Global statistical significance of the model is measured by the Wald test [115] which is a way to find out if explanatory variables in a model are significant or not.

4.3 Dataset

4.3.1 Ethical approval

Ethical approval was obtained from the institutional review board (Ref. No. 17-02-17-10) at Shaukat Khanum Memorial Cancer Hospital and Research Centre (SKMCH&RC) and national bioethics committee (No.4-87/17/NBC-234-Exempt/NBC/2592), Pakistan. All methods and experiments were carried out by following relevant guidelines and regulations. The institutional review board granted exemption from written consent at SKMCH&RC and national bioethics committee because the data and images used in the study were already in existence and were collected and reported in an anonymised way ensuring confidentiality of participants. The study did not involve any intervention or interaction with the participants. The research involved no more than minimal risk to the participants and involved no procedures for which written consent is normally required outside of the research context, and the waiver did not adversely affect the rights and welfare of the participants.



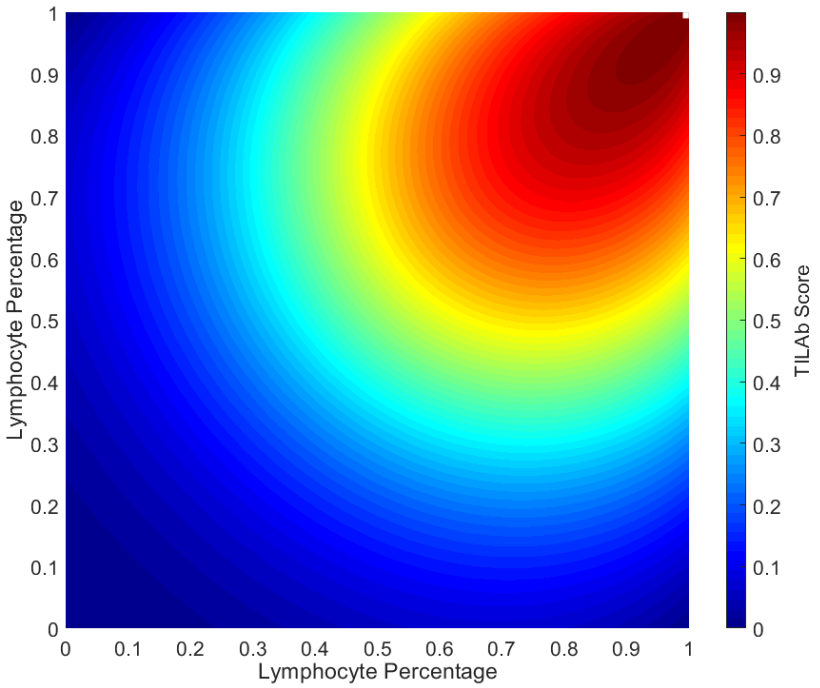


Figure 4.6: Both figures show the distribution of TILAb score with respect to lymphocyte percentage in a grid. (**Top**) TILAb score curve based on the simplest grid with only one cell. (**Bottom**) TILAb score map for a grid with two cells. Lymphocyte percentage in each cell is independent of other cells.

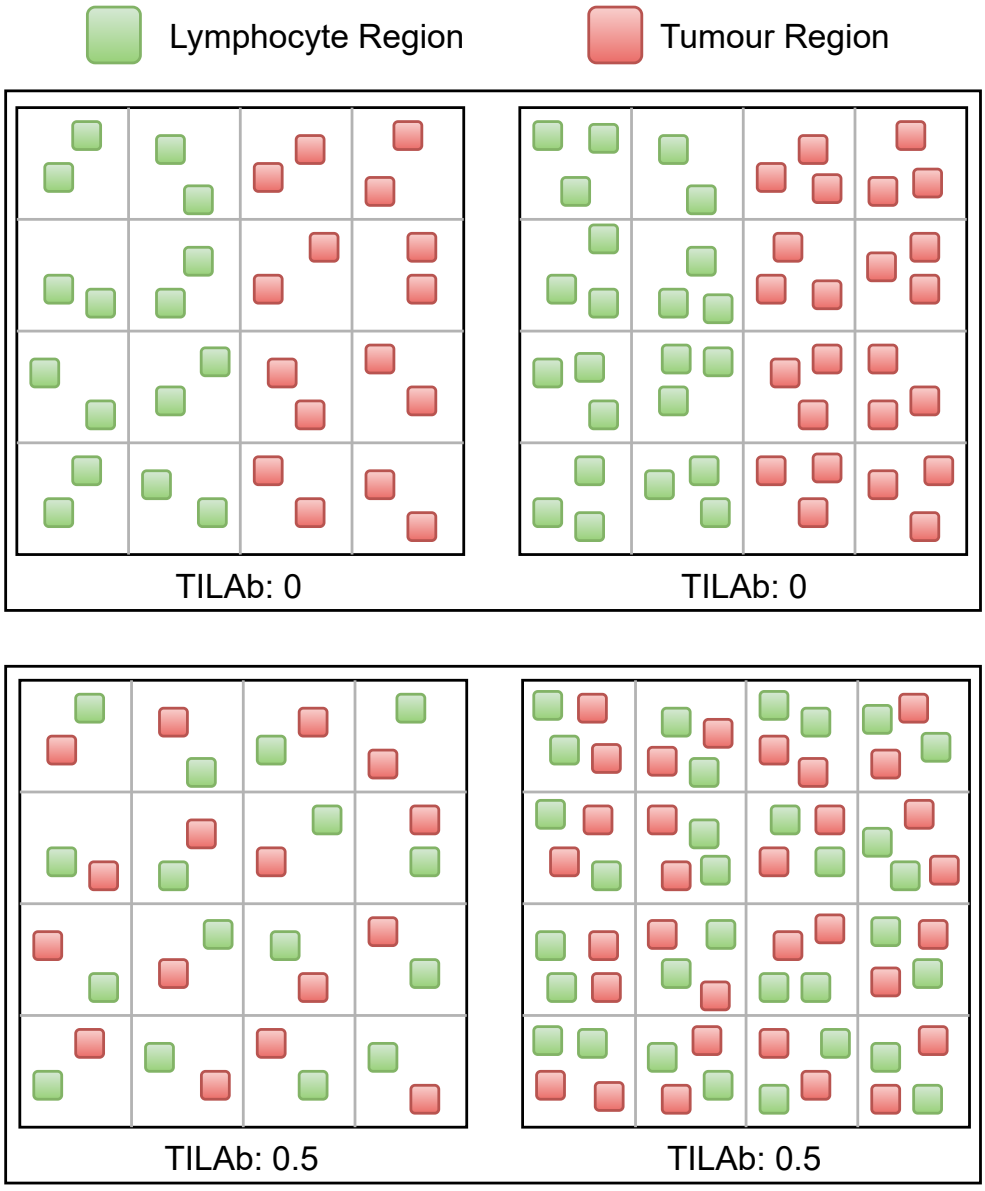


Figure 4.7: The illustration of TILAb score's invariance to tumour and lymphocyte density patterns. Each pair of images has a varying tumour and lymphocyte density but has same TILAb score.

4.3.2 Patient selection

Patients with OSCC diagnosed between 2010-11 at SKMCH&RC were selected from the electronic medical records (hospital information system). SKMCH&RC follows a multidisciplinary approach; therefore, all patients were treated by both a radiation oncologist and by a head and neck surgeon. Cases included both primary and recurrent tumours that underwent complete tumour resection with or without lymph node dissection and for which at least three-year survival data were available. After the initial review of data, 60 patients were selected out of 155 oral cancer cases as per the study protocol (Figure 4.8). The cases excluded were those with either an incomplete resection and those where survival follow up was less than three years. A final cohort of 60 malignant cases and ten controls were finalised where the control cases did not suffer from OSCC. Formalin-fixed paraffin-embedded blocks were retrieved, and representative slides from each case were reviewed by the study pathologists and confirmed to be OSCC. Additionally, slides were reviewed for additional histopathological features such as patterns of invasion, TILs and perineural invasion. For this study, oral cavity cancers were defined as carcinomas of the mouth including lip, tongue, cheeks, the floor of the mouth, hard and soft palate, whereas tumours of the salivary glands were excluded. After compilation of the clinical and pathologic information, including American Joint Committee on Cancer 7th edition stage, clinical and pathological information was retrieved from the electronic medical records, as summarised in Table 4.1. De-identified, tissue slides were digitally scanned in University Hospitals Coventry & Warwickshire using Omnyx Integrated Digital Pathology system at 40× magnification with a resolution of $0.275\mu m$ per pixel. There were 193 tissue sections in 70 digitally scanned WSIs as many WSIs contain multiple tissue sections.

4.3.3 Patient characteristics

Our study cohort consists of 70 cases, including 60 OSCC and ten control cases. Disease-free survival information is available for all the malignant cases where survival time was calculated from the date of diagnosis. Disease-free survival had a census taken at the date of first recurrence or death, whichever occurred first, or the date of the last contact for the patients alive and without recurrent disease. The follow-up period ranged from 3.8 years to a maximum of 6.10 years at the time of data retrieval (2017). Median disease-free survival was 58 months (range, 4 – 86 months) and median age 50 years (range, 25 – 75 years). Approximately 32% ($n = 19$) of patients suffered from disease recurrence whereas 22% ($n = 13$) had died by the time of data retrieval. About 60% ($n = 36$) of the patients were male, while 42% ($n = 25$) are at stage I/II and

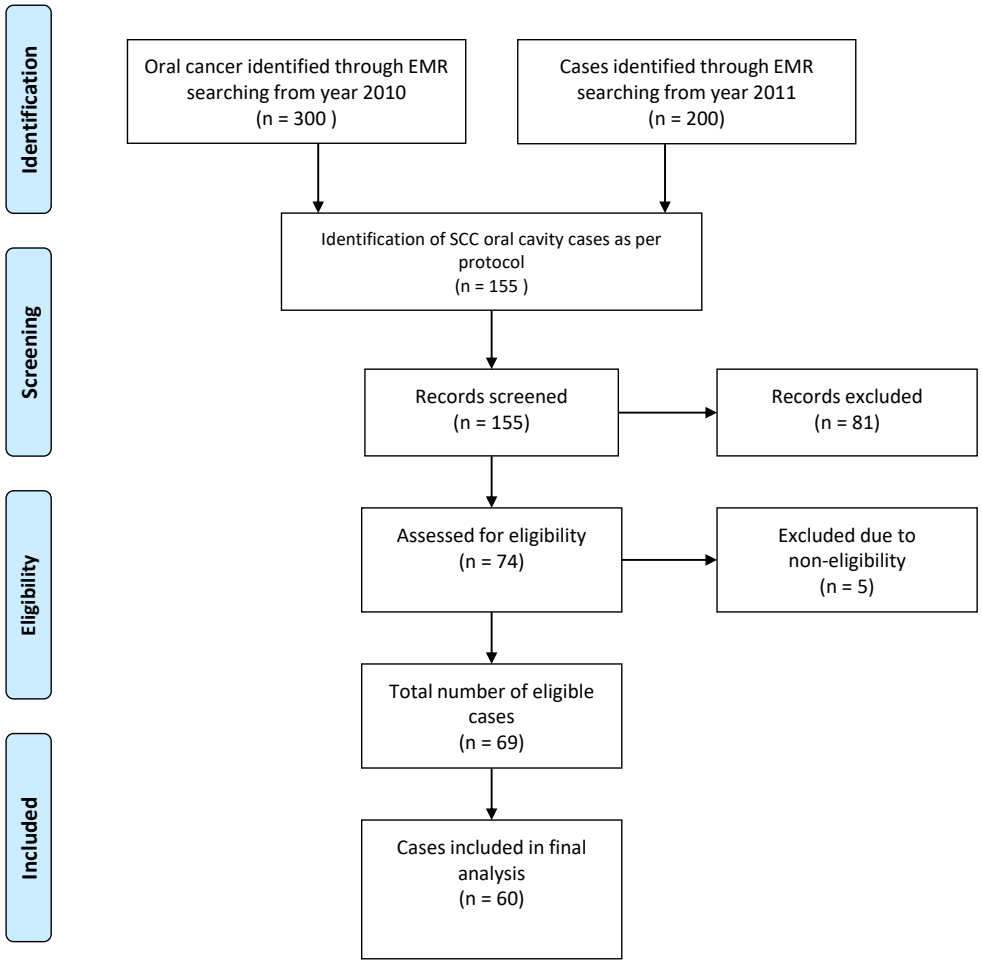


Figure 4.8: The Prisma flow diagram for patient selection. Eligible cases are those cases that underwent complete tumour resection with or without lymph node dissection and for which survival data were available. The cases excluded were those where either a complete resection was not done as this was needed to report all parameters and those where survival follow up of less than three years.

Table 4.1: Summary of clinical parameters of the OSCC Cohort.

Clinical Parameters		Full Cohort	Modelling Set	Test Set
No. of Patients		60	30	30
Age (years)		49.77 ± 10.99	50.57 ± 9.77	48.97 ± 12.03
Survival (months)	Overall	54.78 ± 17.91	56.00 ± 17.75	53.47 ± 17.96
	Disease Free	48.87 ± 22.51	50.00 ± 22.44	48.00 ± 22.55
Gender	Male	36	17	19
	Female	24	13	11
Node-Stage	I/II	25	17	8
	III/IVa	35	13	22
Grade	I/II	48	23	25
	III	12	7	5
Growth Patterns	Type 1/2	21	13	8
	Type 3/4	39	17	22
TILs	Absent/Mild	32	17	15
	Moderate/Severe	28	13	15
Patient Status	Alive	47	24	23
	Dead	13	6	7
Disease Recurrence	Yes	19	9	10
	No	41	21	20

remaining are at stage III/IVa. Further details of all clinical parameters are given in Table 4.1.

4.3.4 Pathologist annotations

I split our OSCC cohort into two equal-sized subsets, one for modelling and the other for the test. Six cases from the modelling set were considered for validation of the proposed tissue region classifier. An oral and maxillofacial pathologist reviewed all the digitised WSIs and marked the ground truth at two different level of abstraction: TIL presence or absence on all the slides, and tumour/lymphocytic regions on the modelling subset. At a high level, the presence or absence of TILs in 193 tissue section from all WSIs was marked where 111 were TIL positive (T^+) and 82 TIL negative (T^-). For the classification of biologically significant regions, more than half a million regions (belonging to different classes such as a tumour, lymphocytes, and other) were marked in all WSIs of the modelling cohort. The annotations were then used for training and validation of the proposed method.

4.4 Results

I evaluate the performance of the proposed framework at three different levels. First, five different tissue region classifiers are compared using different evaluation metrics. Second, TIL detection performance is measured quantitatively. Finally, the prognostic significance of spatial TIL quantification is evaluated by disease-free and disease-specific survival of OSCC patients. The detail of

Table 4.2: Quantitative performance of five different tissue region classifiers on validation dataset of 100,000 patches.

Classifiers	Accuracy	Sensitivity	Specificity	F1-Score	AUC
TRC-1 (ResNet50 [43])	94.04	88.19	96.03	88.08	97.88
TRC-2 (DenseNet [81])	95.40	90.80	96.94	90.78	98.56
TRC-3 (Inception-v3 [44])	95.84	91.91	97.25	91.70	98.86
TRC-4 (Xception [120])	95.94	92.16	97.32	91.88	98.83
TRC-5 (MobileNet [82])	**96.31**	**92.66**	**97.55**	**92.62**	**98.91**

each level of evaluation is presented in the following sections.

4.4.1 Tissue Region Classification

In order to get a best multi-class tissue region classifier (TRC) model, I employ five state-of-the-art convolutional neural network models (ResNet50 [43], DenseNet [81], Inception [44], Xception [120] and MobileNet [82]), denoted as TRC-1 to TRC-5 respectively. Table 4.2 gives the quantitative performance of these TRCs for multi-class patch level classification on the validation dataset, whereas the inter-class confusion results are presented in Figure 4.9. Among the five classifiers, MobileNet (TRC-5) shows superior performance as compared to the other networks. It is a lightweight network that leverages separable convolutions to reduce the number of required parameters and computations, resulting in the consumption of relatively less memory and computational resources and making it an attractive choice for the processing of multi-gigapixel WSIs.

Visual results for tissue region classification obtained with TRC-5 are shown for illustration in Figure 4.10, where tumour, lymphocytic, *other* and non-ROI regions are shown in different colours. Lymphocytic regions are classified with the highest accuracy, whereas tumour and non-ROI regions show a slight overlap. In general, TRC-5 gives the best classification performance, as can also be seen in the precision-recall curve in Figure 4.11, which shows the relatively high true-positive rate and low false-positive rate for both tumour and lymphocyte areas.

4.4.2 TIL Detection

For the evaluation of the proposed framework, I used both the best performing (TRC-5) and the least performing (TRC-1) tissue region classifiers for down-stream analysis. Therefore, the colocalisation of tumour and lymphocytes for each tissue section is calculated based on the equation (4.1) by using both TRC-1 and TRC-5 prediction maps. Five different performance measures are used to evaluate the performance of TIL detection in tissue sections through the colocalisation score, as shown in Table 4.3. Colocalisation score based

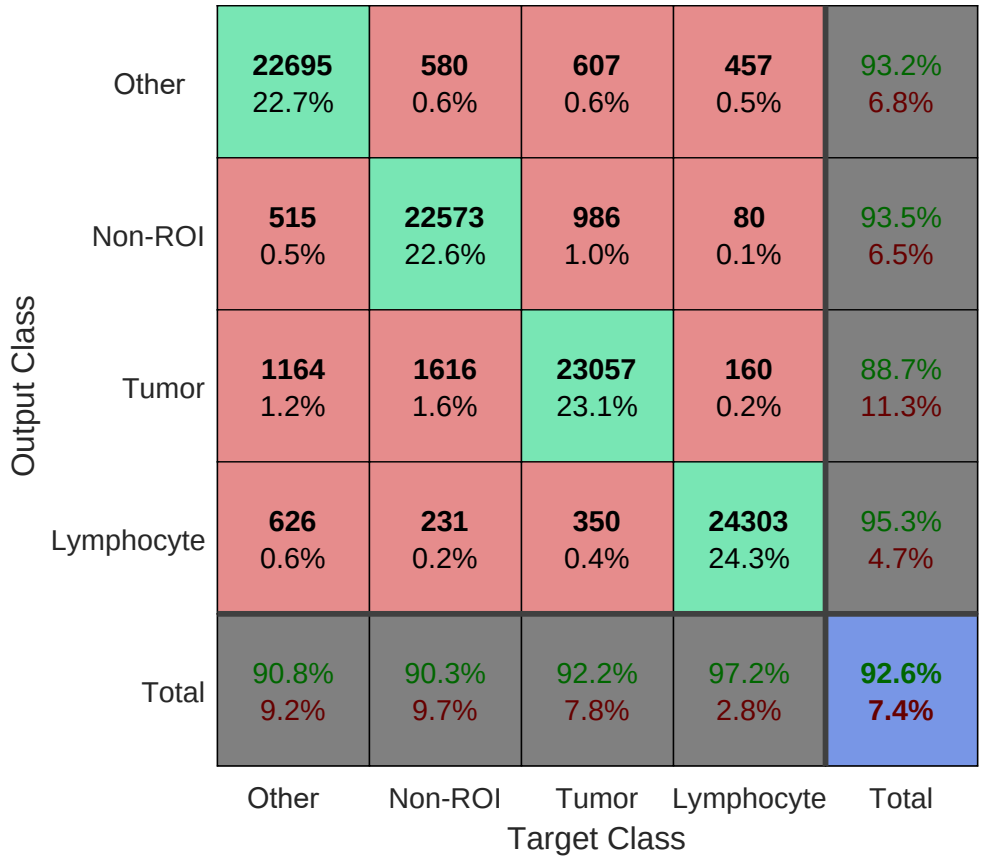


Figure 4.9: Confusion matrix for all four classes using best performing tissue region classifier model. Results show that the classifier classifies the lymphocytes with few false positive and false negative compared to other classes.

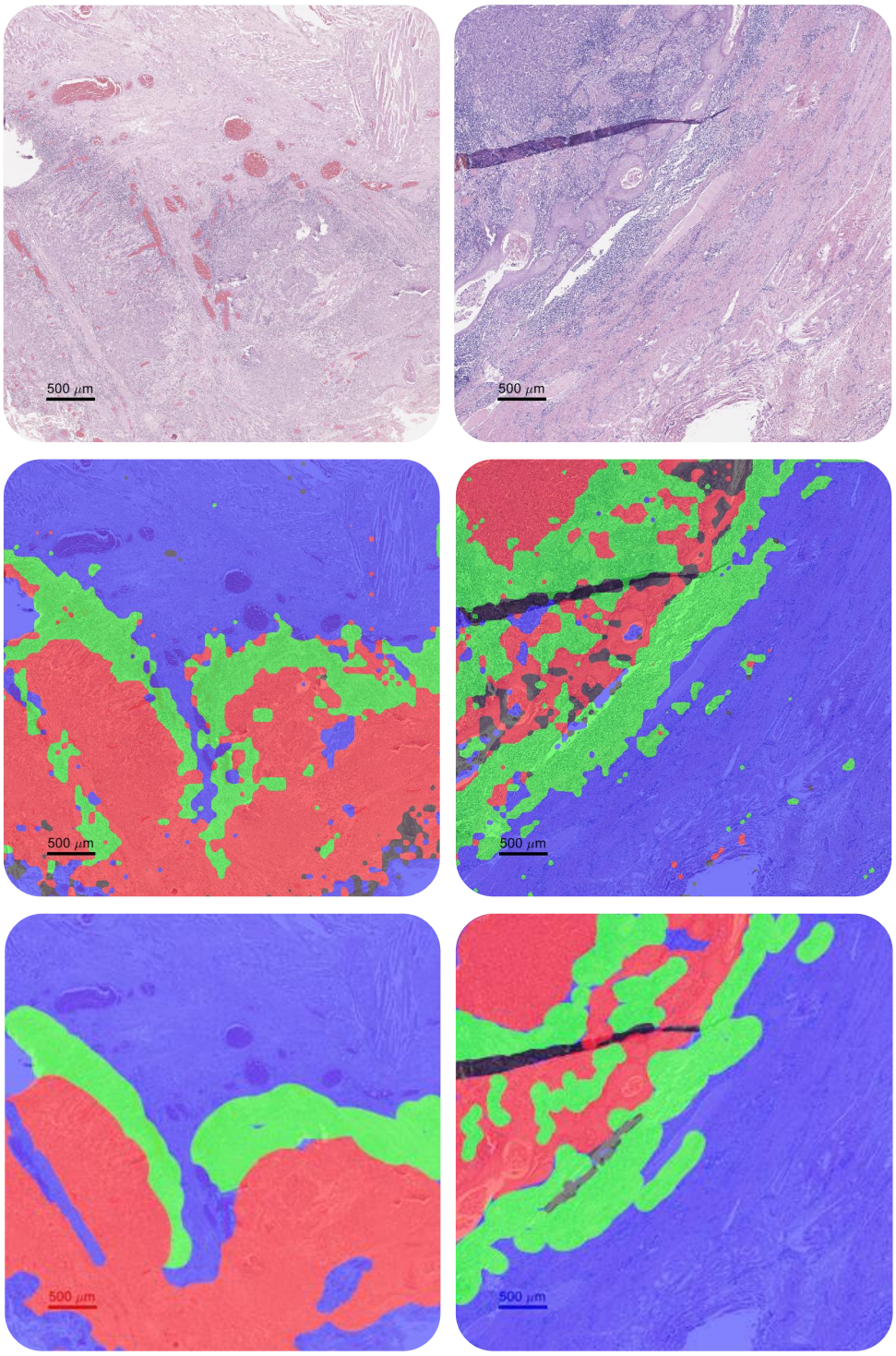


Figure 4.10: Tissue region classification results by TRC-5 where tumour, lymphocytic, other and non-ROI regions are represented by red, green, blue and black colours, respectively. Middle row presents classifier's predictions whereas bottom row represents ground truth labels of different regions.

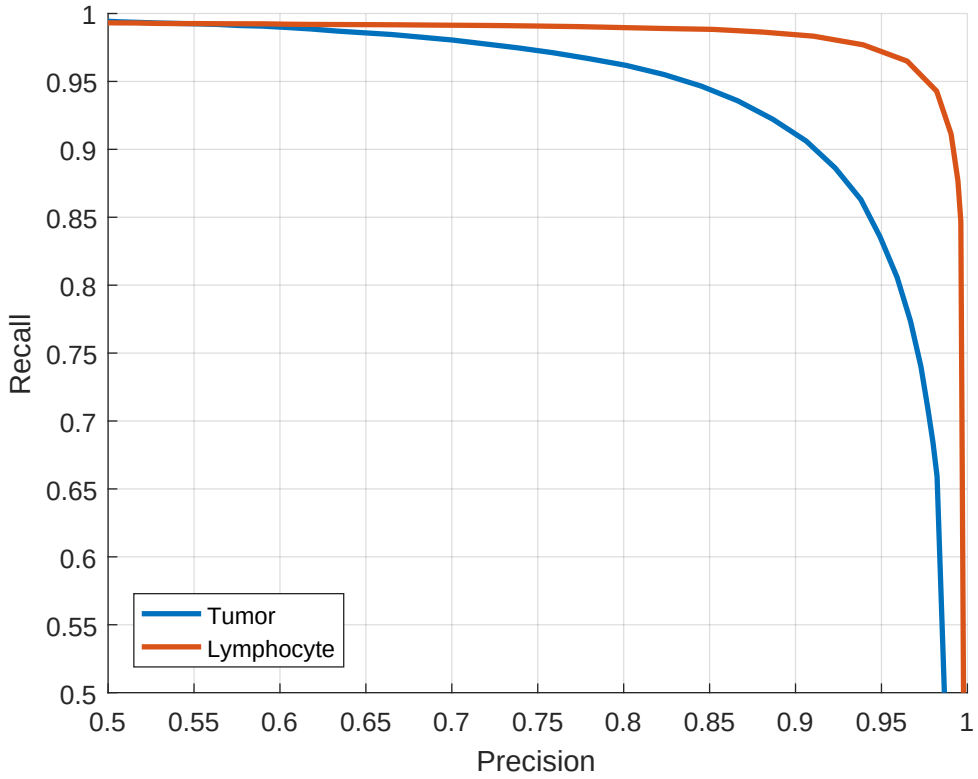


Figure 4.11: Precision-recall curves for tumour and lymphocytic region classification using the TRC-5 classifier. The TRC-5 classifier classifies the lymphocytic regions with better precision and recall compared to tumour regions.

Table 4.3: Performance of tissue section classification into TIL positives and negatives.

Classifiers	Accuracy	Sensitivity	Specificity	F1-Score	AUC (95% CI)
TRC-1	80.19	85.45	74.51	81.74	87.54 (80.95 - 94.12)
TRC-5	79.05	79.69	78.05	82.26	88.96 (82.68 - 95.13)

on the best performing region classifier (TRC-5) achieved 88.96% area under the curve (AUC). It is pertinent to mention that the least performing region classifier (TRC-1) also performed reasonably well with 87.54% AUC.

4.4.3 TIL Quantification

The TILAb score at WSI level is also computed according to equation (4.3) for spatial quantification of TILs. The TILAb score is evaluated by both visual and survival analysis. Figure 4.12 shows the colocalisation heatmaps of different tissue segments in two different WSIs along with WSI level TILAb scores. The WSI at the top shows high colocalisation and TILAb scores which is aligned with the spatial pattern of tumour and lymphocytes in the prediction map. However, the WSI at the bottom shows a lower value of TILAb score as there is less colocalisation of tumour and lymphocytes regions and low lymphocyte to tumour ratio.

The TILAb score based prognostic model is used to classify the OSCC patients into low- and high-risk groups for disease recurrence. The prognostic model finds the optimal cut-off point for TILAb score on the modelling subset and uses that cut-off on the test subset for binary classification. The TIL quantification methods have one hyper-parameter, which is the size of a grid-cell. I experimented with eight different neighbourhood sizes on tissue section to investigate their impact on survival analysis. The size of the smallest grid-cell is $0.28mm \times 0.28mm$, and I used a fixed step size of $0.14mm$ for increment in grid-cell size, up to the largest grid-cell of size $1.2mm \times 1.2mm$. Table 4.4 shows that our proposed TILAb score is statistically significant for all sizes with both best and least performing tissue region classifiers (TRC-1 and TRC-5). However, the Morisita-Horn (MH) and Shannon diversity (SD) indices based TIL quantification scores show significant results only for the smallest grid-cell size when using the best performing region classifier (TRC-5), which indicates the significance of TILAb score for disease-free survival analysis. Moreover, the high concordance indices of the prognostic models are also evidence of the predictive ability of proposed models. Figures 4.13 and 4.14 show the c-index value for both disease-free and disease-specific survival models. Models for disease-free survival achieve high c-index value as compare to the disease-specific survival models. On the other hand, models with different colocalisation methods (TILAb-MH, TILAb-SD) show similar result patterns

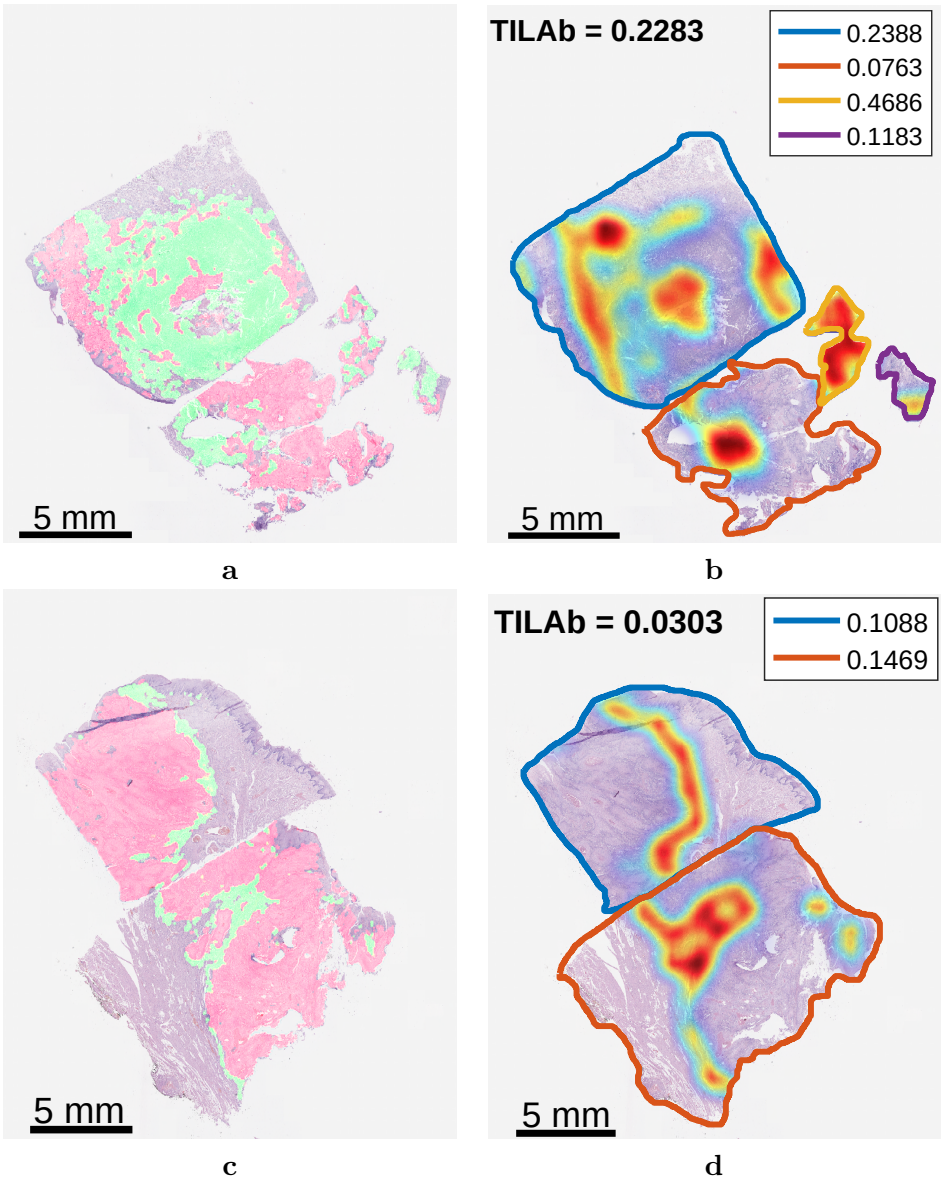


Figure 4.12: Visualisation of colocalisation score as heatmap. (**a, c**) Whole slide images at low resolution (1.5×) with tumour and lymphocytic region predictions overlaid in red and green colours, respectively. (**b, d**) Tumour-lymphocyte colocalization maps along with colocalization score for each tissue section in the upper right corner and WSI level TILAb score. Colour codes map the colocalization score to respective tissue sections.

Table 4.4: Comparison of the different TIL quantification methods based on their prognostic significance (logrank test based p-values) in eight experiments (1-8) with different grid-cell sizes (smallest to largest).

Region Classifier	Quantification Methods	1	2	3	4	5	6	7	8
TRC-1	MH	0.1590	0.1610	0.1560	0.3250	0.2340	0.4760	0.4760	0.4760
	SD	0.1590	0.1610	0.1610	0.3250	0.2190	0.6550	0.4760	0.4760
	TILAb-MH	**0.0146**	**0.0258**	**0.0146**	**0.0012***	**0.0258**	**0.0006***	**0.0020***	**0.0006****
	TILAb-SD	**0.0146**	**0.0146**	**0.0258**	**0.0146**	**0.0258**	**0.0258**	**0.0006****	**0.0006****
TRC-5	MH	**0.0416**	0.0666	0.1030	0.1800	0.0666	0.1800	0.2340	0.1790
	SD	**0.0416**	0.0666	0.0666	0.1160	0.0666	0.1800	0.2340	0.1790
	TILAb-MH	**0.0191**	**0.0191**	**0.0077***	**0.0077***	**0.0258**	**0.0236**	**0.0020***	**0.0038***
	TILAb-SD	**0.0191**	**0.0359**	**0.0146**	**0.0258**	**0.0110**	**0.0146**	**0.0020***	**0.0020***

Significance codes: **0.05**, 0.01*, 0.001**

for disease-free and disease-specific survival.

4.4.4 Survival Analysis

The prognostic significance of TILAb score for disease-free survival is investigated using Kaplan-Meier curves and Cox hazard analyses by conducting the univariate and multivariate analysis of digital, clinical, and pathological parameters. Kaplan-Meier curves in Figure 4.15 show that the proposed TILAb score is significantly associated with long term (low risk) disease-free survival of OSCC patients ($p = 0.00062$). However, the lymphocytic percentage in a WSI without any correlation with tumour does not show any statistical significance. The proposed digital TILAb score has better statistical significance as compared to the manual TIL score given by expert pathologists after visual inspection. Kaplan-Meier curves for other clinical and pathological parameters are shown in Figure 4.16.

Results of the univariate analysis of the prognostic significance of digital, clinical and pathological parameters on the test subset are shown in Figure 4.17. I employed Cox proportional hazards method for univariate analysis for both quantitative and categorical predictor variables. The clinical parameters do not show any significant correlation with disease-free survival, with the confidence interval range of hazard ratios (lower and upper 95% bounds) being quite large, except for age. The pathological parameters show better association with disease-free survival as compared to clinical parameters, especially tumour grade and manual quantification of TILs. Among digital scores, the proposed TILAb score is shown to be statistically significant ($p = 0.0065$) with hazard ratio of 0.0001 ($1.446 \times 10^7 - 0.0769$). I also investigate the prognostic value of the TILAb score in the context of other pathological parameters such as grade, stage and patterns, as shown in Table 4.5. For this purpose, I conduct the multivariate Cox proportional hazards analysis using the TILAb score adjusted by other histological features. The results in Table 4.5 show that TILAb score is independent of pathological parameters, e.g. grade, stage and pattern of

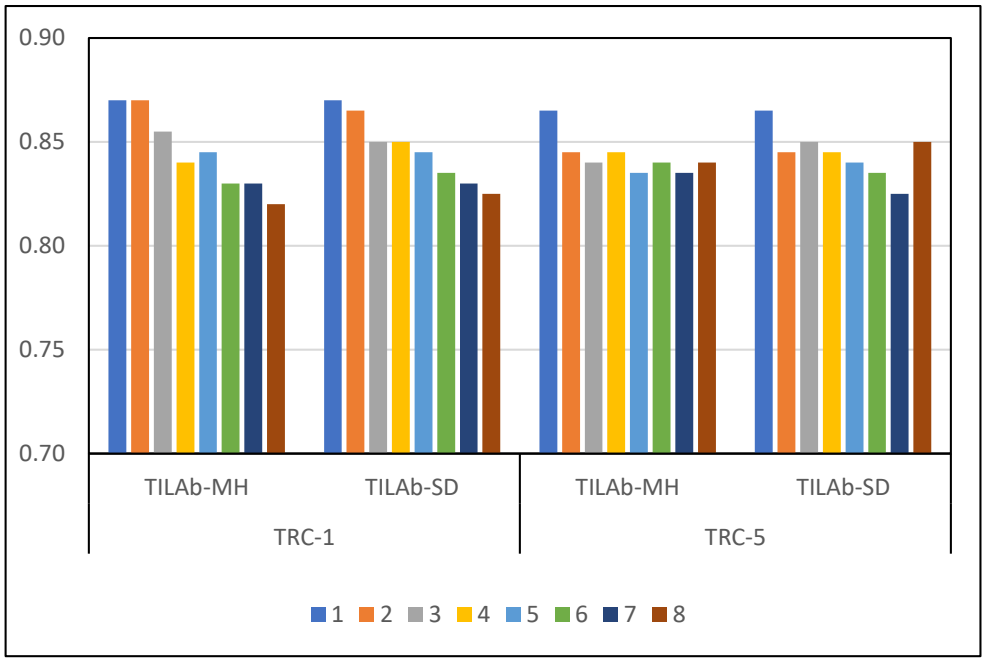


Figure 4.13: C-Indices of TRC-1 and TRC-5 based prognostic models for disease-free survival in eight experiments (1-8) with different grid-cell sizes (smallest to largest).

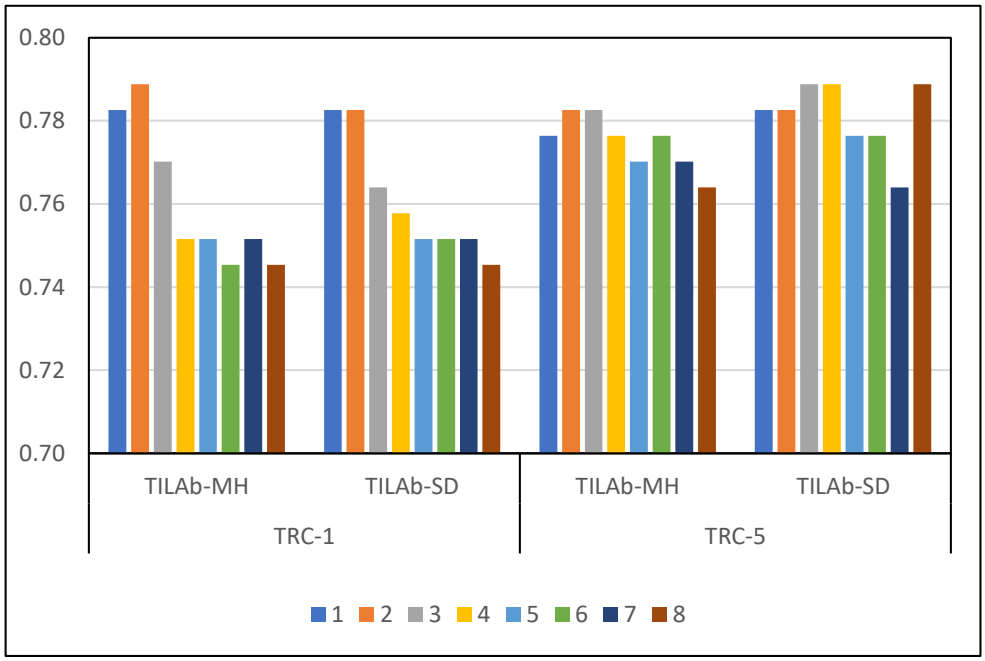


Figure 4.14: C-Indices of TRC-1 and TRC-5 based prognostic models for disease-specific survival in eight experiments (1-8) with different grid-cell sizes (smallest to largest).

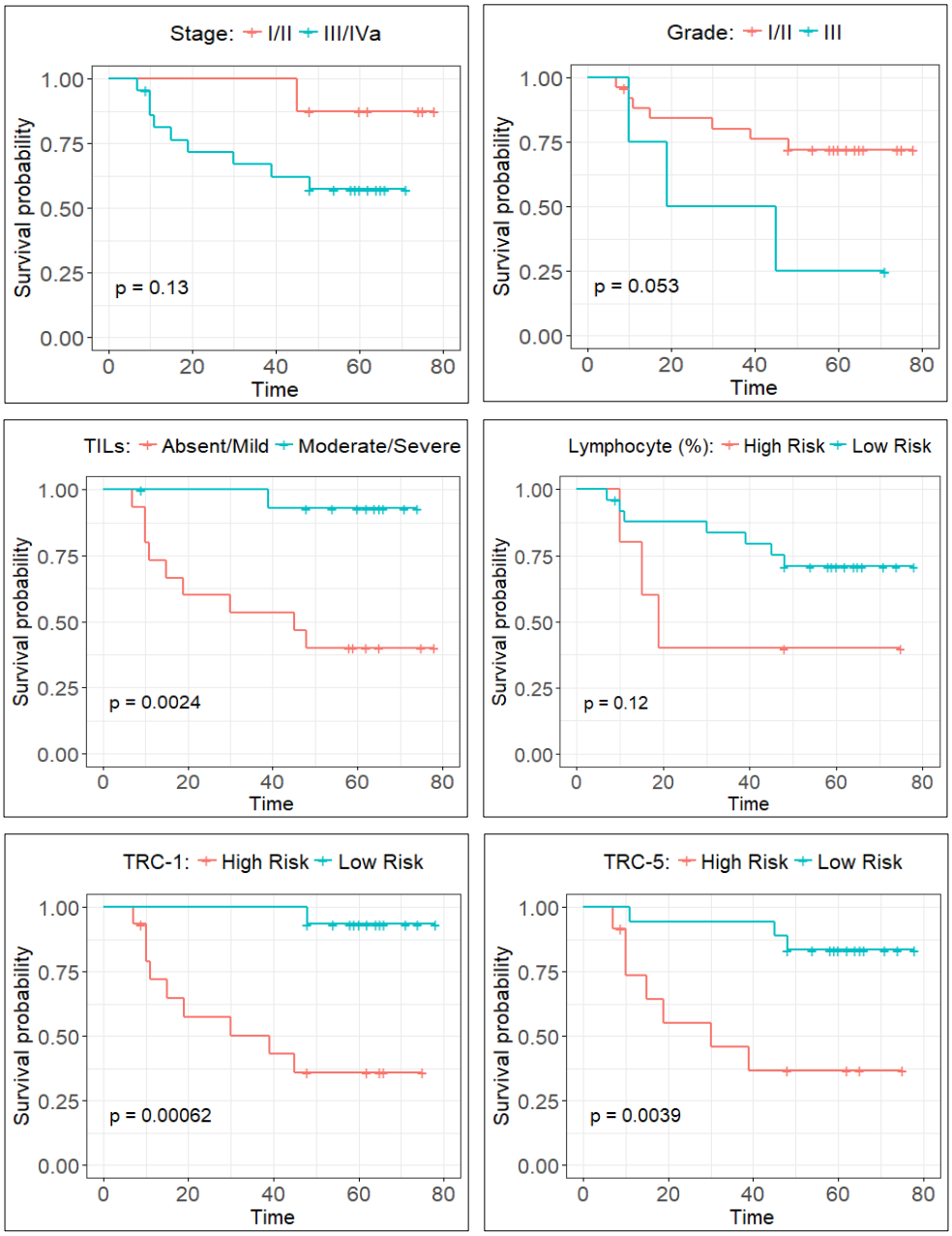


Figure 4.15: Kaplan-Meier curves for disease-free survival of OSCC on test subset. First three are the Kaplan-Meier curves for pathological parameters (stage, grade, and manual TIL quantification) whereas last three are the Kaplan-Meier curves of digital parameters (Lymphocyte percentage in WSI, TILAb score using TRC-1 and TRC-5). It should be noticed that the optimal cut-point values for digital parameters are 0.017, 0.124 and 0.137, respectively.

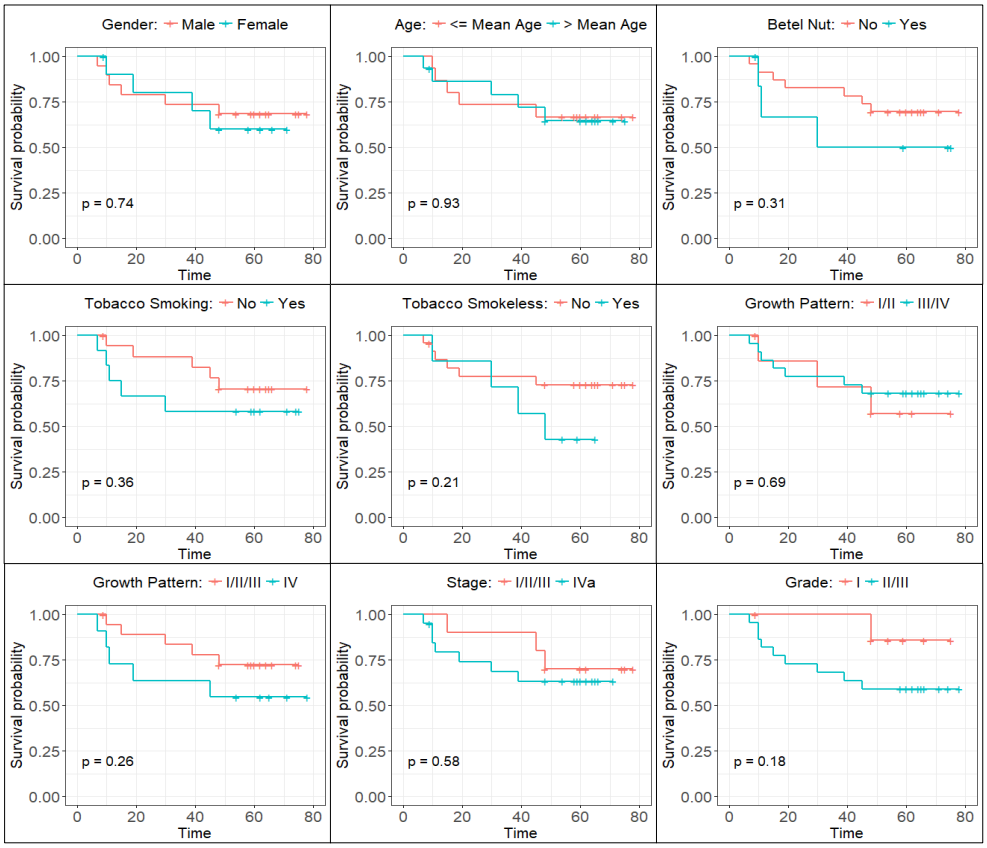


Figure 4.16: Kaplan-Meier curves of clinical and pathological parameters for disease-free survival of OSCC on test subset.

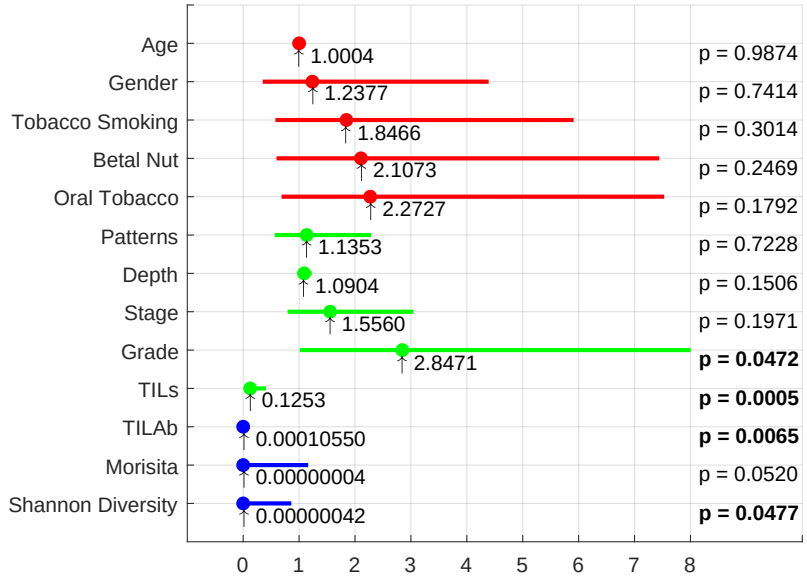


Figure 4.17: Univariate analysis for clinical (red), pathological (green) and digital (blue) parameters. Hazard ratios are represented by a filled circle along the x-axis, whereas the edges of each line represent the lower and upper confidence interval of 0.95%. P-value using the Wald test is shown on the right end for each parameter. Digital parameters are computed using TRC-5 predictions.

invasion.

4.5 Discussion

The presence of lymphocytes in the vicinity of tumour cells has been reported to carry high prognostic value [131, 132]. Quantification of TILs can not only significantly supplement clinical cancer staging information but could be used as an accurate predictor of disease progression [27, 28, 31]. The abundance of TILs in a tissue slide (or its digitised WSI) indicates the host immune response against cancer and/or response to treatment. The density and spatial arrangement of TILs are correlated with improved disease-specific survival and longer disease-free survival. However, the manual quantification of TIL is subjective, leading to inter-/intra- observer variability and lacking diagnostic reproducibility.

I propose a deep learning based approach for the identification and quantification of TILs in OSCC cases. A digital score of TIL abundance is computed, and its prognostic potential is investigated for disease-free survival in OSCC patients. The biologically significant regions in tissue such as tumour and lymphocytes are classified using a CNN. Several techniques are available in literature for detection and classification of histological structures in WSI images [48, 50, 90, 139–141]. However, very few are used for downstream prognostic

Table 4.5: Multivariate analysis of TILAb score along with other clinical parameters. TILAb score is computed with the Morisita-Horn as colocalization measure on TRC-5 predictions, while the p-value is computed using the Wald test.

	p	HR	Lower 95%	Upper 95%
	A - Overall Significance (**0.0334**)			
TILAb	**0.0103**	3.423×10^5	1.321×10^8	0.0887
Grade	0.3625	1.9720	0.4574	8.5003
Stage	0.2307	1.5280	0.7637	3.0587
Pattern	0.9387	1.0370	0.4105	2.6196
	B - Overall Significance (**0.0090**)			
TILAb	**0.0085**	6.701×10^5	5.232×10^8	0.0858
Grade	0.0745	2.3810	0.9177	6.1798
	C - Overall Significance (**0.0128**)			
TILAb	**0.0061**	3.267×10^5	2.044×10^8	0.0522
Stage	0.1377	1.6610	0.8499	3.2466
	D - Overall Significance (**0.0105**)			
TILAb	**0.0038**	2.243×10^5	1.610×10^8	0.0313
Pattern	0.1324	1.6750	0.8555	3.2803

analysis for disease-free survival. In this study, the results of the tumour and lymphocytic region classification are used to compute the TILAb followed by its evaluation as a prognostic marker. For tissue region classification, I experiment with different state-of-the-art CNN based image classifiers. I have chosen the classifiers giving the highest (TRC-5) and the lowest (TRC-1) patch level classification accuracy for further analysis of TIL detection, computing the abundance score and survival analysis. The results obtained by both of the classifiers are statistically significant.

The prognostic significance of TILAb score for disease-free survival is investigated by employing univariate and multivariate analyses using clinicopathological parameters. I analyse the prognostic significance of the TILAb score using the Cox proportional hazard model. The TILAb score shows good statistical significance in both univariate (Table 4.4) and multivariate (Table 4.5) analyses ($p < 0.05$). Therefore, the TILAb score can be used as an independent prognostic parameter in OSCC patients. The Kaplan-Meier curves showed the ability of TILAb score to stratify patients into long-term (low risk) and short-term (high risk) disease-free survival ($p = 0.0006$). Although the main focus of the work is on disease-free survival, the prognostic significance of TILAb score for disease-specific survival is also investigated. The Kaplan Meier curves are shown in Figure 4.18, which illustrates that the TILAb score gives good separation for disease-specific survival too. The 3-fold cross-validation with random initial image selection is used for disease-free survival to highlight the independence of the proposed model on initial image

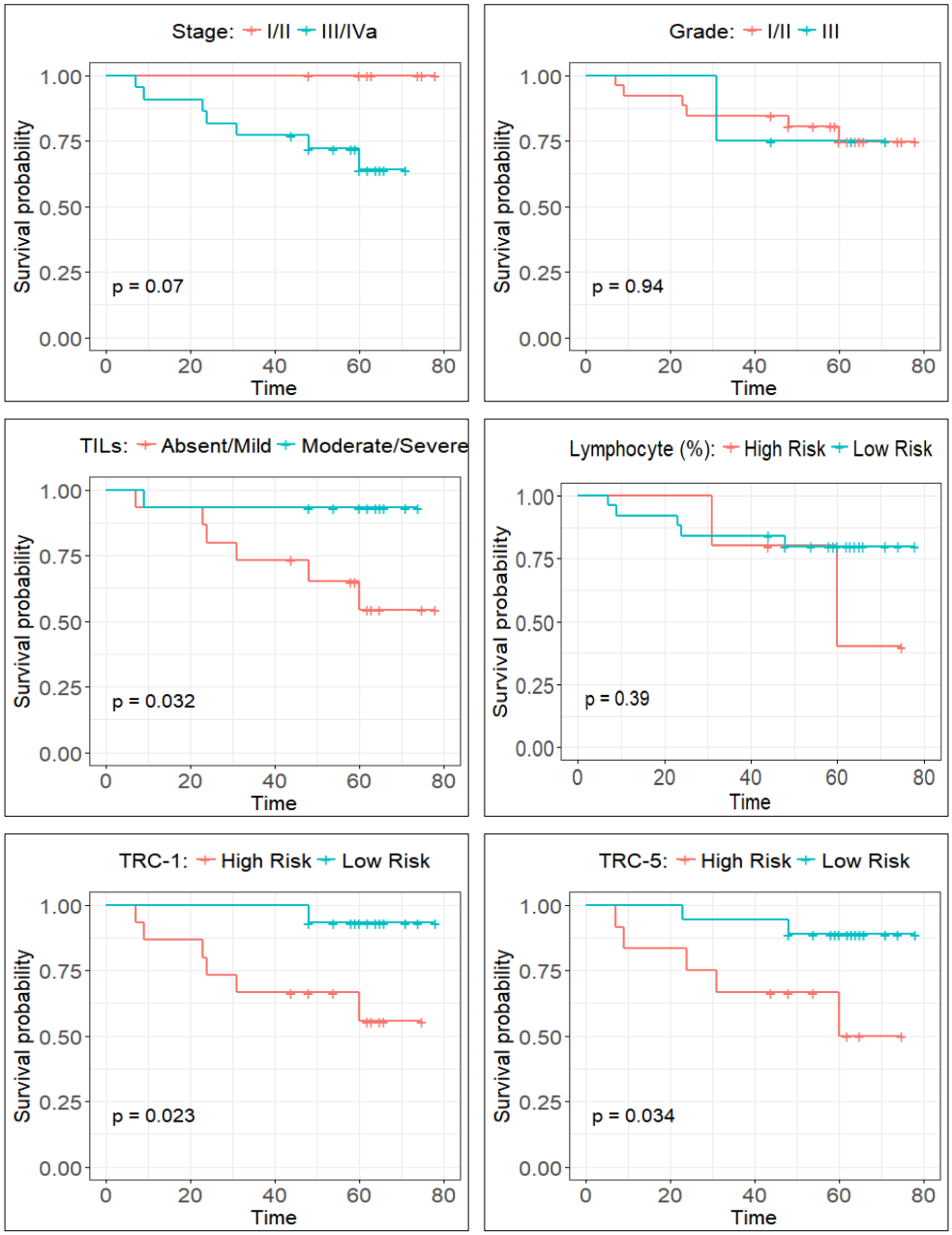


Figure 4.18: Kaplan-Meier curves for disease-specific survival of OSCC on test subset. Top row contains the Kaplan-Meier curves for pathological parameters (stage, grade and manual TIL quantification) whereas bottom bottom row shows the Kaplan-Meier curves of digital parameters (Lymphocyte percentage in WSI, TILAb score using TRC-1 and TRC-5). It should be noticed that the optimal cut-point values for digital parameters are 0.017, 0.124 and 0.137, respectively.

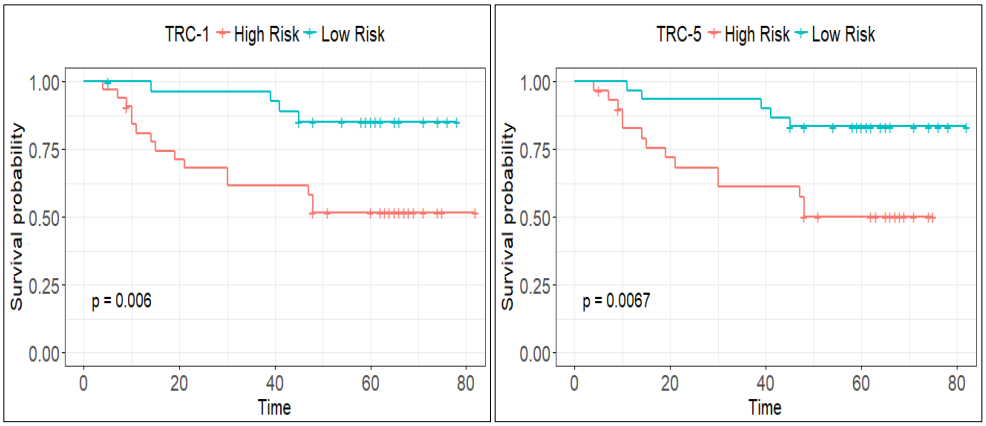


Figure 4.19: Kaplan-Meier curves for disease-free survival of OSCC on 3-fold cross-validation using TRC-1 and TRC-5.

selection for discovery and validation splits. The Kaplan-Meier curves and associated p-values, shown in Figure 4.19, illustrate that the proposed models (TRC-1 and TRC-5) are prognostically significant for different sets of discovery and validation splits. Moreover, to show the robustness of our method, the C-indices (with 95% confidence intervals) of prognostic models for disease-free survival and disease-specific survival are also computed and shown in Figures 4.13 and 4.14, respectively. The results of the proposed method are also in agreement with previous findings based on manual and immunohistochemistry based TIL quantification [135, 142, 143] in OSCC. Fang *et al.* [143] analysed the prognostic significance of tumour infiltrating immune cell in OSCC. The immune cells were identified by their specific markers (CD8, CD4, T-bet, CD68 and CD57). High CD8 (T-cells) and CD57 (NK-cell) expression were significantly associated with longer survival.

Hematoxylin and eosin staining is routinely used in pathology labs around the globe in clinical practice for cancer diagnostics. Automated methods for extracting information related to TILs from the whole slide images can help in treatment planning according to the immune response. The proposed framework for automated quantification of TILs, computation of their abundance score, and its prognostic analysis of patient survival using OSCC histology images is the first of its kind. Even though the total number of cases involved in this study is limited ($n = 70$), some other studies have reported results on smaller cohorts [144] ($n = 48$) or using tissue microarrays [145], which contain much smaller snapshots of tumour and lymphocytes characteristics as compared to the whole slide images. Having said that, the results of this study need to be cross-validated on data from sizeable multi-centric patient cohorts before they can be adopted in clinical practice.

In addition to the application to cancer resections and information about future behaviour, our proposed TILAb score can be applied to the initial biopsy

specimen undertaken before surgical resection or chemoradiotherapy. A biopsy and histological assessment is the gold standard for pre-operative diagnosis and a prerequisite for staging. As part of this assessment, pathologists report the presence/absence and comment on the density of the host lymphocytic response. The TILAb score can provide an objective quantification on this initial biopsy providing vital information about prognosis to the clinical team with the potential to guide treatment decisions and risk stratification.

Chapter 5

Coarse Segmentation of Histology Images for Profiling of Tumour Microenvironment

5.1 Introduction

Tumour microenvironment (TME) is the environment around the tumour which consists of stromal cells, immune cells and extracellular matrix components [146]. TME is known to influence the tumour growth positively or negatively depending on the state of its components and their interaction with each other [147–149]; therefore, TME profiling becomes important for better patient prognosis. Objective quantification of different TME constituents helps to profile the TME, which may then lead to the prediction of tumour behaviour [60, 150].

Profiling of TME in histology images requires localisation of its components, followed by quantification of their abundance and spatial interactions with each other. Convolutional neural networks (CNNs) have been used for the segmentation of varying tissue objects and components in histology images such as nuclei [86], cells [89], glands [53], and tissue sub-types [151]. Generally, segmentation of large objects, e.g. glands, require large contextual information whereas segmentation of small objects, e.g. nuclei and cells, require high-resolution appearance for precise segmentation. High precision in object segmentation is crucial for morphometric analysis [58]. However, patch-based segmentation of different tissue types is enough for whole slide image (WSI) level analysis of histology tissues such as TME profiling [97, 152], mutation prediction [59].

Patch-based segmentation of histology images is the most commonly used method for histology image analysis [50, 57, 153]. In patch-based segmentation, a single label is assigned to each patch instead of each pixel of a histology

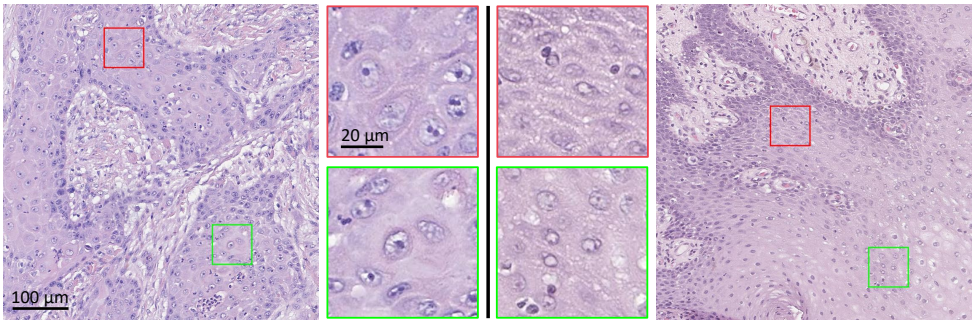


Figure 5.1: Small images in the middle show the amount of context captured by a patch of 256 × 256 at 40× magnification. These patches with limited context are less discriminative as compared to their corresponding images with larger context.

image and most common patch size is 224 × 224 pixels [50, 153]. However, the visual appearance of some tissue types in a patch is quite similar to other tissue types; therefore, their spatial context becomes key for correct prediction. For instance, classification of the normal and malignant epithelium in head and neck squamous cell carcinoma (HNSCC) requires a broader spatial context than the spatial context captured by the input patch of a standard patch classifier, as illustrated in Figure 5.1. The use of lower resolution/magnification patches is the most straightforward approach to increase the spatial context within input patches. However, this approach will result in less certain segmentation maps as lower resolution patch may contain other types of tissue as well (see Figure 5.2). In histology landscape, different tissue components appear in various sizes at different locations. Whenever the size of these components is much smaller than the patch size, then the less certain segmentation issue will arise.

I proposed a novel coarse segmentation method to overcome the issue of limited context and less certain segmentation issue. Unlike patch-based segmentation methods, the proposed method predicts a label for each 32 × 32 pixel region in a patch of size 256 × 256 pixels, which generates 64 times denser prediction map than a standard patch classifier. The dense prediction ability of our method enables it to take patches at low resolution (e.g. 20×, or 10×) to incorporate a broader context without introducing noise in tissue segmentation. The proposed method does not require pixel-level ground truth and the use of sparse weighted loss function enable it to learn from partially annotated images during training. The proposed method takes the same amount of memory and time as compared to standard patch-based segmentation methods but with the added advantage of better and denser segmentation. Our method achieved 4% better segmentation performance as compared to standard patch classifiers. I used the proposed method for coarse segmentation of HNSCC WSIs into

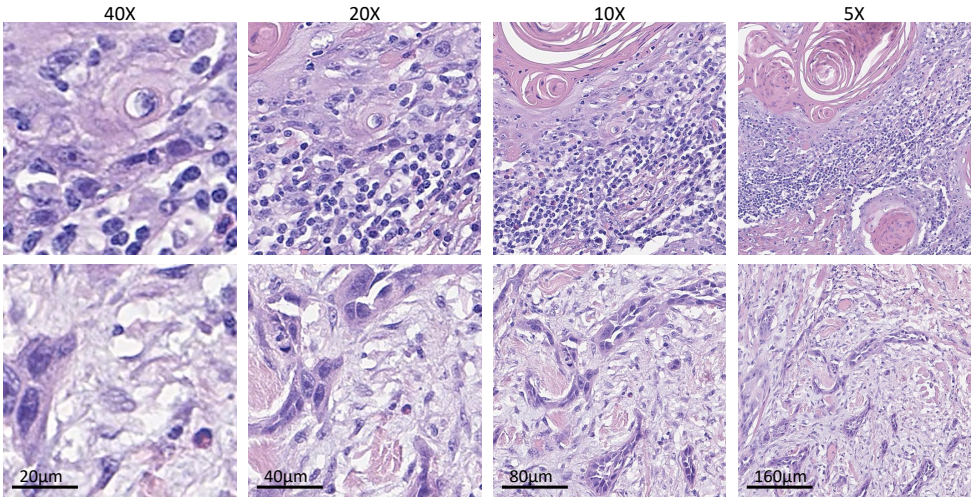


Figure 5.2: Illustration of two 256×256 patches at different resolution with different types of tissue regions. Patch based segmentation by using lower resolution patches will results in less certain segmentation map as each patch may contains multiple type of tissue regions (tumour, tumour-associated stroma, and lymphocyte).

clinically significant tissue types to profile TME of HNSCC.

I quantify different spatial patterns of the tumour, tumour-associated stroma, and lymphocytes to profile the TME of HNSCC. I proposed novel quantification measures for the quantification of lymphocyte infiltration into tumour and tumour-associated stroma. The proposed measure calculates the ratio between lymphocytes colocalisation with the tumour or tumour-associated stroma and overall tumour or tumour-associated stroma. I evaluate the prognostic significance of the proposed quantification method on three patient cohorts. Our proposed measure based tumour-associated stroma infiltrating lymphocyte (TASIL-Ratio) score shows prognostic significance (p-value=0.002) for better disease-specific survival of HNSCC patients. The TASIL-Ratio score remains prognostic indicator for disease-specific and disease-free survival of oral squamous cell carcinoma (OSCC) and oropharyngeal squamous cell carcinoma (OPSCC). I also compared the predictive ability of TASIL-Ratio based survival model with existing quantification methods through concordance index measure where TASIL-Ratio achieved the highest concordance score as compared to its counterparts. The TASIL-Ratio also shows a positive correlation with molecular estimates of CD8 T cells which kill the cancerous cells in the human body. The main highlights of this work are as follows:

- I propose a new coarse segmentation network which addresses the issues of limited context and less certain segmentation in patch-based segmentation methods.

- Our method does not require pixel-level ground truth and can learn from

partially annotated images as well.

- I profile the TME using different quantification methods including a novel TASIL-Ratio score.
- I evaluate the prognostic significance of the proposed quantification method on three different patient cohorts.
- Our proposed TASIL-Ratio score shows prognostic significance for disease-specific and disease-free survival on all three patient cohorts.

5.2 Method

Our TME profiling approach consists of two stages. First, segmentation of WSIs into different tissue types using a novel coarse segmentation method. Second, the calculation of different quantitative measures from clinically significant tissue types for TME profiling. The detailed descriptions of both stages are presented in the following sections.

5.2.1 Coarse Segmentation

There are two main motives behind the proposal of the novel coarse segmentation method as an alternative approach to current pixel or patch-based segmentation algorithms. First, the use of pixel-based segmentation methods add precise ground truth dependency for network training and requires longer inference time due to their memory and computation-intensive network architectures. Therefore, most of the existing work on TME analysis either used cell-based [60, 150] or patch-based [97] quantification measures. Second, the patch-based segmentation methods, usually faster than the pixel-based segmentation methods, are bound to predict only one label regardless of input image size. Therefore, the segmentation precision decreases with the increase of input image size or use of lower resolution image to incorporate larger contextual information. Hence, the prediction of each patch becomes less certain, as illustrated in Figure 5.2.

The concept of coarse segmentation is generic in term of network design and can be implemented by using any existing state-of-the-art patch classification network, e.g. DenseNet [81], ResNet[43], or MobileNet [82]. It takes an $M \times N$ patch as an input just like standard patch classifiers, but its output is an $m \times n$ coarse segmentation map of the input patch unlike a single patch label of standard patch classifiers. The m and n are eight times smaller than the M and N, respectively. Although its output is more like the output of segmentation networks, it does not contain any decoder module or up-sampling layers unlike

segmentation architectures [87, 154]. Figure 5.3 shows the DenseNet [81] based coarse segmentation network (CSNet), which is explained in the following sections.

Network Architecture

The proposed architecture presented in Figure 5.3 consists of multiple convolution, pooling, and feature concatenation layers. All the convolution layers preceded by a batch norm and ReLU based activation layers apart from the first one where these two layers are used after the convolution layer. The main building block of the proposed network is the dense block which consists of multiple pair of convolution layers where the depth of a block depends on the number of iterations selected for the block. In each iteration, the pair of convolution layers converts the input feature-map into 32 channels feature-map and concatenate it with the input feature-map. The last convolution layer followed by softmax layer takes the output feature-maps of all the dense blocks through skip connections after spatial average pooling, if required, and outputs the probability maps for the given number of classes.

Network Variants

I consider three variants of the proposed coarse segmentation approach which differs only in network architecture. The first variant, CSNet-121, is the simplest variant with a minimal difference from standard DenseNet-121 [81]. It replaces the last average-pooling and fully connected layer of DenseNet-121 with a 1×1 convolution layer which enables it to produce coarse prediction map instead of the single label for each input image. The second variant, CSNet-121-SC, is an extension of the former which feeds the average pooling based down-sampled features of the intermediate layers to the final convolution layer through skip connections (SC), as shown in Figure 5.3. The last variant, CSNet-61-SC, is a lightweight version of CSNet-121-SC as it uses DenseNet architecture with only 61 layers as a baseline.

Weighted Sparse Loss Function

The loss function of the most segmentation methods [87, 154] requires fully annotated images for error calculation during the model training. However, annotation of every region of histology images is not a trivial task. There are many regions which lie on the boundary of two different classes such as dysplastic regions which are neither normal nor malignant. Annotation of these regions introduces noise in the ground truth due to inter- and intra-observer variability. I use a sparse loss function which does not require fully annotated images and enable us to use partially annotated images with high confidence

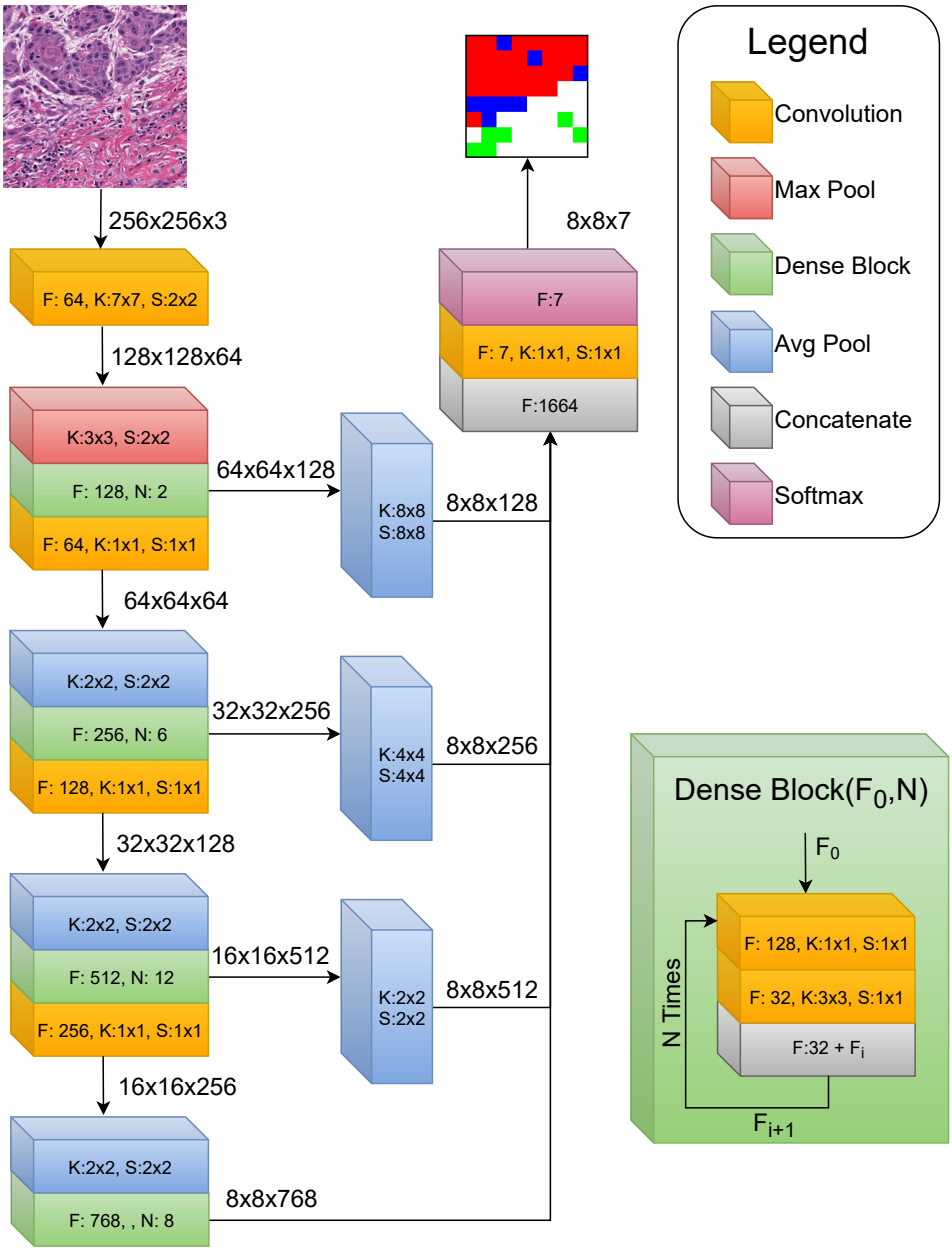


Figure 5.3: The architecture of the proposed coarse segmentation network using DenseNet as a baseline. Each box in the prediction map represents the prediction of a 32×32 corresponding region in the input patch. The letters N, F, K, and S represent the dense block depth, output feature maps, kernel size, and stride size, respectively.

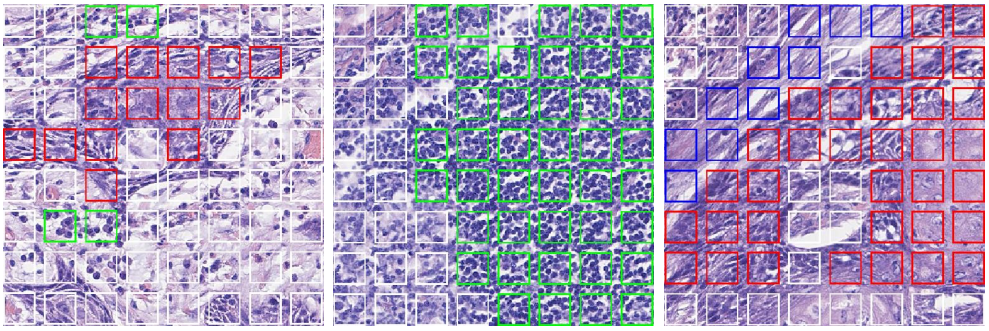


Figure 5.4: Three partially annotated images where red, green, blue, and white boxes represent tumour, lymphocytes, tumour-associated stroma and unannotated regions, respectively.

(Figure 5.4). However, the use of sparse loss worsens the natural class imbalance issue in any histology dataset. Therefore, I introduce a weighted sparse loss function which consists of categorical cross-entropy loss weighted by a weight-map. Zero weight is assigned to all unannotated regions, whereas batch-based weights are calculated for all the annotated regions in all images of a batch. More weight is given to classes with fewer number of regions to address the class imbalance issue in a batch. Let E be the expected count of the annotated number of regions of each class in a batch,

$$E = \frac{\sum_{i=1}^{C} r_i}{C'} \tag{5.1}$$

where r_i, and C represent the number of annotated regions of i^{th} class and total number classes in a dataset, respectively. C' denotes the number of classes which have more than one annotated region in the current batch. The weight of each region of i^{th} class is defined as:

$$W_i = \frac{E}{r_i} \tag{5.2}$$

W_i will be greater than one if the total number of regions of i^{th} class are less than the expected number of regions (E) and vise-versa. The final loss is the sum of the product of categorical cross-entropy loss and the weight-map.

Model Training

Each variant of the proposed coarse segmentation network is trained on 256×256 input patches at $10\times$ magnification and predicts a label for each 32×32 region in the input patches. The size of tissue in each input patch is $280 \times 280 \mu m$ and in each predicted region is $35 \times 35 \mu m$. Input patches are randomly augmented during the training with random rotation (0, 90, 180, and 270 degrees), random flipping (horizontal, vertical), random jittering (0-128 pixels) and random

colour perturbation. All models are trained with RMSProp optimiser for at least 1 million iterations (optimisation steps).

5.2.2 TME Profiling

I have used both existing and new methods for quantification of different spatial patterns in histology images to profile the TME of HNSCC. I only explore the spatial patterns of the tumour, tumour-associated stroma, and lymphocytes as some of these patterns have shown prognostic significance in many clinical studies of different tumour types [95, 105]. I quantify the percentage of a tissue type, the ratio of one tissue type to another tissue type, colocalisation of two different tissue types, the abundance of lymphocytes in the vicinity of another tissue type, and different patterns of adjacent tissue types.

Percentage and Ratio

The percentage and ratio based measures are the most straightforward quantification measures for TME profiling which rely on WSI level statistics and do not consider spatial patterns of tissue types. The percentage measure has shown prognostic significance in ovarian cancer [108]. However, ratio based quantification measures have shown prognostic significance in more than one tumour types [155, 156]. I calculate the percentage of a tumour (T-Percentage), tumour-associated stroma (TAS-Percentage), and lymphocytes (L-Percentage) to the total tissue in a WSI. For ratio based quantification, I consider tumour-associated stroma to tumour ratio (TAST-Ratio), lymphocyte to tumour ratio (LT-Ratio) and lymphocyte to tumour-associated stroma ratio (LTAS-Ratio).

Colocalisation

The co-occurrence of tumour and other tissue types has also shown the prognostic significance for a range of tumour types. Therefore, I quantify the co-occurrence of two different tissue types in a WSI using Morisita-Horn index [106], which is a measure of colocalisation in the ecological domain. The colocalisation measure calculates the co-occurrence of two different tissue types in fixed-size regions of a WSI to capture the spatial patterns of their co-occurrence. A given WSI is first divided into $r \times s$ small regions, and then the percentage of both tissue types is calculated for each region. The Morisita-Horn index is defined as:

$$M = \frac{2\sum_{i=1}^{r}\sum_{j=1}^{s}(p_{ij}^{c_1} \times p_{ij}^{c_2})}{\sum_{i=1}^{r}\sum_{j=1}^{s}(p_{ij}^{c_1})^2 + \sum_{i=1}^{r}\sum_{j=1}^{s}(p_{ij}^{c_2})^2}, \tag{5.3}$$

where $p_{ij}^{c_1}$ and $p_{ij}^{c_2}$ represent the percentage of two tissue types in the $(i,j)^{th}$ region. The value of M ranges from 0 to 1, where zero and one represent no and maximum colocalisation, respectively. I consider the colocalisation of

tumour-associated stroma and tumour (TAST-Col), lymphocyte and tumour (LT-Col), and lymphocyte and tumour-associated stroma (LTAS-Col) for TME profiling.

Lymphocyte Abundance

The higher lymphocytes infiltration in the tumour is associated with better patient survival. Therefore, I quantify the abundance of lymphocytes in the vicinity of another tissue type for TME profiling. The measure of lymphocyte abundance relies on both spatial colocalisation and ratio of lymphocyte to other tissue types. The abundance of lymphocytes in the vicinity of tumour or tumour-associated stroma is calculated by the abundance score [152]. The abundance score quantifies the abundance of lymphocytes in the vicinity of a given tissue type such as a tumour or tumour-associated stroma. It is defined as the product of lymphocytes to given tissue type ratio and their colocalisation. The formulation of abundance score using Morisita-Horn index as colocalisation measure is given below,

$$\mathrm{A} = \begin{cases} \frac{\sum_{i=1}^{r}\sum_{j=1}^{s}(p_{ij}^{l} \times p_{ij}^{c})}{\sum_{i=1}^{r}\sum_{j=1}^{s}(p_{ij}^{l})^2 + \sum_{i=1}^{r}\sum_{j=1}^{s}(p_{ij}^{c})^2} \times \frac{\sum_{i=1}^{r}\sum_{j=1}^{s} p_{ij}^{l}}{\sum_{i=1}^{r}\sum_{j=1}^{s} p_{ij}^{c}}, & \text{if} \sum_{i=1}^{r}\sum_{j=1}^{s} p_{ij}^{c} > 0 \\ 1, & \text{otherwise} \end{cases} \tag{5.4}$$

where p_{ij}^{l} and p_{ij}^{c} represents the percentage of lymphocyte and the given tissue type in $(ij)^{th}$ region. The value of A ranges from 0 to 1, where zero and one represent no and maximum lymphocyte abundance, respectively. I consider tumour infiltrating lymphocytes abundance (TILAb) and tumour-associated stroma infiltrating lymphocytes abundance (TASILAb) for TME profiling.

Proposed Quantification Measure

I formulate objective and automated scores for the quantification of tumour infiltrating lymphocytes to tumour ratio (TIL-Ratio) and tumour-associated stroma infiltrating lymphocytes to tumour-associated stroma ratio (TASIL-Ratio) using the statistics of adjacent tissue types in a WSI. First, six different patterns of adjacent tissue type are defined based on three clinically significant tissue types, as shown in Figure 5.5. Then the TIL-Ratio is defined as:

$$\text{TIL-Ratio} = \frac{TL}{TT + TL + ST} \tag{5.5}$$

where TL represents the number of times tumour and lymphocyte regions appear adjacent to each other in a WSI. Similarly, TT and ST denotes the number of times tumour regions appear adjacent to another tumour and tumour-associated stroma region, respectively. The TASIL scores is defined as:

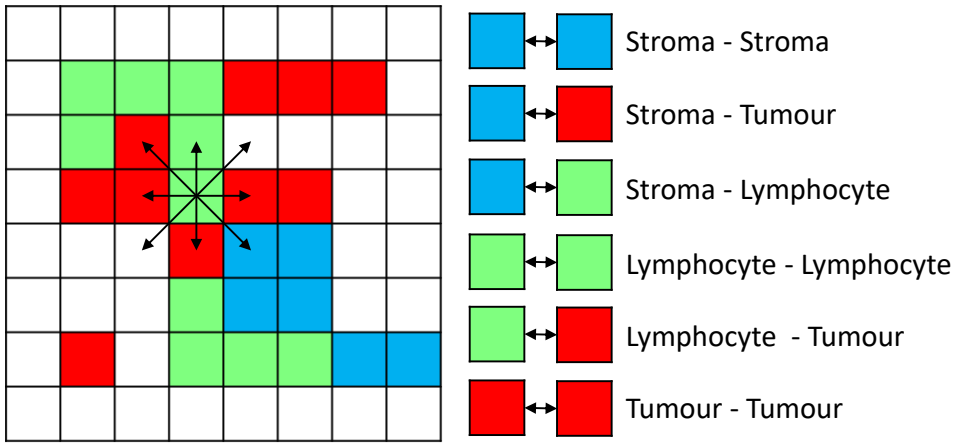


Figure 5.5: Visual illustration of different patterns of adjacent regions over a synthetic image. Right half of the figure list down the six different clinically significant patterns appeared in the synthetic image. Here, term stroma is used to refer tumour-associated stroma for the sake of brevity.

$$\text{TASIL-Ratio} = \frac{SL}{SS + SL + ST} \tag{5.6}$$

where SL and SS represent the number of times tumour-associated stroma regions appear adjacent to lymphocyte and tumour-associated stroma, respectively. Both TIL-Ratio and TASIL-Ratio range from zero to one, where zero represents no infiltration of lymphocytes, and one represents high infiltration of lymphocytes in tumour and tumour-associated stroma, respectively.

5.2.3 Statistical Analysis

The concordance index is used to compare the predictive ability of different automated quantification scores based survival models. Survival analysis is performed with disease-specific and disease-free survival data. The Kaplan–Meier estimator is used, and the log-rank test is performed to test differences among low and high-risk patient groups where log-rank test p-value < 0.05 is considered significant. The Cox proportional hazards regression model is fitted for univariate and multivariate analysis, and 95% confidence intervals computed to determine prognostic values. Spearman correlation is used for correlation analysis between TASIL-Ratio and molecular estimates.

5.3 Datasets

Three different patient cohorts (TCGA-HN [157], SKMCH&RC, PredicTR2) are used in this study. The TCGA-HN is a publicly available cohort of 528 HNSCC cases, whereas the other two are internal cohorts. The SKMCH&RC

cohort consists of 100 OSCC cases collected from one centre in Pakistan. PredicTR2 cohort contains 95 OPSCC cases collected from 6 different centres across the United Kingdom. The detailed description of patient selection criteria, patient characteristics and available clinical, pathological, and survival data is presented in the following sections.

5.3.1 Patient Selection

The TCGA-HN cohort contains diagnostic slides of 450 squamous cell carcinoma cases from different sites of head and neck. However, there are many slides with severe slide preparation and scanning artefacts. Therefore, I excluded the cases with poor quality of slides from overall cases. Our final cohort consists of 342 cases with one slide per case. The tissue slides of these cases are stained with H&E stains, and most of the cases are scanned at 40× magnification with slightly varying micron per pixel resolution, which varies from $0.23\mu m$ to $0.25\mu m$. However, some cases are scanned at 20× magnification with around $0.50\mu m$ per pixel.

The SKMCH&RC cohort is curated for our previous study on oral cavity cases, whereas PredicTR2 cohort is part of another multi-centre study on cases from the oropharyngeal site. The representative tissue sections from formalin-fixed and paraffin-embedded tissue blocks were collected for all cases. Tissue slides of all cases were stained with H&E stains and scanned at 40× magnification with $0.275\mu m$ per pixel resolution.

5.3.2 Patient Characteristics

The patients' survival information is available for most of the cases in three cohorts. The disease-specific survival time is calculated from the date of diagnosis to the date of death or the date of the last follow-up in case of censored data. The disease-free survival is censored at the date of first recurrence or death, whichever occurred first, or the date of the last contact for the patients alive without recurrent disease. The information about different clinical and pathological parameters is also available for all cohorts.

In TCGA-HN cohort, most of the patients were diagnosed between 2007 to 2013, and the average age of the patients is 61.09 years with a standard deviation of 11.82. There are more male patients ($n = 252$) as compared to female ($n = 90$) patients. The distribution of TNM-stage of the cases is a bit skewed toward higher stages with 52% cases of stage IVa. The ratio of alive and deceased patients is also imbalanced, with only 25% deceased cases, and the average disease-specific survival of all patients is 28.69 months with a standard deviation of 24 months. Detailed statistics of the cohort are presented in Table 5.1.

Table 5.1: Summary of available parameters of the TCGA-HN cohort along with log-rank test based p-values for disease-specific survival.

Categorical Parameters		Count	Percentage	DSS p-value
Number of Patients		342	100%	-
Gender	Male	252	74%	0.401
	Female	90	26%	
TNM Stage	Stage I	20	6%	0.143
	Stage II	48	14%	0.0851
	Stage III	50	14%	**0.0212**
	Stage IVa	177	52%	**0.00341**
	Stage IVb&c	10	3%	**0.00255**
	Not Reported	37	11%	-
Patient Status	Alive	255	74.6%	-
	Deceased	85	24.8%	
	Not Reported	2	0.6%	
Continuous Parameters		Mean	Standard Dev	p-value
Age (years)		61.09	11.82	0.647
Survival (Months)		28.69	24.00	-

In SKMCH&RC cohort, all the patients were diagnosed between 2010 to 2013, and the average age of the patients is around 50 years, with 11.12 years of standard deviation. The grade, growth pattern, and pathologists' manual TIL score information along with TNM stage is available for almost all the patients. The most dominant stage is stage-IVa, and grade is grade-II in the cohort. The manual TIL score was assigned to each case by an expert pathologist based on the amount of TIL infiltration, and it is categorised into four groups (absent, low, moderate, and high). The low and moderate TIL groups show the prognostic significance for both disease-specific and disease-free survival. In terms of survival, most patients were alive until the last follow-up; however, 32 patients have suffered from disease recurrence. Table 5.2 presents a detailed description of all the available parameters of the cohort and their prognostic significance, if applicable.

Patients in PredicTR2 cohort were diagnosed between 2000 to 2010, and the patients were tracked until 2014. The cases in this cohort are collected from six different data centres with a minimum of nine and a maximum of 23 cases from a data centre. The disease-specific survival information is available for 84 cases, whereas disease-free information is available only for 77 cases. There are 24 recurrent and 25 deceased cases out of cases with survival data. Unlike SKMCH&RC cohort, TILs are manually scored into only three categories low, moderate and high. Presence of lymphocytes in 80% or more of tumour/stroma is categorised as high TILs and lymphocytes in less than 20% of tumour/stroma is denoted by mild TILs. Unlike SKMCH&RC cohort, the low and high groups have prognostic significance in this cohort instead of low and moderate, which

Table 5.2: Summary of available parameters of the SKMCH&RC cohort along with log-test based p-values for disease-specific survival (DSS) and disease-free survival (DFS).

Categorical Parameters		Count	Percentage	DSS p-value	DFS p-value
Number of Patients		100	100%	-	-
Gender	Male	57	57%	0.259	0.196
	Female	43	43%		
TNM Stage	I	25	25%	**0.0203**	0.183
	II	14	14%	0.982	0.738
	III	15	15%	0.813	0.914
	IVa	43	43%	**0.0211**	0.124
	Not Reported	3	3%	-	-
Grade	I	35	35%	0.484	0.54
	II	50	50%	0.996	0.769
	III	15	15%	0.363	0.688
Growth Pattern	I	18	18%	0.25	0.0792
	II	18	18%	0.212	0.156
	III	35	35%	0.238	0.24
	IV	28	28%	0.209	0.0805
	Not Reported	1	1%	-	-
Manual TIL Score	Absent	10	10%	0.574	0.23
	Low	34	34%	**0.00013**	**0.0188**
	Moderate	47	47%	**0.00019**	**0.0148**
	High	9	9%	0.792	0.503
Patient Status	Alive	86	86%	-	-
	Deceased	14	14%		
Disease Recurrence	Yes	32	32%	-	-
	No	68	68%		
Continuous Parameters		Mean	Standard Dev	DSS p-value	DFS p-value
Age (years)		49.57	11.12	0.861	0.364
Survival (Months)	Overall	60.10	17.75	-	-
	Disease Free	53.56	22.29		

Table 5.3: Summary of available parameters of the PredicTR2 cohort along with log-rank test based p-values for disease-specific survival and disease-free survival.

Categorical Parameters		Count	Percentage	DSS p-value	DFS p-value
Number of Patients		95	100%	-	-
Gender	Male	61	64%	0.13	0.239
	Female	34	36%		
Manual TIL Score	Low	22	23%	**0.0189**	**0.00948**
	Moderate	37	39%	0.414	0.445
	High	36	38%	**0.00811**	**0.00632**
Patient Status	Alive	59	62%	-	-
	Deceased	25	26%		
	Not Reported	11	12%		
Disease Recurrence	Yes	24	25%	-	-
	No	53	56%		
	Not Reported	18	19%		
Continuous Parameters		Mean	Standard Dev	DSS p-value	DFS p-value
Age (years)		57.74	11.50	0.0649	0.134
Survival (Months)	Overall	47.22	30.33	-	-
	Disease Free	47.25	29.12		

may be due to the inter-observer variability in TIL scoring. Table 5.3 presents a detailed description of all the available parameters of the cohort along with the prognostic significance where applicable.

5.3.3 Pathologist Annotations

I used 24 cases, one WSI per case, for training and evaluation of the coarse segmentation method. Half of the cases are taken from the TCGA-HN cohort, and the remaining half are selected from SKMCH&RC cohort. Multiple visual fields of size 256×256 at $10\times$ magnification (280×280 μm) are extracted from each case for multi-class tissue annotation by an expert pathologist. The pathologist then assigned a label to each 32×32 (35×35 μm) region, 64 per visual field, in all the visual fields from the pre-defined set of seven tissue types: Tumour, Lymphocyte/Inflammatory, Tumour-associated stroma, Keratin, Epithelium, Artifacts, and Other for remaining tissue regions. All the annotated visual fields are then split into training and validation sets where all visual fields from a case lie only in one set. Training and validation sets consist of 141,541 and 38,893 annotated regions, respectively. Table 5.4 presents the detailed distribution of annotated regions in each set.

5.3.4 Stain Invariance

Stain variation is the most common issue in histology datasets, especially when datasets are curated from multiple centres. I normalize [158] the visual fields in training set using multiple target images to make proposed coarse

Table 5.4: Distribution of annotated regions for each class in each training and validation sets.

Classes	Training (%)	Validation (%)	Total
Tumour	35,627 (73)	12,863 (27)	48,490
Lymphocytes	10,488 (65)	5736 (35)	16,224
Tumour-associated stroma	15,248 (88)	2161 (12)	17,409
Keratin	6735 (65)	3552 (35)	10,287
Epithelium	17,884 (80)	4354 (20)	22,238
Others	39,331 (84)	7450 (16)	46,781
Artifacts	16,228 (85)	2777 (15)	19,005
Total	141,541 (78)	38,893 (22)	180,434

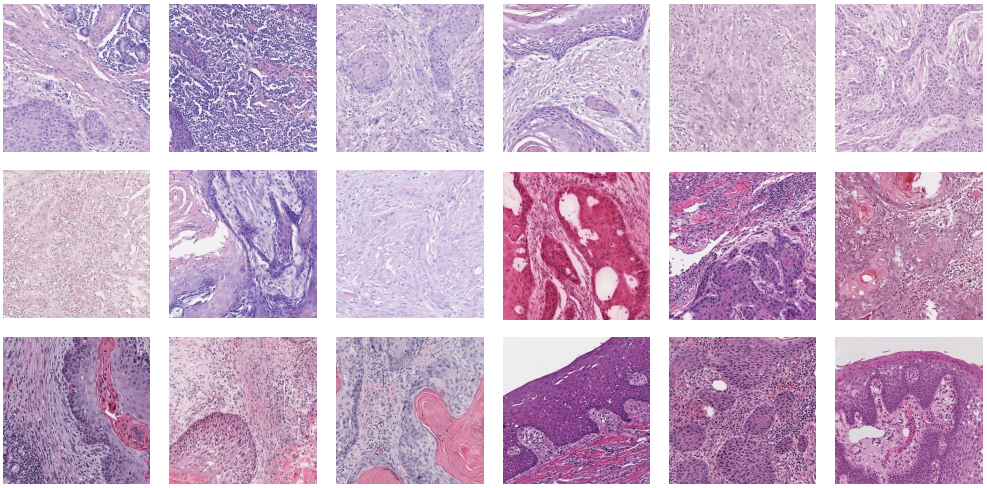


Figure 5.6: The images of target stains used for stain normalisation of the training cohort.

segmentation model invariant to stain variations (Figure 5.6). All visual fields from SKMCH&RC cohort are normalised using nine target images with diverse stains from TCGA-HN cohort, which results in a 10-time increase in the dataset, one original and nine normalised copies of each visual field. The same process is repeated for TCGA-HN visual fields using 9 SKMCH&RC visual fields as target images.

5.4 Results

I evaluate our proposed method for TME profiling using different evaluation metrics. The coarse segmentation method is evaluated in term of accuracy and time complexity, whereas prognostic significance measures are used to evaluate different spatial quantification methods. The detailed analysis of the performance of the proposed method is given in the following sections.

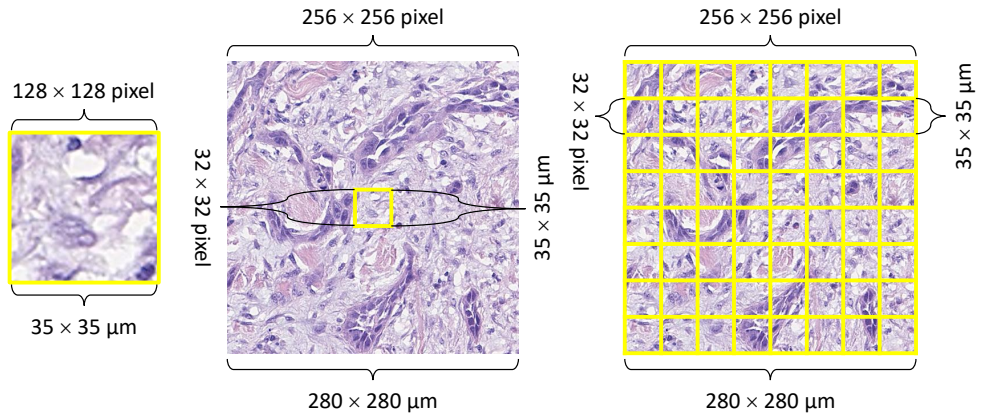


Figure 5.7: Input patches with yellow rectangles representing the regions corresponding to the predicted labels. Left, centre, and right patches are the input of standard patch classifiers, patch classifiers with context, and coarse segmentation network, respectively.

5.4.1 Coarse Segmentation Evaluation

I evaluate the performance of our proposed coarse segmentation method on 38K+ annotated regions from seven different tissue types, as explained in the dataset section. Average accuracy and average F1-score metrics are used to summarise the performance across seven tissue types, whereas box-plots are used to illustrate the variation in F1-score of different tissue types. The results of our proposed method are compared with three standard patch classifiers and their variants with a larger context.

Comparative Methods

Three standard patch classification methods (ResNet-50, MobileNet-1.0, and DenseNet-121) are used for comparison. These methods are trained on patches of size 128×128 pixels at $40\times$ magnification which is equivalent to $35 \times 35 \mu m$ tissue region used for coarse segmentation prediction in the proposed method. The input patches do not contain the same amount of context as the CSNet's input patches. Therefore, I train another set of models for these methods using the same CSNet input patches, 256×256 pixels at $10\times$ magnification, but the predicted label only represents the central 32×32 pixels ($35 \times 35 \mu m$). This training strategy enables us to make a fair comparison of the proposed method with patch classification methods as both types of methods are trained using the same amount of contextual information as shown in Figure 5.7. For the sake of clarity, I renamed these classifiers by adding 'Context' as post-fix to discriminate them (ResNet-50-Context, MobileNet-1.0-Context, and DenseNet-121-Context) from their standard version.

Table 5.5: Comparison of different variants of the proposed method with existing patch classifier methods in term of average accuracy and average F1-Score.

Methods	Accuracy	F1-Score
ResNet-50	0.6108	0.5797
MobileNet-1.0	0.6714	0.6373
DenseNet-121	0.6954	0.6610
ResNet-50-Context	0.7615	0.7323
MobileNet-1.0-Context	0.7662	0.7478
DenseNet-121-Context	0.8125	0.7876
CSNet-121	0.8165	0.7928
CSNet-121-SC	**0.8511**	**0.8311**
CSNet-61-SC	0.8205	0.8056

Comparative Analysis

The comparative results are presented in Table 5.5 and Figure 5.8. All variants of the proposed method outperformed the standard patch classifiers with and without context on both accuracy and F1-score metrics. The results of the proposed variants justify the need for different architectural modification in the baseline DenseNet-121 architecture. The CSNet-121 is the simplest variant and significantly similar to DenseNet-121; therefore, its performance is just above the DenseNet-121. However, the CSNet-121-CS achieves the highest performance due to the use of skip connection to link the features of intermediate layers to the final convolution layer. The CSNet-61-CS which consist of almost half of the parameters as compared to CSNet-121-CS and CSNet-121 but it outperformed the CSNet-121 variant just because of efficient architecture. Moreover, it shows the least variance in the F1-score of all tissue types.

In summary, the performance gain achieved by the proposed method is due to the use of broader context, more network parameters, and efficient network design. Similar performance patterns can be observed in the standard patch classifiers with and without context. The context-based patch classifiers perform significantly better than the one without context but with extra processing cost in term of time due to the requirement of overlapping patch-based prediction of whole slide images.

Time Comparison

WSIs are large images, and processing of these images may take from 1 minute to 1 hour. The processing time depends on several factors such as image magnification (e.g. 40×, 20×), type of problem, model complexity, inference pipeline, and available computational and memory resources. I compared the

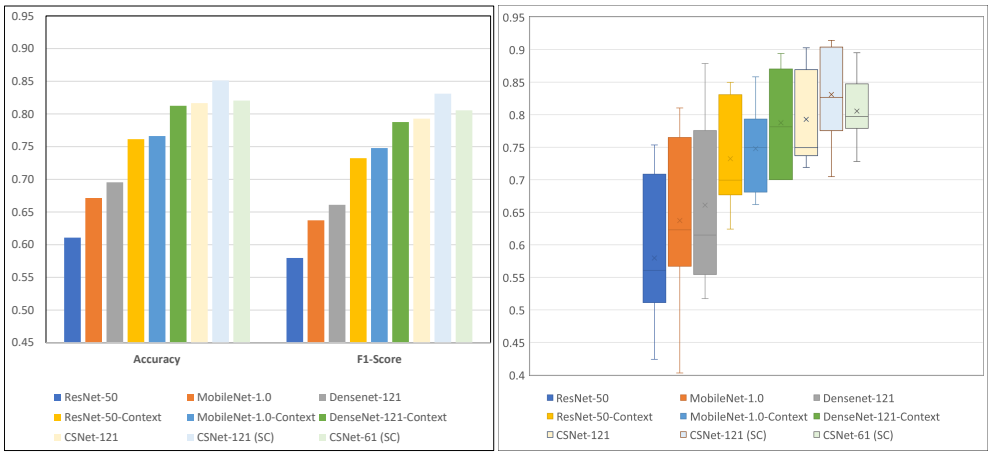


Figure 5.8: (Left) Bar-chart representing average accuracy and F1-score of 9 different methods. (Right) Boxplot based illustration of F1-score variation across different classes for each method.

Table 5.6: Comparison of time that different methods took to process a WSI at 40×. Sizes are given in pixels and time is reported in minutes.

	Size				Time		
Methods	**Batch**	**Patch**	**Stride**	**Prediction**	**Loading**	**Processing**	**Total**
DenseNet-121	512	128×128	128×128	128×128	10.33	10.61	20.94
DenseNet-121-Context	512	256×256	32×32	32×32	544.79	663.49	1208.28
DenseNet-121-Context	1	7168×7168	6944×6944	32×32	12.41	11.32	23.73
CSNet-121-SC	512	256×256	256×256	32×32	10.34	10.64	20.98
Deeplab-V3+	128	512×512	512×512	1×1	10.57	16.90	27.47

time taken by the proposed method and its counterparts to process a WSI with dimensions $76,608 \times 111,328$ at 40×. The proposed method and the standard Densenet-121 took a similar amount of time (21 minutes) whereas DenseNet-121-Context took significantly longer time (20+ hours) to process the WSI due to requirement of small stride size to produce a complete prediction map. However, efficient implementation of inference pipeline helps to reduce this stride overhead by using a large patch and stride size. I also reported the time a pixel-based segmentation method (Deeplab-V3+) took to process the WSI. Table 5.6 presents the time comparison conducted on same machine with 12GB TitanX GPU.

Visual Results

The visual results of the best performing coarse segmentation method (CS-121-SC) are presented in Figure 5.9 on SKMCH&RC visual fields. Each row shows the original and overlaid visual field in two separate columns. The visual field in the first row highlights the segmentation of tumour and tumour-associated stroma, where CS-121-SC has reliably segment the two tissue types. The tumour region in the upper left corner of the visual field in the second row has similarity with the border of the epithelium appeared in the visual field in

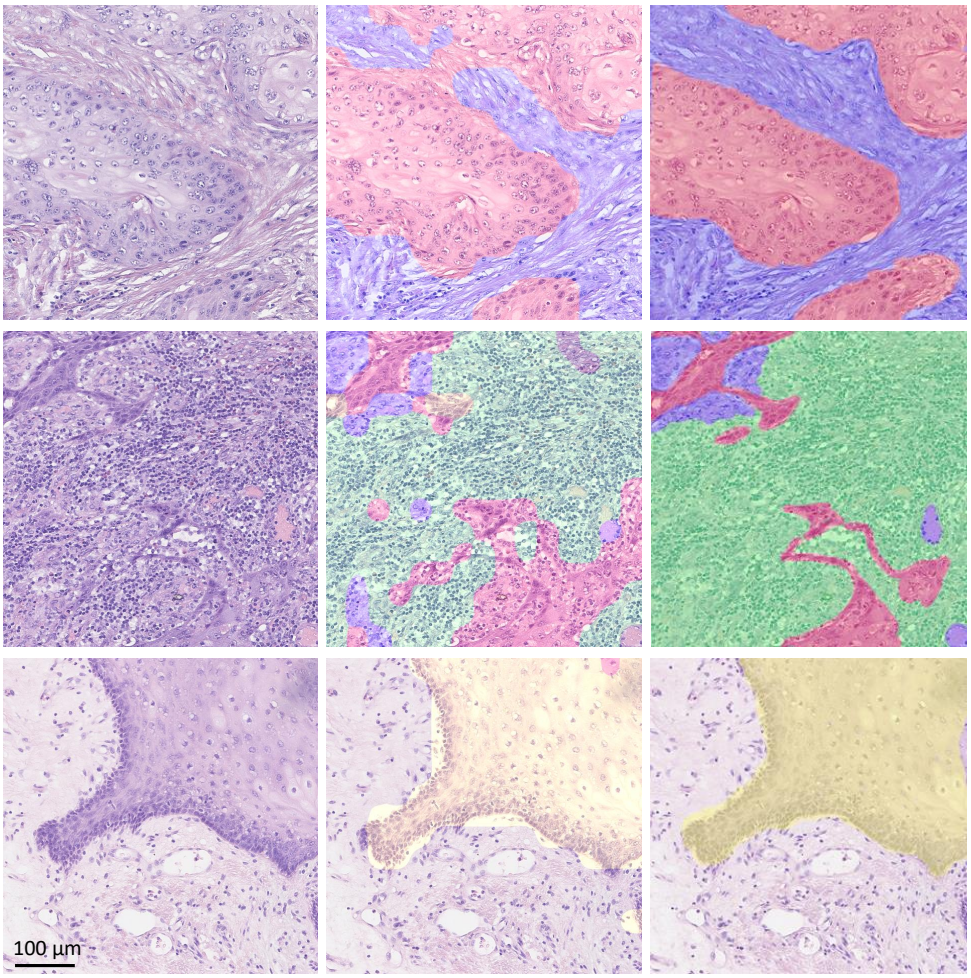


Figure 5.9: Visual results of the coarse segmentation method on three SKMCH&RC visual fields. The middle column shows the overlay of predicted tissue type and right column shows the ground truth tissue types in different colours where tumour, lymphocyte, tumour-associated stroma, and normal epithelium regions are represented by red, green, blue, and yellow colours. Non-overlaid regions belong to other tissue types.

the third row. The proposed methods segmented the majority of the tissue regions correctly. In general, the proposed method has performed well with some exception of small mispredictions which do not have a significant effect on the downstream analysis.

5.4.2 TME Profiling Analysis

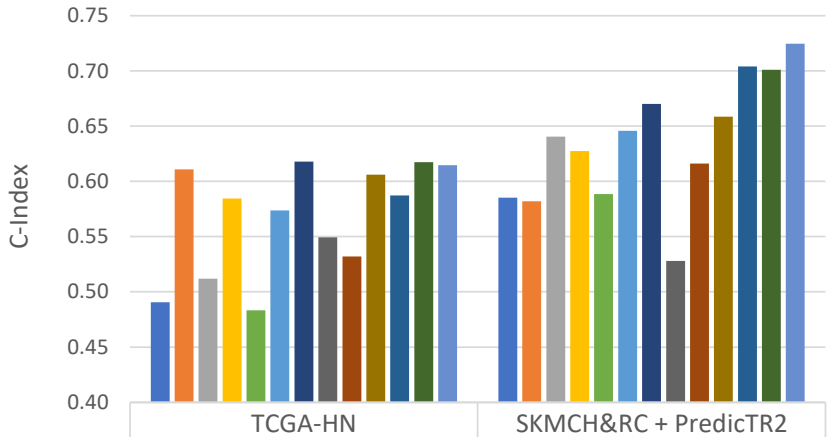
I analyse the prognostic significance of different spatial quantification methods based survival models using Log-rank test based p-value, concordance index. Kaplan Meier curves are used to illustrate the difference between low and high-risk patient groups in univariate analysis. Cox regression model is employed to investigate the potentially interacting clinical and pathological covariates. Disease-specific survival analysis is conducted using TCGA-HN as discovery and SKMCH&RC and PredicTR2 as a joint validation cohort (SKMCH&RC + PredicTR2). However, disease-free analysis is performed only on SKMCH&RC, and PredicTR2 cohorts separately as TCGA-HN does not has disease recurrence data.

Concordance Analysis

I use Harrell's concordance index (C-Index) [116] to evaluate the predictive ability of different quantification scores based survival models. Figure 5.10 presents the C-Index of all quantification measures on both TCGA-HN and SKMCH&RC + PredicTR2 cohorts when used as a validation cohort for disease-specific survival. C-Index results show that our proposed TIL-Ratio and TASIL-Ratio measures have better predictive ability for quantification of different spatial patterns. For the quantification of tumour-associated stroma and lymphocytes based spatial patterns, the proposed TASIL-Ratio achieved higher C-Index score as compared to LTAS-Ratio, LTAS-Col, and TASILAb based quantification methods. Similarly, for the quantification of tumour and lymphocytes based spatial patterns, the proposed TIL-Ratio achieved comparable C-Index score as compared to LT-Ratio, LT-Col, and TILAb based quantification methods.

Univariate Analysis

I further explore the prognostic significance of each quantification method independent to other clinical and pathological parameters. Patients in the validation cohorts are divided into two groups based on quantification scores using an optimal threshold value selected using the discovery cohort. Table 5.7 presents the hazard ratio (HR) with 95% confidence interval (CI) and Log-rank test based p-values of different quantification methods on both validation cohorts for disease-specific survival. Our proposed quantification methods



	TCGA-HN	SKMCH&RC + PredicTR2
T-Percentage	0.49	0.59
L-Percentage	0.61	0.58
TAS-Percentage	0.51	0.64
LT-Ratio	0.58	0.63
TAST-Ratio	0.48	0.59
LTAS-Ratio	0.57	0.65
LT-Col	0.62	0.67
LTAS-Col	0.55	0.53
TAST-Col	0.53	0.62
TILAb	0.61	0.66
TASILAb	0.59	0.70
TIL-Ratio	0.62	0.70
TASIL-Ratio	0.61	0.72

Validation Cohorts

Figure 5.10: C-Index based comparison of different quantification methods across validation cohorts for disease-specific survival.

Table 5.7: The hazard ratio with 95% confidence interval and Log-rank test based p-values of different quantification methods across different validation cohorts for disease-specific survival.

	TCGA-HN				SKMCH&RC + PredicTR2			
Methods	HR	CI-Lower	CI-Upper	p-value	HR	CI-Lower	CI-Upper	p-value
T-Percentage	0.96	0.60	1.50	0.863	2.70	0.97	7.70	**0.048**
TAS-Percentage	0.29	0.09	0.92	**0.025**	0.62	0.26	1.50	0.287
L-Percentage	0.70	0.31	1.60	0.407	1.50	0.36	6.20	0.587
LT-Ratio	0.63	0.41	0.97	**0.035**	0.41	0.21	0.80	**0.008**
LTAS-Ratio	0.76	0.50	1.20	0.209	0.32	0.10	1.00	**0.046**
TAST-Ratio	1.30	0.47	3.50	0.623	2.00	1.00	4.00	**0.043**
LT-Col	0.67	0.44	1.00	0.069	0.37	0.19	0.69	**0.001**
LTAS-Col	0.80	0.48	1.30	0.379	0.98	0.52	1.90	0.959
TAST-Col	1.10	0.72	1.70	0.640	2.30	0.88	5.80	0.082
TILAb	0.83	0.50	1.40	0.461	0.32	0.17	0.62	**0.0003**
TASILAb	0.62	0.37	1.00	0.073	0.42	0.22	0.79	**0.005**
TIL-Ratio	0.57	0.36	0.91	**0.018**	0.48	0.23	1.00	**0.049**
TASIL-Ratio	0.49	0.30	0.78	**0.002**	0.20	0.10	0.43	**0.000003**

(TIL-Ratio and TASIL-Ratio) remains prognostic on both validation cohorts. Patient group with higher TASIL-Ratio score shows significantly better disease-specific survival (p=0.00239, HR = 0.49, 95% CI 0.30–0.78) on TCGA-HN cohort. Similar pattern (p=0.000003, HR = 0.20, 95% CI 0.10–0.43) was observed on our joint cohort (SKMCH&RC + PredicTR2). The LT-Ratio is the only existing method which shows prognostic significance on both validation cohorts.

I use the Kaplan Meier curves to visualise the difference between the survival probability of low and high-risk patients for TL-Ratio, TIL-Ratio, and TASIL-Ratio based methods. Figure 5.11 presents the survival curve along with log-rank test based p-values for disease-specific survival of HNSCC patients from both cohorts. The TASIL-Ratio based Kaplan Meier curve shows a clear separation between low and high-risk patients on both cohorts as compared to TL-Ratio and TIL-Ratio based quantification methods.

SKMCH&RC and PredicTR2 cohorts are curated from the oral cavity and oropharynx, respectively. Therefore, I also investigate the prognostic significance of TASIL-Ratio for patients of a specific HNSCC site. First, oral (SKMCH&RC) cohort is considered as a discovery cohort for validation on oropharyngeal (PredicTR2) cohort. Similar to our previous finding, TASIL-Ratio remains prognostic (p=0.000159, HR = 0.20, 95% CI 0.08–0.49) for oropharyngeal squamous cell carcinoma patient stratification into low and high-risk groups for disease-specific survival. Second, I consider (SKMCH&RC) cohort as validation while using oropharyngeal (PredicTR2) cohort as the discovery cohort. I found that TASIL-Ratio based oral squamous cell carcinoma patient stratification again proved prognostic (p=0.000935, HR = 0.08, 95% CI 0.01–0.65). I repeated the same set of the experiment to evaluate the prognostic significance of TASIL-Ratio for disease-free survival. The results follow the same

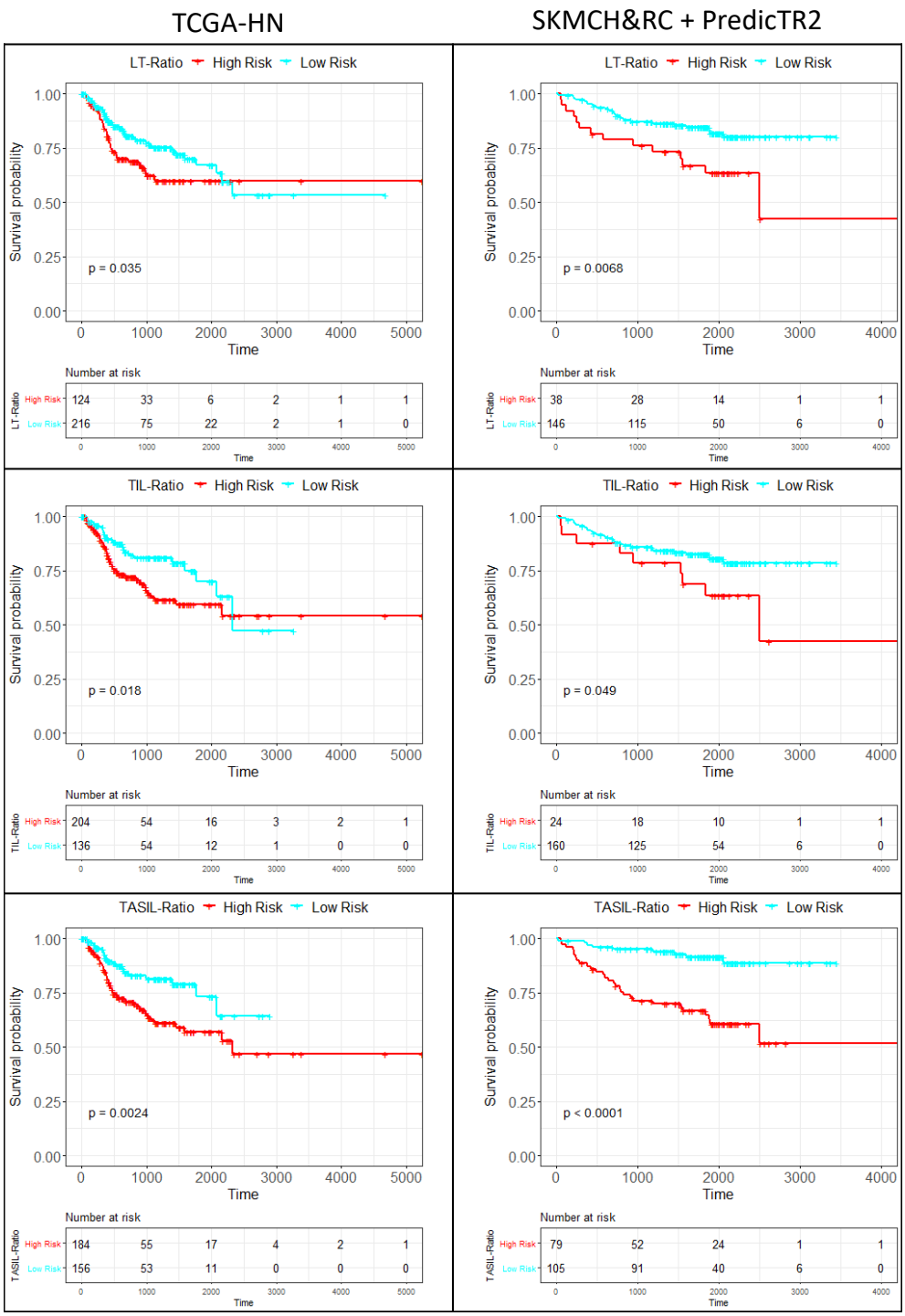


Figure 5.11: Kaplan Meier curves along with log-rank test based p-values for disease-specific survival of three automated scores.

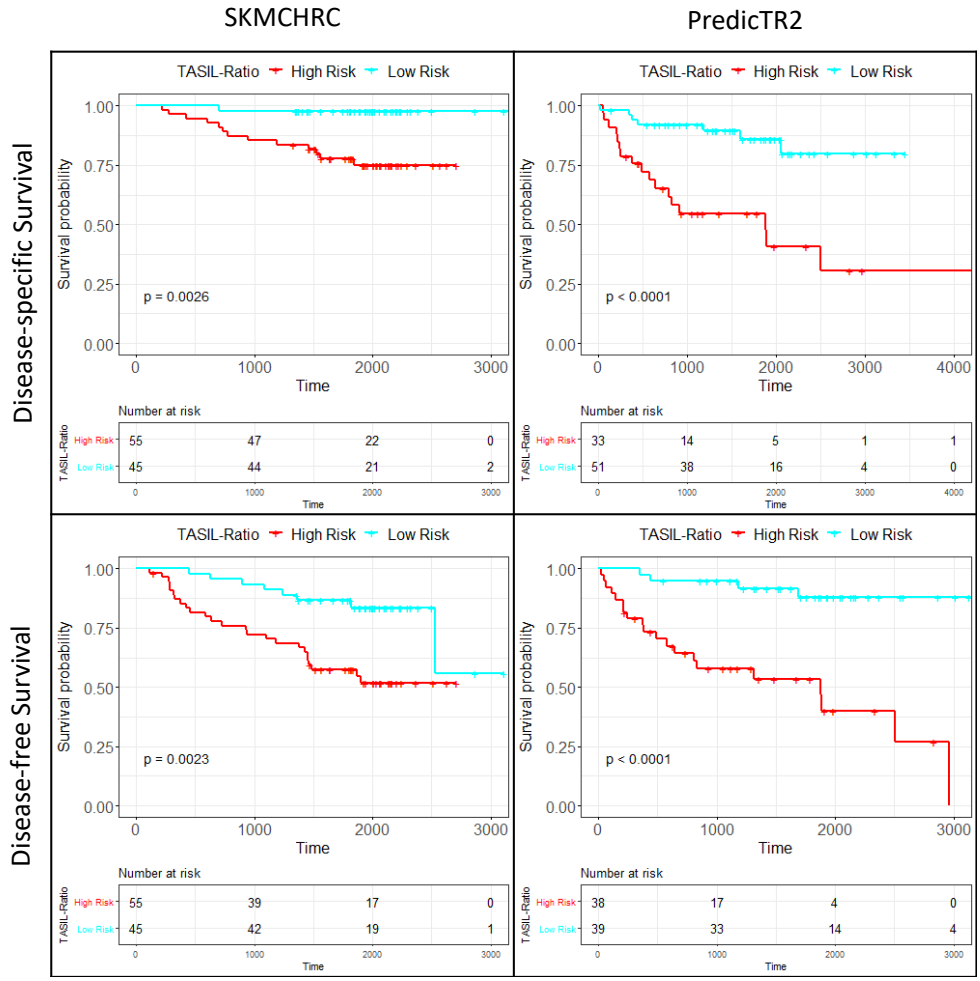


Figure 5.12: Kaplan Meier curves along with log-rank test based p-values for disease-specific and disease-free survival of using TASIL-Ratio based quantification method.

pattern where TASIL-Ratio stratifies the patients in prognostically significant low and high-risk groups. Patient stratification into low and high-risk groups is presented Figure 5.12 through Kaplan Meier curves for all four experiments.

Comparison with Pathologist Score

Pathologist score for tumour/stroma infiltrating lymphocytes is a categorical score with absent, low, moderate, and high infiltration categories. Only some categories show prognostic significance in SKMCH&RC and PredicTR2 cohorts for disease-specific and disease-free survival, as shown in Tables 5.2 and 5.3. Therefore, I group these categories into two categories, where one group consists of absent and low, and other contains moderate and high categories. I present the comparison of most prognostic quantification method (TASIL-Ratio) and pathologist manual TIL score in Figure 5.13. TASIL-Ratio shows the better separation between low and high-risk groups in 3 out of 4 experiments as

compared to pathologist score.

Multivariate Analysis

I investigate the prognostic significance of TASIL-Ratio in the presence of clinical and pathological variables whose information is available for TCGA-HN cohort for disease-specific survival. TASIL-Ratio remains prognostic (p=0.043, HR = 0.58, 95% CI 0.34–0.98) in presence of other clinicopathological variables: age, gender, grade and pathological stage. Although stage IVb and IVc show high prognostic values, the total number of patients in stage IVb and IVC are 9 and 1, respectively, which is quite small as compared to the total number of patients (Figure 5.14). I further evaluated the independence TASIL-Ratio in SKMT&CH cohort, which has more clinical and pathological parameters as compare to TCGA-HN, and both disease-specific and disease-free survival information. In disease-specific survival, I found a similar pattern as in TCGA-HN cohort. Both TASIL-Ratio (p=0.027, HR = 0.10, 95% CI 0.01–0.76) and pathological stage (p=0.043, HR = 2.02, 95% CI 1.02–3.97) turned-out as independent variables against all other variables (Figure 5.15). However, in disease-free survival, TASIL-Ratio is the only independent variable (p=0.004, HR = 0.29, 95% CI 0.12–0.67) against age, gender, smoke and smokeless tobacco status, grade, patterns of invasion, and pathological stage.

Correlation with Molecular Estimates of Immune Subtypes

I further investigate the correlation of proposed TASIL-Ratio with molecular estimates of immune cell fractions in TCGA-HN cohort. Throsson *et al.* [159] have estimated the fraction of 22 immune cell types in the histology sample of each patient in the TCGA cohort using CIBERSORT. I used those estimates for the correlation analysis with our TASIL-Ratio. The immune subtypes were grouped based on nine different immune cell types: dendritic, mast, neutrophils, eosinophils, monocytes, macrophages, natural killer cells, T cells and B cells. The TASIL-Ratio shows a moderate but highly significant positive correlation with T cells estimates and negative correlation with macrophages estimates (Figure 5.16). This correlation pattern indirectly indicates the correctness of lymphocyte segmentation by our coarse tissue segmentation method as lymphocytes largely comprise of T cells and B cells. CD8 T cell fraction shows the highest positive correlation among all immune subtypes (Table 5.8), which may indicate that the lymphocytes in the vicinity of the tumour-associated stroma are mainly CD8 T cells. A very high correlation between TASIL-Ratio is not expected as TASIL-Ratio and molecular estimates are computed from formalin-fixed paraffin-embedded and fresh frozen tissue sections, respectively. Although both tissues sections belong to the tissue block of the same patient,

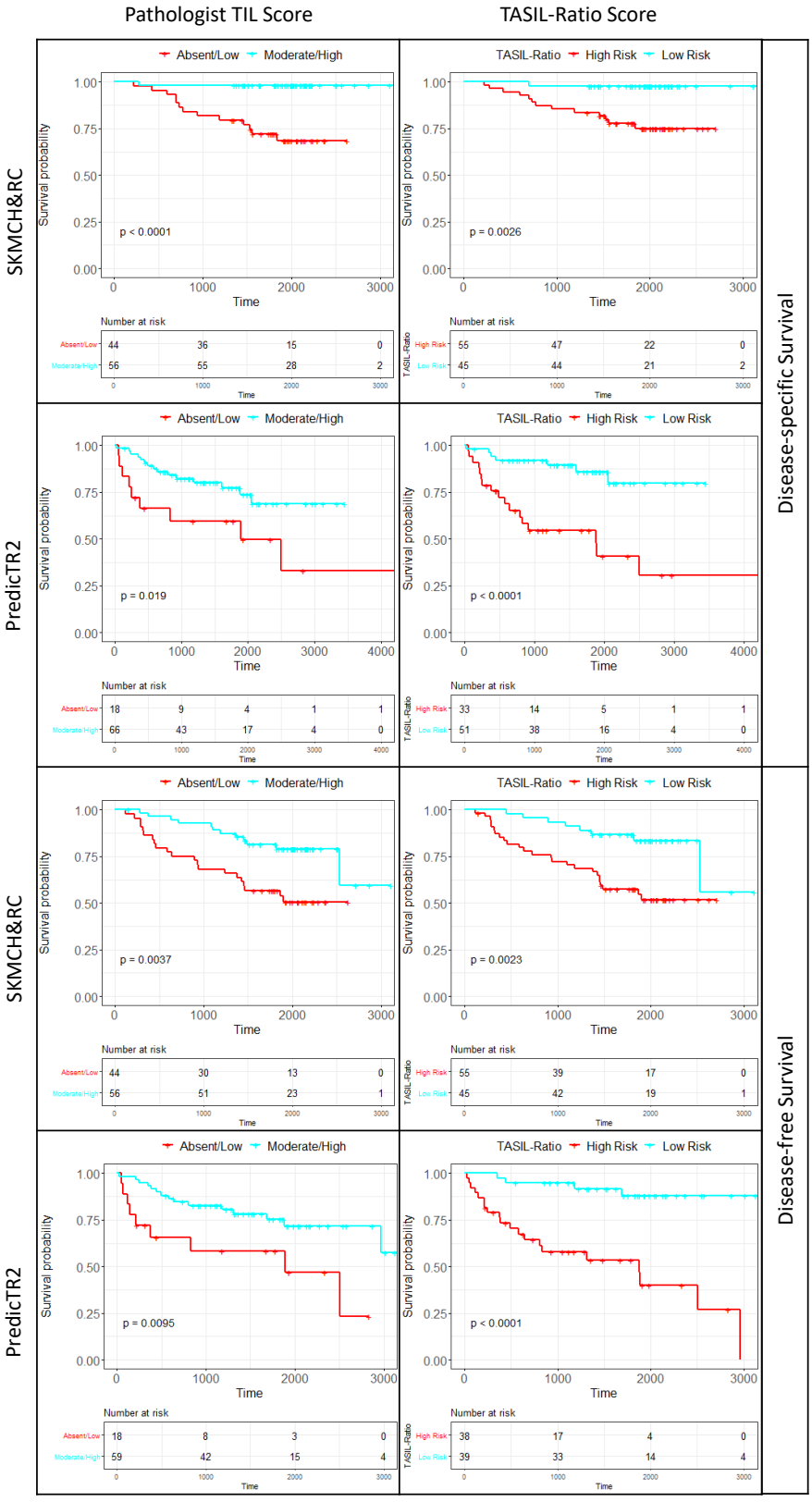


Figure 5.13: Comparison of manual pathologist TIL score and proposed TASIL-Ratio.

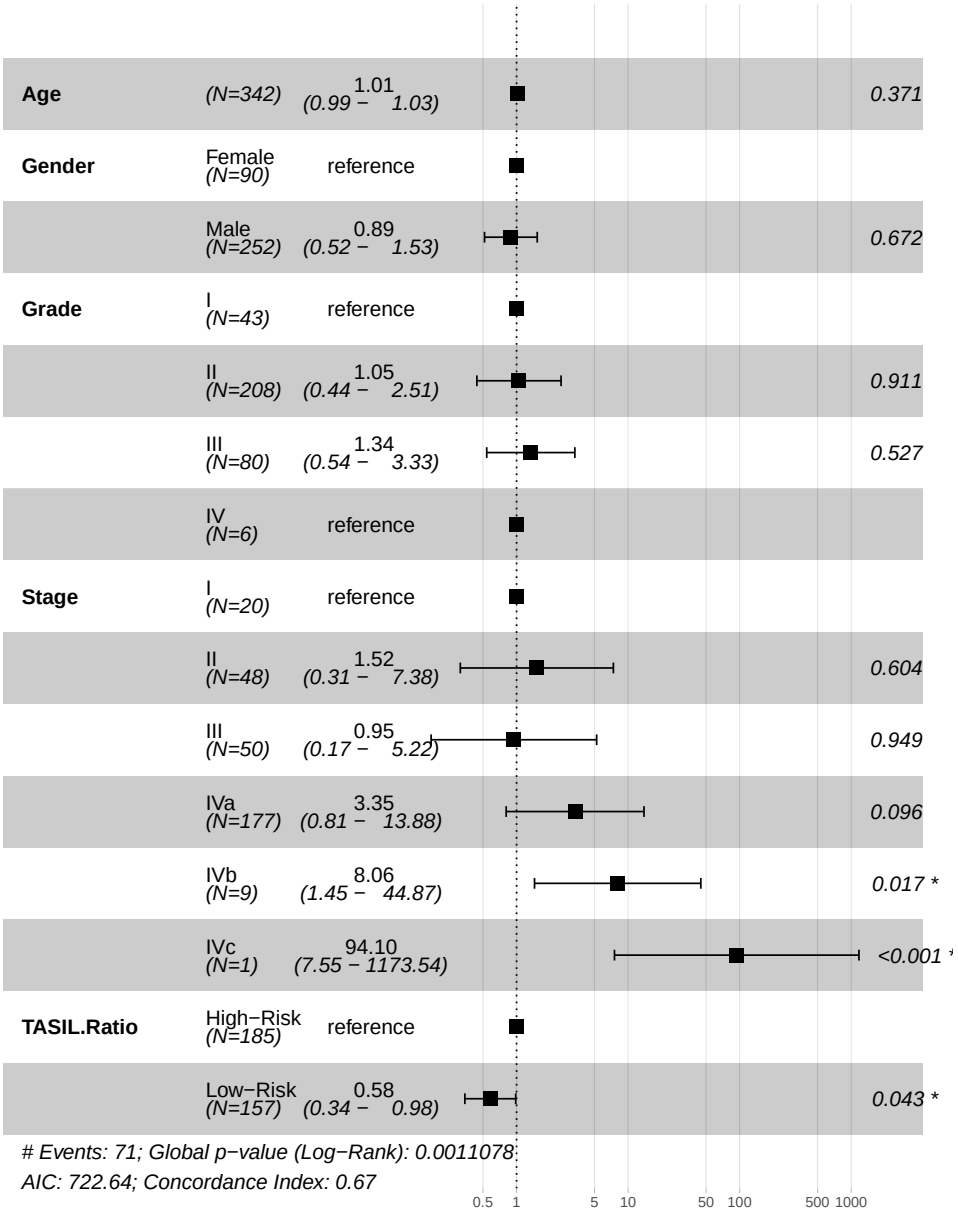


Figure 5.14: Multivariate analysis of TASIL-Ratio in the presence of available clinical and pathological variables of TCGA-HN cohort.

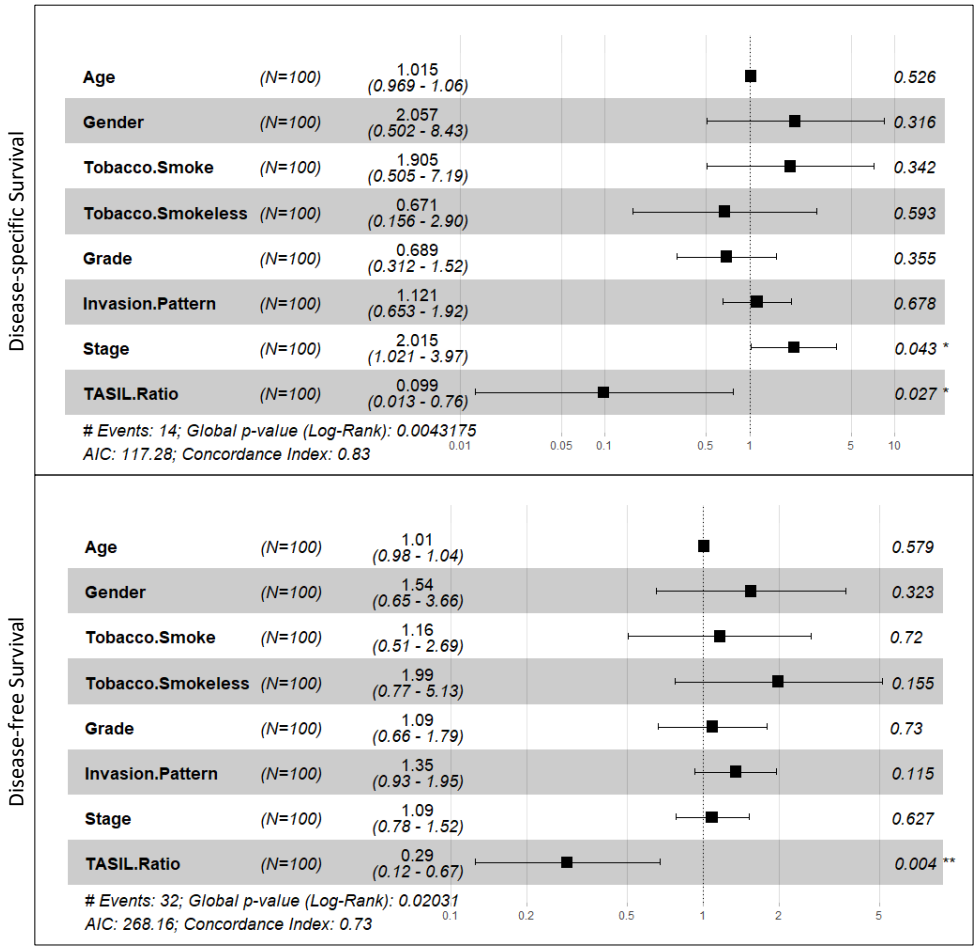


Figure 5.15: Multivariate analysis of TASIL-Ratio in the presence of available clinical and pathological variables of SKMCH&RC cohort.

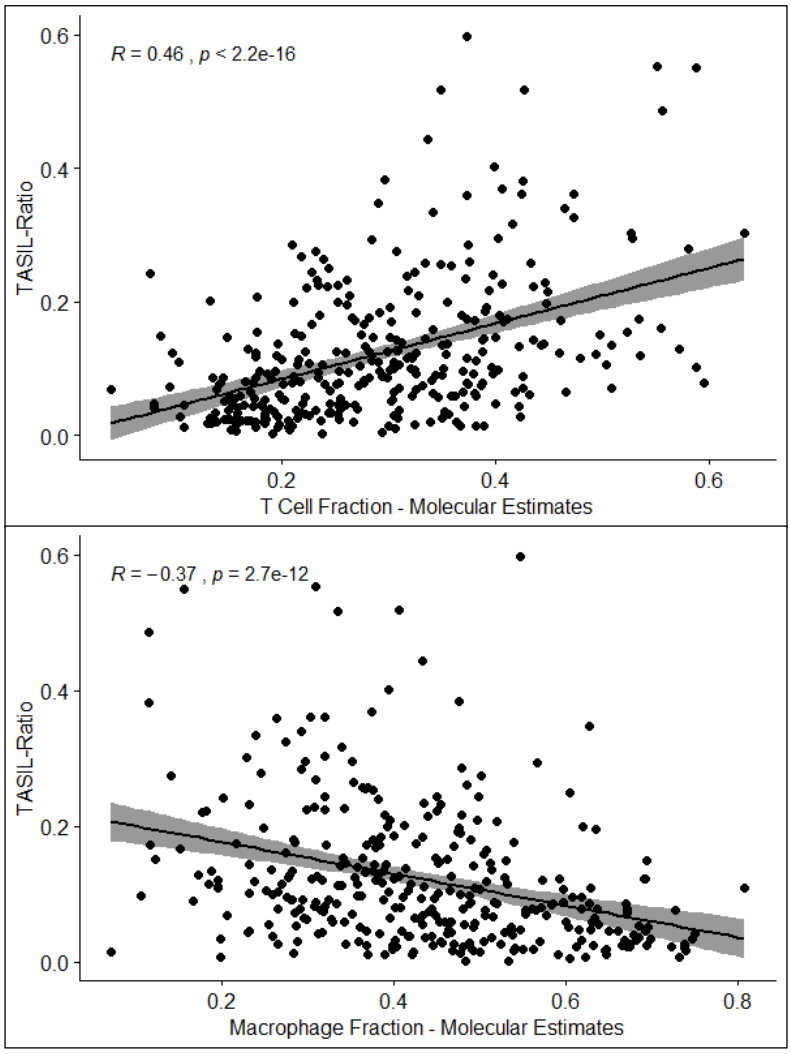


Figure 5.16: Spearman correlation between TASIL-Ratio and molecular estimates of macrophages and T Cell fractions.

their exact spatial relation is unknown.

5.5 Discussion

In this chapter, I profile the TME of HNSCC in the context of the tumour, stroma, lymphocytes. The role of stroma and lymphocytes in TME has been explored in many clinical studies [160–162] on HNSCC. The consensus is that the tumour-associated stroma helps in tumour development whereas tumour infiltrating lymphocytes act against the tumour. The standardised quantification of stroma and lymphocytes in the tumour is still an open challenge in HNSCC, just like many other cancer types. However, some guidelines have been proposed for TILs to make the quantification process more objective and to reduce the inter- and intra-observer variability [163–165].

Table 5.8: Spearman correlation between TASIL-Ratio and molecular estimates of immune subtypes.

Cell Types	ρ	p-value	Cell Subtypes	ρ	p-value
Dendritic Cells	0.01	8.17×10^{-1}	Activated	0.01	8.34×10^{-1}
			Resting	0.01	8.83×10^{-1}
Mast Cells	-0.18	8.22×10^{-4}	Activated	-0.18	1.19×10^{-3}
			Resting	0.12	2.84×10^{-2}
Neutrophils	0.09	9.73×10^{-2}	Neutrophils	0.09	9.73×10^{-2}
Eosinophils	-0.10	6.52×10^{-2}	Eosinophils	-0.10	6.52×10^{-2}
Monocytes	0.18	7.70×10^{-4}	Monocytes	0.18	7.70×10^{-4}
Macrophages	-0.37	2.67×10^{-12}	M0	-0.39	2.08×10^{-13}
			M1	0.32	2.62×10^{-9}
			M2	-0.24	1.05×10^{-5}
Natural Killer Cells	0.05	3.52×10^{-1}	Activated	0.05	3.44×10^{-1}
			Resting	0.00	9.94×10^{-1}
T-Cells	**0.46**	1.32×10^{-18}	CD4 Memory Activated	0.32	2.04×10^{-9}
			CD4 Memory Resting	-0.09	9.66×10^{-2}
			CD4 Naive	-0.20	3.11×10^{-4}
			CD8	**0.45**	9.28×10^{-18}
			Follicular Helper	0.26	2.37×10^{-6}
			Gammadelta	0.07	1.94×10^{-1}
			Regulatory	0.26	1.16×10^{-6}
B-Cells	0.22	3.97×10^{-5}	Memory	-0.02	7.37×10^{-1}
			Naive	0.22	4.32×10^{-5}
			Plasma	0.11	4.77×10^{-2}

Despite these efforts, manual quantification remains as a subjective process which leads to a lack of reproducibility.

In recent years, researchers have developed several automated quantification methods for TME analysis [60, 97, 150, 152]. The use of automated methods eliminate the issue of subjectivity and outputs objective and reproducible quantification scores. Most of the methods either rely on nucleus/cell detection and classification [60, 150] or used patch-based segmentation methods for the segmentation of different TME components [97, 152]. The nucleus/cell detection and classification work on higher image magnifications (40× or 20×) which requires more time for WSI processing, usually around an hour. Moreover, training for these methods requires a large number of annotated cells of different types which is a tedious and error-prone task. On the other hand, patch-based segmentation methods lose the precision in the segmentation of different TME components due to the large size of the patch. The use of small patch size results in the lack of contextual information which is essential for correct segmentation of some TME components.

Our proposed framework for automated quantification of TME is based on a novel coarse segmentation method which is more precise and accurate than the standard patch-based segmentation methods. The proposed method has achieved higher precision with 64× dense predictions as compare to standard patch classifiers whereas higher accuracy is achieved by the use of broader

spatial context and addition of skip connections in the network architecture. In Table 5.5, it can be seen that even standard patch classifiers can achiever up to 10% of performance improvements by using input patches with larger spatial context. However, classification with larger context requires more WSI processing time due to the small stride size as compare to patch size. Our proposed method predicts a label for each region of the input patch; therefore, it does not process a WSI using an overlapping sliding window based approach which makes it faster.

Our proposed coarse segmentation method segment a WSI into seven tissue types/classes. Although I were only interested in the tumour, tumour-associated stroma, and lymphocyte classes, the remaining four classes (normal epithelium, keratin, others and artefacts) were also important. The normal epithelium is very similar to the malignant epithelium (tumour); therefore, it could easily be misclassified as tumour if I put it in the other class which already consists of many different tissue regions. Similarly, artefacts class contains tissue regions from different classes with some blurring, tissue folding, and staining artefacts. However, keratin as separate class was considered to explore its prognostic significance in survival analysis, but it did not show any significance for any type of quantification method; therefore, I dropped it from the downstream analysis.

In the literature, several methods have been developed for quantification of different pairs of TME components such as stroma to tumour ratio in breast and ovarian cancer [95, 166], lymphocyte and tumour colocalisation in breast cancer [105], and abundance of tumour infiltrating lymphocyte in oral cancer [152]. I explored the significance of 13 differ spatial patterns of the tumour, tumour-associated stroma, and lymphocytes, along with the existing quantification methods in HNSSC. The percentage (T-Percentage, L-Percentage, and TAS-Percentage) and ratio (LT-Ratio, TAST-Ratio, LTAS-Ratio) based quantification scores are shallow scores as they just calculate the overall percentage or ratio and ignore the spatial co-occurrences of the tumour, tumour-associated stroma, and lymphocytes. The colocalisation base scores (LT-Col, LTAS-Col, and TAST-Col) contain more information about the spatial patterns of the pair of tissue types. Although high colocalisation score represents a relatively equal ratio of two tissue types, the low colocalisation score does give any information about the actual ratio between the pair of tissue types. It just represents that one tissue type is more prevalent as compared to another tissue type. However, the lymphocyte abundance scores (TILAb and TASILAb) capture both colocalisation and ratio information simultaneously. A higher lymphocyte abundance score represents high infiltration of lymphocytes in tumour or stroma regions, whereas lower score represents low lymphocytes infiltration. Both colocalisation and abundance-based quantification measures

calculate the colocalisation using the percentage of given tissue types in small pre-defined regions, which results in some loss of spatial patterns. However, our proposed quantification measures (TIL-Ratio, and TASIL-Ratio) captures the spatial patterns from the lowest level of coarse segmentation maps of WSIs, as illustrated in Figure 5.5.

In survival analysis, I found that higher infiltration of lymphocytes in tumour-associated stroma is associated with better disease-free survival of HNSCC patients. Our proposed automated TASIL-Ratio measure quantifies the extent of tumour-associated stroma infiltrating lymphocytes. The TASIL-Ratio is independent of HNSCC site and shows prognostic significance in both oral and oropharyngeal cohorts. Furthermore, the TASIL-Ratio is independent of clinical and pathological parameters, including grade, and stage. I compared TASIL-Ratio with the current spatial quantification methods for tumour, tumour-associated stroma and lymphocyte quantification. The TASIL-Ratio achieve high concordance score as compared to its counterparts. The TASIL-Ratio also shows a moderate but highly significant correlation with molecular estimates of 22 immune subtypes. In general, our quantification score based findings are aligned with the clinical knowledge with the added advantage of objectivity and reproducibility. Although I validated our method on relatively large cohorts, a comprehensive evaluation on a sizeable multicentric cohort is required before adopting the proposed digital biomarkers in clinical practice.

Chapter 6

Conclusions and Future Directions

In this thesis, I proposed a set of computational methods for spatial context based automated analysis of haemotoxylin and eosin (H&E) digitised histology images. First, a context-aware convolutional neural network (CNN) based method was proposed to capture the spatial architecture of the colorectal adenocarcinoma (CRA) glands for better CRA grading. Second, a statistical measure was formulated for spatial quantification of lymphocyte abundance in the vicinity of tumour in oral squamous cell carcinoma (OSCC). Third, a coarse segmentation method was developed for precise tissue segmentation which was then used for the profiling of tumour microenvironment (TME) of head and neck squamous cell carcinoma (HNSCC).

I evaluated thoroughly the performance of the proposed methods on both internal and public cohorts and also compared the results with existing approaches for the respective tasks. Context-aware CNN based CRA grading methods outperformed all its counterparts, whereas spatial quantification method for tumour infiltrating lymphocytes (TILs) have shown prognostic significance for disease-free survival of OSCC patients. The coarse segmentation of digitised tissue into multiple tissue types enabled us to simultaneously quantify different spatial patterns of the tumour, tumour-associated stroma, and lymphocytes in HNSCC cohorts. I have shown that our novel quantification of tumour-associated stroma infiltrating lymphocytes is a statistically significant prognostic indicator for disease-specific survival of HNSCC patients as compared to existing quantification methods for TILs and tumour stroma ratio.

I present a summary of each proposed method, along with a set of future directions in the following sections.

6.1 Context-aware Convolutional Neural Network

I have presented a novel context-aware CNN, which can incorporate 64 times larger context than standard CNN based patch classifiers. The proposed network is well-suited for the CRA grading task as glandular structures in CRA vary in size and shape which does not fit in an input patch of standard patch classifiers. The proposed context-aware CNN is comprised of two stacked CNNs. The first local representation CNN is used for learning the local representation of the patches in a histology image. Second, representation aggregation CNN (RA-CNN) predicts the CRA grade by aggregating local representation of the patches and their spatial context. The proposed method has been evaluated on two CRA grading datasets. A comprehensive analysis of different variations of the proposed method and comparison with existing approaches has been presented. The qualitative and quantitative results have demonstrated that our method has outperformed the patch-based classification methods, domain-specific CRA grading techniques, and existing context-based methods. The proposed approach is also suitable for other tasks which require broader contextual information, such as Gleason grading in prostate cancer and tumour growth pattern classification in lung cancer.

The RA-CNN in the proposed context-aware CNN method incorporates spatial contextual information through convolutional layers. One potential future direction could be to develop a representation aggregation network using recurrent neural network (RNN) instead of convolutional neural networks. However, RNNs are generally prone to overfitting (especially vanilla RNNs). Two-dimensional long short-term memory based RNNs have potential to capture bi-directional context, therefore, suitable for context-aware learning in histology images [167].

The idea of context-aware learning through CNN is generic, and I only explored it in term of spatial context. However, contextual information from other image modalities (e.g. immunohistochemistry) and sources (e.g. clinical or genomics data) can be incorporated in the CNN networks. Recently, some works [168, 169] have been done in this direction where authors tried to fuse data from multiple sources to build deep learning based cancer prognosis model.

6.2 Coarse Segmentation of Histology Images

I have proposed a novel coarse segmentation network (CSNet) to eliminate the issues of noisy patch-based segmentation of histology images due to patch size and limited context. The CSNet leverages the spatial context of the input image to segment each 32×32 region of the input image. The segmentation map of CSNet is $64\times$ denser than standard patch-based image segmentation methods.

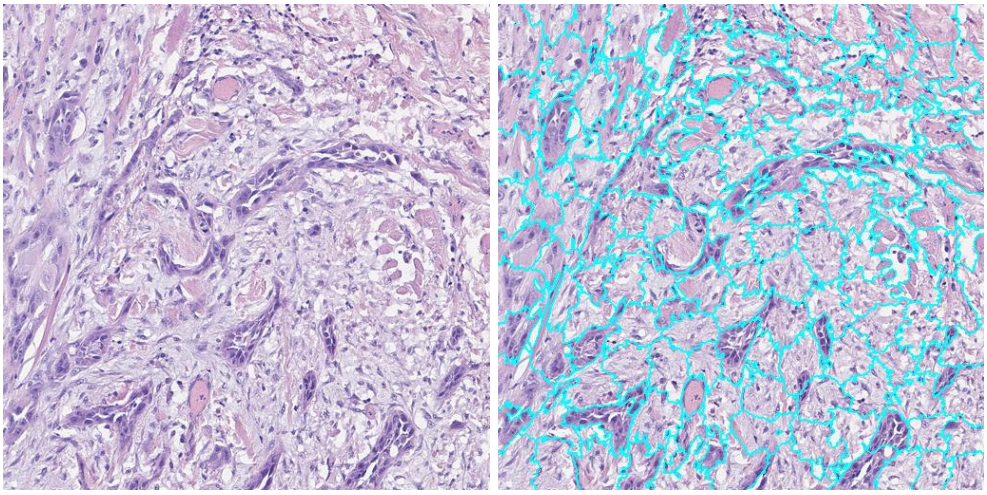


Figure 6.1: Illustration of super-pixel based segmentation of histology image. Each cyan colored region in the right image represent one super pixel.

The CSNet is a fully convolutional neural network with skip connections from intermediate layers to the final layer for better spatial segmentation of a given input image. Unlike pixel-based segmentation methods, our proposed method does not require pixel-level ground truth and high computation requirements of network training. The use of weighted sparse loss function has enabled CSNet to learn from partially annotated images which further ease the process of ground truth marking in large datasets. The proposed network has achieved superior performance for HNSCC tissue segmentation task when compared with its counterparts. The CSNet is also faster in time as compared to pixel-based segmentation methods. The proposed method is suitable for histology image segmentation problems where segmentation of the small region of tissue requires larger contextual information such as segmentation of normal and malignant epithelium, and normal and tumour associated stroma.

The current architecture of the proposed method only supports 64 times denser prediction map as compared to a standard patch-based classifier. The architecture of the proposed method can be further improved by the use of spatial pyramid pooling layers [170] to adjust (increase or decrease) the density of the prediction map. However, a larger/denser prediction map also requires more precise ground truth for the training of the proposed method. Another future direction could be the use of super-pixel based unsupervised segmentation method (6.1) as postprocessing step where each super-pixel will get the label from the corresponding prediction in the prediction map generated by the proposed method. This postprocessing approach will help to increase the precision in final tissue segmentation without any extra annotation.

6.3 Profiling of Tumour Microenvironment

I have profiled the tumour microenvironment by quantifying the spatial patterns of the tumour, tumour-associated stroma, and lymphocytes using novel quantification methods. In the fourth chapter, I have proposed tumour infiltrating lymphocytes abundance (TILAb) score which is the product of spatial colocalisation of tumour and lymphocytes, and lymphocyte to tumour ratio. I have shown that our proposed TILAb score is a prognostic indicator for disease-free survival of OSCC patients. The TILAb score has also shown independence from tumour invasion pattern, grade, and stage in multivariate analysis.

In the fifth chapter, I have developed another automated score for tumour-associated stroma infiltrating lymphocytes (TASIL-Ratio) which is the ratio between tumour associated stroma colocalised with lymphocytes and overall tumour associated stroma. The TASIL-Ratio has shown better predictive ability when compared with other existing automated quantification methods for HNSCC. The TASIL score is also a prognostic indicator for disease-specific survival of HNSCC patients. It has shown a moderate but highly significant positive correlation with molecular estimates of CD8 T cells. I have also demonstrated the prognostic significance of TASIL score for disease-free and disease-specific survival of OSCC and pharyngeal squamous cell carcinoma patients.

Both TILAb and TASIL-Ratio scores are based on a two-step approach where histology images are first segmented into clinically significant tissue types. Then spatial patterns are quantified using statistics such as TILAb and TASIL-Ratio scores. However, one future direction could be to predict risk score directly from the histology image through deep learning based methods using survival information as ground truth information. Furthermore, attention-based strategies could be used to highlight image regions associated with a high or low-risk score to improve the interpretability of the deep learning methods. One potential challenge to train such networks is the limited availability of large patient cohort, with reliable survival data.

6.4 Concluding Remarks

The proposed methods have shown generalisability with promising results on reasonably large datasets. The most common factors that impact the generalisability of a method across multiple cohorts are related to tissue preparation (e.g. tissue shrinkage, fixation artefacts, staining artefacts) and digitisation of tissue slides (e.g. out of focus scanning). I have used different strategies (stain normalisation and stain invariant training) to overcome the

staining variabilities in the available datasets. Similarly, most of the artefacts are detected and excluded by the use of a separate class. Further evaluation on large cohorts of patients with diverse demographics information may increase the generalisability of the proposed methods, especially in the context of personalised healthcare. However, a rigorous independent evaluation of the proposed methods on multiple cohorts is required before the adoption of these methods in clinical practice.

Bibliography

[1] R. Pazdur, L. R. Coia, W. J. Hoskins, and L. D. Wagman, *Cancer Management: A Multidisciplinary Approach; Medical, Surgical, & Radiation Oncology*. FA Davis Company, 2004.

[2] K. D. Alsibai and D. Meseure, "Significance of tumor microenvironment scoring and immune biomarkers in patient stratification and cancer outcomes," *Histopathology: An Update*, vol. 11, 2018.

[3] M. Hayat, *Methods of cancer diagnosis, therapy, and prognosis: liver cancer*, vol. 5. Springer Science & Business Media, 2009.

[4] 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.

[5] M. Dollinger and E. H. Rosenbaum, *Everyone's Guide to Cancer Therapy;: How Cancer Is Diagnosed, Treated, and Managed Day to Day*. Andrews McMeel Publishing, 2002.

[6] P. Trott, "International classification of diseases for oncology," *Journal of clinical pathology*, vol. 30, no. 8, p. 782, 1977.

[7] A. Jemal, R. Siegel, E. Ward, T. Murray, J. Xu, and M. J. Thun, "Cancer statistics, 2007," *CA: a cancer journal for clinicians*, vol. 57, no. 1, pp. 43–66, 2007.

[8] "Bowel cancer facts: About bowel cancer." url=https://www.bowelcanceruk.org.uk/about-bowel-cancer/bowel-cancer/. Accessed: 2019-03-10.

[9] M. Fleming, S. Ravula, S. F. Tatishchev, and H. L. Wang, "Colorectal carcinoma: Pathologic aspects," *Journal of gastrointestinal oncology*, vol. 3, no. 3, p. 153, 2012.

[10] "Bowel cancer survival statistics." url=https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer/survival. Accessed: 2019-03-10.

[11] M. B. Amin, F. L. Greene, S. B. Edge, C. C. Compton, J. E. Gershenwald, R. K. Brookland, L. Meyer, D. M. Gress, D. R. Byrd, and D. P. Winchester, "The eighth edition ajcc cancer staging manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging," *CA: a cancer journal for clinicians*, vol. 67, no. 2, pp. 93–99, 2017.

[12] C. C. Compton, L. P. Fielding, L. J. Burgart, B. Conley, H. S. Cooper, S. R. Hamilton, M. E. H. Hammond, D. E. Henson, R. V. Hutter, R. B. Nagle, *et al.*, "Prognostic factors in colorectal cancer: College of american pathologists consensus statement 1999," *Archives of pathology & laboratory medicine*, vol. 124, no. 7, pp. 979–994, 2000.

[13] C. C. Compton, "Updated protocol for the examination of specimens from patients with carcinomas of the colon and rectum, excluding carcinoid tumors, lymphomas, sarcomas, and tumors of the vermiform appendix: a basis for checklists," *Archives of pathology & laboratory medicine*, vol. 124, no. 7, pp. 1016–1025, 2000.

[14] N. Bannister and J. Broggio, "Cancer survival by stage at diagnosis for england (experimental statistics): Adults diagnosed 2012, 2013, 2014 and followe d up to 2015," *Produced in collaboration with Public Health England*, 2016.

[15] W. Blenkinsopp, S. Stewart-Brown, L. Blesovsky, G. Kearney, and L. Fielding, "Histopathology reporting in large bowel cancer.," *Journal of clinical pathology*, vol. 34, no. 5, pp. 509–513, 1981.

[16] "Head and neck cancers incidence statistics - cancer research uk." `https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/head-and-neck-cancers#heading-Zero`. Accessed 18 March 2020.

[17] K. D. Shield, J. Ferlay, A. Jemal, R. Sankaranarayanan, A. K. Chaturvedi, F. Bray, and I. Soerjomataram, "The global incidence of lip, oral cavity, and pharyngeal cancers by subsite in 2012," *CA: a cancer journal for clinicians*, vol. 67, no. 1, pp. 51–64, 2017.

[18] "Head and neck cancers incidence statistics - cancer research uk." `https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/`

head-and-neck-cancers/incidence#heading-Four. Accessed: 2019-03-10.

[19] A. Krishna, R. Singh, S. Singh, P. Verma, U. Pal, and S. Tiwari, "Demographic risk factors, affected anatomical sites and clinicopathological profile for oral squamous cell carcinoma in a north Indian population," *Asian Pacific Journal of Cancer Prevention*, vol. 15, no. 16, pp. 6755–6760, 2014.

[20] L. Liu, S. K. Kumar, P. P. Sedghizadeh, A. N. Jayakar, and C. F. Shuler, "Oral squamous cell carcinoma incidence by subsite among diverse racial and ethnic populations in California," *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology and Endodontics*, vol. 105, no. 4, pp. 470–480, 2008.

[21] M. L. C. Oliveira, V. P. Wagner, M. Sant'Ana Filho, V. C. Carrard, F. N. Hugo, and M. D. Martins, "A 10-year analysis of the oral squamous cell carcinoma profile in patients from public health centers in Uruguay," *Brazilian oral research*, vol. 29, no. 1, pp. 1–8, 2015.

[22] B. J. Braakhuis, R. H. Brakenhoff, and C. R. Leemans, "Second field tumors: a new opportunity for cancer prevention?," *The Oncologist*, vol. 10, no. 7, pp. 493–500, 2005.

[23] A. Mohanta and P. Mohanty, "Pattern of keratin expression and its impact on nuclear-cytoplasmic ratio in plump keratinized squamous cells during oral carcinogenesis," *J Med Diagn Meth*, vol. 5, no. 1, pp. 1–7, 2016.

[24] J. Pindborg, P. Reichart, C. Smith, and I. Van der Waal, "Definitions and explanatory notes. histological typing of cancer and precancer of the oral mucosa," 1997.

[25] J. D. Brierley, M. K. Gospodarowicz, and C. Wittekind, *TNM classification of malignant tumours*. John Wiley & Sons, 2016.

[26] C. Rivera, D. Droguett, U. Kemmerling, and B. Venegas, "Chronic restraint stress in oral squamous cell carcinoma," *Journal of dental research*, vol. 90, no. 6, pp. 799–803, 2011.

[27] H. Kurokawa, M. Zhang, S. Matsumoto, Y. Yamashita, T. Tomoyose, T. Tanaka, H. Fukuyama, and T. Takahashi, "The high prognostic value of the histologic grade at the deep invasive front of tongue squamous cell carcinoma," *Journal of oral pathology & medicine*, vol. 34, no. 6, pp. 329–333, 2005.

[28] M. Bryne, H. S. Koppang, R. Lilleng, and Å. Kjærheim, "Malignancy grading of the deep invasive margins of oral squamous cell carcinomas has high prognostic value," *The Journal of pathology*, vol. 166, no. 4, pp. 375–381, 1992.

[29] M. Peled, A. Onn, and R. S. Herbst, "Tumor-infiltrating lymphocytes—location for prognostic evaluation," *Clinical Cancer Research*, vol. 25, no. 5, pp. 1449–1451, 2019.

[30] C. Zhou, Y. Wu, L. Jiang, Z. Li, P. Diao, D. Wang, W. Zhang, L. Liu, Y. Wang, H. Jiang, *et al.*, "Density and location of cd 3+ and cd 8+ tumor-infiltrating lymphocytes correlate with prognosis of oral squamous cell carcinoma," *Journal of Oral Pathology & Medicine*, vol. 47, no. 4, pp. 359–367, 2018.

[31] M. Bryne, "Is the invasive front of an oral carcinoma the most important area for prognostication?," *Oral diseases*, vol. 4, no. 2, pp. 70–77, 1998.

[32] W. L. Dissanayaka, G. Pitiyage, P. V. R. Kumarasiri, R. L. P. R. Liyanage, K. D. Dias, and W. M. Tilakaratne, "Clinical and histopathologic parameters in survival of oral squamous cell carcinoma," *Oral surgery, oral medicine, oral pathology and oral radiology*, vol. 113, no. 4, pp. 518–525, 2012.

[33] A. K. Chaturvedi, E. A. Engels, R. M. Pfeiffer, B. Y. Hernandez, W. Xiao, E. Kim, B. Jiang, M. T. Goodman, M. Sibug-Saber, W. Cozen, *et al.*, "Human papillomavirus and rising oropharyngeal cancer incidence in the united states," *Journal of clinical oncology*, vol. 29, no. 32, p. 4294, 2011.

[34] E. J. Junor, G. R. Kerr, and D. H. Brewster, "Fastest increasing cancer in scotland, especially in men," *Bmj*, vol. 340, p. c2512, 2010.

[35] M. Ward, S. Thirdborough, T. Mellows, C. Riley, S. Harris, K. Suchak, A. Webb, C. Hampton, N. Patel, C. Randall, *et al.*, "Tumour-infiltrating lymphocytes predict for outcome in hpv-positive oropharyngeal cancer," *British journal of cancer*, vol. 110, no. 2, p. 489, 2014.

[36] C. Anderson, "Tumour-infiltrating lymphocytes influence prognosis in human papillomavirus-positive cervical and oropharyngeal cancer: A systematic review," *European Journal of Cancer*, vol. 110, p. S31, 2019.

[37] N. Farahani, A. V. Parwani, and L. Pantanowitz, "Whole slide imaging in pathology: advantages, limitations, and emerging perspectives," *Pathol Lab Med Int*, vol. 7, pp. 23–33, 2015.

[38] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE reviews in biomedical engineering*, vol. 2, pp. 147–171, 2009.

[39] D. Komura and S. Ishikawa, "Machine learning methods for histopathological image analysis," *Computational and structural biotechnology journal*, vol. 16, pp. 34–42, 2018.

[40] H. Mahmood, M. Shaban, B. Indave, A. Santos-Silva, N. Rajpoot, and S. Khurram, "Use of artificial intelligence in diagnosis of head and neck precancerous and cancerous lesions: A systematic review," *Oral Oncology*, vol. 110, p. 104885, 2020.

[41] C. L. Srinidhi, O. Ciga, and A. L. Martel, "Deep neural network models for computational histopathology: A survey," *arXiv preprint arXiv:1912.12378*, 2019.

[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

[45] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[46] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1915–1929, 2013.

[47] N. A. Koohababni, M. Jahanifar, A. Gooya, and N. Rajpoot, "Nuclei detection using mixture density networks," in *International Workshop on Machine Learning in Medical Imaging*, pp. 241–248, Springer, 2018.

[48] 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.

[49] T.-H. Song, V. Sanchez, H. ElDaly, and N. Rajpoot, "Simultaneous cell detection and classification in bone marrow histology images," *IEEE journal of biomedical and health informatics*, 2018.

[50] 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.

[51] N. A. Koohbanani, T. Qaisar, M. Shaban, J. Gamper, and N. Rajpoot, "Significance of hyperparameter optimization for metastasis detection in breast histology images," in *Computational Pathology and Ophthalmic Medical Image Analysis*, pp. 139–147, Springer, 2018.

[52] H. Lin, H. Chen, S. Graham, Q. Dou, N. Rajpoot, and P.-A. Heng, "Fast scannet: Fast and dense analysis of multi-gigapixel whole-slide images for cancer metastasis detection," *IEEE transactions on medical imaging*, 2019.

[53] 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.

[54] E. Arvaniti, K. S. Fricker, M. Moret, N. J. Rupp, T. Hermanns, C. Fankhauser, and et al., "Automated gleason grading of prostate cancer tissue microarrays via deep learning," *bioRxiv*, p. 280024, 2018.

[55] N. Ing, Z. Ma, J. Li, H. Salemi, C. Arnold, B. S. Knudsen, and et al., "Semantic segmentation for prostate cancer grading by convolutional neural networks," in *Medical Imaging 2018: Digital Pathology*, vol. 10581, p. 105811B, International Society for Optics and Photonics, 2018.

[56] R. Awan, K. Sirinukunwattana, D. Epstein, S. Jefferyes, U. Qidwai, Z. Aftab, and et al., "Glandular morphometrics for objective grading of colorectal adenocarcinoma histology images," *Scientific reports*, vol. 7, no. 1, p. 16852, 2017.

[57] R. Awan, N. A. Koohbanani, M. Shaban, A. Lisowska, and N. Rajpoot, "Context-aware learning using transferable features for classification of breast cancer histology images," in *International Conference Image Analysis and Recognition*, pp. 788–795, Springer, 2018.

[58] N. Alsubaie, K. Sirinukunwattana, S. E. A. Raza, D. Snead, and N. Rajpoot, "A bottom-up approach for tumour differentiation in whole slide

images of lung adenocarcinoma," in *Medical Imaging 2018: Digital Pathology*, vol. 10581, p. 105810E, International Society for Optics and Photonics, 2018.

[59] 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.

[60] S. Nawaz, A. Heindl, K. Koelble, and Y. Yuan, "Beyond immune density: critical role of spatial heterogeneity in Estrogen Receptor-negative breast cancer," *Modern Pathology*, vol. 28, no. 6, p. 766, 2015.

[61] G. Finak, N. Bertos, F. Pepin, S. Sadekova, M. Souleimanova, H. Zhao, H. Chen, G. Omeroglu, S. Meterissian, A. Omeroglu, *et al.*, "Stromal gene expression predicts clinical outcome in breast cancer," *Nature medicine*, vol. 14, no. 5, pp. 518–527, 2008.

[62] A. Calon, E. Lonardo, A. Berenguer-Llergo, E. Espinet, X. Hernando-Momblona, M. Iglesias, M. Sevillano, S. Palomo-Ponce, D. V. Tauriello, D. Byrom, *et al.*, "Stromal gene expression defines poor-prognosis subtypes in colorectal cancer," *Nature genetics*, vol. 47, no. 4, p. 320, 2015.

[63] C. Isella, A. Terrasi, S. E. Bellomo, C. Petti, G. Galatola, A. Muratore, A. Mellano, R. Senetta, A. Cassenti, C. Sonetto, *et al.*, "Stromal contribution to the colorectal cancer transcriptome," *Nature genetics*, vol. 47, no. 4, p. 312, 2015.

[64] S. S. Chennamsetty, M. Safwan, and V. Alex, "Classification of breast cancer histology image using ensemble of pre-trained neural networks," in *International Conference Image Analysis and Recognition*, pp. 804–811, Springer, 2018.

[65] M. Kohl, C. Walz, F. Ludwig, S. Braunewell, and M. Baust, "Assessment of breast cancer histology using densely connected convolutional networks," in *International Conference Image Analysis and Recognition*, pp. 903–913, Springer, 2018.

[66] I. Koné and L. Boulmane, "Hierarchical resnext models for breast cancer histology image classification," in *International Conference Image Analysis and Recognition*, pp. 796–803, Springer, 2018.

[67] Y. Liu, K. Gadepalli, M. Norouzi, G. E. Dahl, T. Kohlberger, A. Boyko, and et al., "Detecting cancer metastases on gigapixel pathology images," *arXiv preprint arXiv:1703.02442*, 2017.

[68] N. Alsubaie, M. Shaban, D. Snead, A. Khurram, and N. Rajpoot, "A multi-resolution deep learning framework for lung adenocarcinoma growth pattern classification," in *Annual Conference on Medical Image Understanding and Analysis*, pp. 3–11, Springer, 2018.

[69] A. BenTaieb, H. Li-Chang, D. Huntsman, and G. Hamarneh, "A structured latent model for ovarian carcinoma subtyping from histopathology slides," *Medical image analysis*, vol. 39, pp. 194–205, 2017.

[70] X. Wang, H. Chen, C. Gan, H. Lin, Q. Dou, Q. Huang, and et al., "Weakly supervised learning for whole slide lung cancer image classification," 2018.

[71] F. G. Zanjani, S. Zinger, *et al.*, "Cancer detection in histopathology whole-slide images using conditional random fields on deep embedded spaces," in *Medical Imaging 2018: Digital Pathology*, vol. 10581, p. 105810I, International Society for Optics and Photonics, 2018.

[72] Y. Li and W. Ping, "Cancer metastasis detection with neural conditional random field," *arXiv preprint arXiv:1806.07064*, 2018.

[73] B. E. Bejnordi, G. Zuidhof, M. Balkenhol, M. Hermsen, P. Bult, B. van Ginneken, and et al., "Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images," *Journal of Medical Imaging*, vol. 4, no. 4, p. 044504, 2017.

[74] A. Agarwalla, M. Shaban, and N. M. Rajpoot, "Representation-aggregation networks for segmentation of multi-gigapixel histology images," *arXiv preprint arXiv:1707.08814*, 2017.

[75] B. Kong, X. Wang, Z. Li, Q. Song, and S. Zhang, "Cancer metastasis detection via spatially structured deep network," in *International Conference on Information Processing in Medical Imaging*, pp. 236–248, Springer, 2017.

[76] K. Sirinukunwattana, N. K. Alham, C. Verrill, and J. Rittscher, "Improving whole slide segmentation through visual context-a systematic study," *arXiv preprint arXiv:1806.04259*, 2018.

[77] S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology," in *Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on*, pp. 284–287, IEEE, 2008.

[78] R. Farjam, H. Soltanian-Zadeh, K. Jafari-Khouzani, and R. A. Zoroofi, "An image analysis approach for automatic malignancy determination of

prostate pathological images," *Cytometry Part B: Clinical Cytometry*, vol. 72, no. 4, pp. 227–240, 2007.

[79] K. Nguyen, B. Sabata, and A. K. Jain, "Prostate cancer grading: Gland segmentation and structural features," *Pattern Recognition Letters*, vol. 33, no. 7, pp. 951–961, 2012.

[80] S. Rathore, M. A. Iftikhar, A. Chaddad, T. Niazi, T. Karasic, and M. Bilello, "Segmentation and grade prediction of colon cancer digital pathology images across multiple institutions," *Cancers*, vol. 11, no. 11, p. 1700, 2019.

[81] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, IEEE, 2017.

[82] 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.

[83] 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*, 2019.

[84] K. Sirinukunwattana, J. P. Pluim, H. Chen, X. Qi, P.-A. Heng, Y. B. Guo, L. Y. Wang, B. J. Matuszewski, E. Bruni, U. Sanchez, *et al.*, "Gland segmentation in colon histology images: The glas challenge contest," *Medical image analysis*, vol. 35, pp. 489–502, 2017.

[85] Y. Zhou, O. F. Onder, Q. Dou, E. Tsougenis, H. Chen, and P.-A. Heng, "Cia-net: Robust nuclei instance segmentation with contour-aware information aggregation," in *International Conference on Information Processing in Medical Imaging*, pp. 682–693, Springer, 2019.

[86] 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.

[87] 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*, pp. 234–241, Springer, 2015.

[88] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[89] S. E. A. Raza, L. Cheung, M. Shaban, S. Graham, D. Epstein, S. Pelengaris, M. Khan, and N. M. Rajpoot, "Micro-net: A unified model for segmentation of various objects in microscopy images," *Medical image analysis*, vol. 52, pp. 160–173, 2019.

[90] H. Chen, X. Qi, L. Yu, Q. Dou, J. Qin, and P.-A. Heng, "Dcan: Deep contour-aware networks for object instance segmentation from histology images," *Medical image analysis*, vol. 36, pp. 135–146, 2017.

[91] Z. Guo, H. Liu, H. Ni, X. Wang, M. Su, W. Guo, K. Wang, T. Jiang, and Y. Qian, "A fast and refined cancer regions segmentation framework in whole-slide breast pathological images," *Scientific reports*, vol. 9, no. 1, pp. 1–10, 2019.

[92] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," *arXiv preprint arXiv:1606.05718*, 2016.

[93] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[94] G. Aresta, T. Araújo, S. Kwok, S. S. Chennamsetty, M. Safwan, V. Alex, B. Marami, M. Prastawa, M. Chan, M. Donovan, *et al.*, "Bach: Grand challenge on breast cancer histology images," *Medical image analysis*, vol. 56, pp. 122–139, 2019.

[95] O. G. Geessink, A. Baidoshvili, J. M. Klaase, B. E. Bejnordi, G. J. Litjens, G. W. van Pelt, W. E. Mesker, I. D. Nagtegaal, F. Ciompi, and J. A. van der Laak, "Computer aided quantification of intratumoral stroma yields an independent prognosticator in rectal cancer," *Cellular Oncology*, vol. 42, no. 3, pp. 331–341, 2019.

[96] J. N. Kather, J. Krisam, P. Charoentong, T. Luedde, E. Herpel, C.-A. Weis, T. Gaiser, A. Marx, N. A. Valous, D. Ferber, *et al.*, "Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study," *PLoS medicine*, vol. 16, no. 1, 2019.

[97] J. Saltz, R. Gupta, L. Hou, T. Kurc, P. Singh, V. Nguyen, D. Samaras, K. R. Shroyer, T. Zhao, R. Batiste, *et al.*, "Spatial organization and mo-

lecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images," *Cell reports*, vol. 23, no. 1, p. 181, 2018.

[98] L. E. Tracy, R. A. Minasian, and E. Caterson, "Extracellular matrix and dermal fibroblast function in the healing wound," *Advances in wound care*, vol. 5, no. 3, pp. 119–136, 2016.

[99] E. Sahai, I. Astsaturov, E. Cukierman, D. G. DeNardo, M. Egeblad, R. M. Evans, D. Fearon, F. R. Greten, S. R. Hingorani, T. Hunter, *et al.*, "A framework for advancing our understanding of cancer-associated fibroblasts," *Nature Reviews Cancer*, pp. 1–13, 2020.

[100] N. Erez, M. Truitt, P. Olson, and D. Hanahan, "Cancer-associated fibroblasts are activated in incipient neoplasia to orchestrate tumor-promoting inflammation in an nf-κb-dependent manner," *Cancer cell*, vol. 17, no. 2, pp. 135–147, 2010.

[101] E. Giannoni, F. Bianchini, L. Masieri, S. Serni, E. Torre, L. Calorini, and P. Chiarugi, "Reciprocal activation of prostate cancer cells and cancer-associated fibroblasts stimulates epithelial-mesenchymal transition and cancer stemness," *Cancer research*, vol. 70, no. 17, pp. 6945–6956, 2010.

[102] Y. Yuan, "Modelling the spatial heterogeneity and molecular correlates of lymphocytic infiltration in triple-negative breast cancer," *Journal of The Royal Society Interface*, vol. 12, no. 103, p. 20141153, 2015.

[103] Y. Yuan, H. Failmezger, O. M. Rueda, H. R. Ali, S. Gräf, S.-F. Chin, R. F. Schwarz, C. Curtis, M. J. Dunning, H. Bardwell, *et al.*, "Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling," *Science translational medicine*, vol. 4, no. 157, pp. 157ra143–157ra143, 2012.

[104] A. Getis and J. K. Ord, "The analysis of spatial association by use of distance statistics," *Geographical analysis*, vol. 24, no. 3, pp. 189–206, 1992.

[105] C. C. Maley, K. Koelble, R. Natrajan, A. Aktipis, and Y. Yuan, "An ecological measure of immune-cancer colocalization as a prognostic factor for breast cancer," *Breast Cancer Research*, vol. 17, no. 1, p. 131, 2015.

[106] H. S. Horn, "Measurement of overlap in comparative ecological studies," *The American Naturalist*, vol. 100, no. 914, pp. 419–424, 1966.

[107] R. M. Bremnes, T. Dønnem, S. Al-Saad, K. Al-Shibli, S. Andersen, R. Sirera, C. Camps, I. Marinez, and L.-T. Busund, "The role of tumor

stroma in cancer progression and prognosis: emphasis on carcinoma-associated fibroblasts and non-small cell lung cancer," *Journal of thoracic oncology*, vol. 6, no. 1, pp. 209–217, 2011.

[108] C. Lan, A. Heindl, X. Huang, S. Xi, S. Banerjee, J. Liu, and Y. Yuan, "Quantitative histology analysis of the ovarian tumour microenvironment," *Scientific reports*, vol. 5, p. 16317, 2015.

[109] P. Haase, "Spatial pattern analysis in ecology based on ripley's k-function: Introduction and methods of edge correction," *Journal of vegetation science*, vol. 6, no. 4, pp. 575–582, 1995.

[110] H. Failmezger, S. Muralidhar, A. Rullan, C. E. de Andrea, E. Sahai, and Y. Yuan, "Topological tumor graphs: a graph-based spatial model to infer stromal recruitment for immunosuppression in melanoma histology," *Cancer Research*, vol. 80, no. 5, pp. 1199–1209, 2020.

[111] A. Heindl, C. Lan, D. N. Rodrigues, K. Koelble, and Y. Yuan, "Similarity and diversity of the tumor microenvironment in multiple metastases: critical implications for overall and progression-free survival of high-grade serous ovarian cancer," *Oncotarget*, vol. 7, no. 44, p. 71123, 2016.

[112] K. Sirinukunwattana, D. Snead, D. Epstein, Z. Aftab, I. Mujeeb, Y. W. Tsang, I. Cree, and N. Rajpoot, "Novel digital signatures of tissue phenotypes for predicting distant metastasis in colorectal cancer," *Scientific reports*, vol. 8, no. 1, pp. 1–13, 2018.

[113] C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[114] N. Mantel, "Evaluation of survival data and two new rank order statistics arising in its consideration," *Cancer Chemother. Rep.*, vol. 50, pp. 163–170, 1966.

[115] A. Wald, "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Transactions of the American Mathematical society*, vol. 54, no. 3, pp. 426–482, 1943.

[116] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, "Evaluating the yield of medical tests," *Jama*, vol. 247, no. 18, pp. 2543–2546, 1982.

[117] T. Qaiser, K. Sirinukunwattana, K. Nakane, Y.-W. Tsang, D. Epstein, and N. Rajpoot, "Persistent homology for fast tumor segmentation in whole slide histology images," *Procedia Computer Science*, vol. 90, pp. 119–124, 2016.

[118] B. Stewart, C. P. Wild, *et al.*, "World cancer report 2014," *Health*, 2017.

[119] J. Jass, W. Atkin, J. Cuzick, H. Bussey, B. Morson, J. Northover, and et al., "The grading of rectal cancer: historical perspectives and a multivariate analysis of 447 cases," *Histopathology*, vol. 10, no. 5, pp. 437–459, 1986.

[120] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *arXiv preprint*, 2016.

[121] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang, "Mdnet: A semantically and visually interpretable medical image diagnosis network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6428–6436, 2017.

[122] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*, pp. 160–167, ACM, 2008.

[123] X. Chu, W. Ouyang, W. Yang, and X. Wang, "Multi-task recurrent neural network for immediacy prediction," in *Proceedings of the IEEE international conference on computer vision*, pp. 3352–3360, 2015.

[124] B. McCann, N. S. Keskar, C. Xiong, and R. Socher, "The natural language decathlon: Multitask learning as question answering," *arXiv preprint arXiv:1806.08730*, 2018.

[125] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3712–3722, 2018.

[126] G. Hinton, N. Srivastava, and K. Swersky, "Neural networks for machine learning lecture 6a overview of mini-batch gradient descent,"

[127] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[128] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2424–2433, 2016.

[129] M. M. Kokar, J. A. Tomasik, and J. Weyman, "Data vs. decision fusion in the category theory framework," *FUSION 2001*, 2001.

[130] I. Catacchio, N. Silvestris, E. Scarpi, L. Schirosi, A. Scattone, and A. Mangia, "Intratumoral, rather than stromal, cd8+ t cells could be a potential negative prognostic marker in invasive breast cancer patients," *Translational oncology*, vol. 12, no. 3, pp. 585–595, 2019.

[131] H. Angell and J. Galon, "From the immune contexture to the Immunoscore: the role of prognostic and predictive immune markers in cancer," *Current opinion in immunology*, vol. 25, no. 2, pp. 261–267, 2013.

[132] W. H. Fridman, F. Pagès, C. Sautès-Fridman, and J. Galon, "The immune contexture in human tumours: impact on clinical outcome," *Nature Reviews Cancer*, vol. 12, no. 4, p. nrc3245, 2012.

[133] G. Corredor, X. Wang, Y. Zhou, C. Lu, P. Fu, K. N. Syrigos, D. L. Rimm, M. Yang, E. Romero, K. A. Schalper, *et al.*, "Spatial architecture and arrangement of tumor-infiltrating lymphocytes for predicting likelihood of recurrence in early-stage non-small cell lung cancer," *Clinical Cancer Research*, pp. clincanres–2013, 2018.

[134] A. Andrion, C. Magnani, P. Betta, A. Donna, F. Mollo, M. Scelsi, P. Bernardi, M. Botta, and B. Terracini, "Malignant mesothelioma of the pleura: interobserver variability.," *Journal of clinical pathology*, vol. 48, no. 9, pp. 856–860, 1995.

[135] E. J. de Ruiter, M. L. Ooft, L. A. Devriese, and S. M. Willems, "The prognostic role of tumor infiltrating T-lymphocytes in squamous cell carcinoma of the head and neck: A systematic review and meta-analysis," *OncoImmunology*, vol. 6, no. 11, p. e1356148, 2017.

[136] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, pp. 91–99, 2015.

[137] A. R. McIntosh, B. L. Peckarsky, and B. W. Taylor, "Predator-induced resource heterogeneity in a stream food web," *Ecology*, vol. 85, no. 8, pp. 2279–2290, 2004.

[138] J. D. Scalon, M. B. L. Avelar, G. d. F. Alves, and M. S. Zacarias, "Spatial and temporal dynamics of coffee-leaf-miner and predatory wasps in organic coffee field in formation," *Ciência Rural*, vol. 41, no. 4, pp. 646–652, 2011.

[139] J. Xu, L. Xiang, Q. Liu, H. Gilmore, J. Wu, J. Tang, and A. Madabhushi, "Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images," *IEEE transactions on medical imaging*, vol. 35, no. 1, pp. 119–130, 2016.

[140] C. Li, X. Wang, W. Liu, and L. J. Latecki, "Deepmitosis: Mitosis detection via deep detection, verification and segmentation networks," *Medical image analysis*, vol. 45, pp. 121–133, 2018.

[141] S. Javed, M. M. Fraz, D. Epstein, D. Snead, and N. M. Rajpoot, "Cellular community detection for tissue phenotyping in histology images," in *Computational Pathology and Ophthalmic Medical Image Analysis*, pp. 120–129, Springer, 2018.

[142] S. Partlová, J. Bouček, K. Kloudová, E. Lukešová, M. Zábrodskỳ, M. Grega, J. Fučíková, I. Truxová, R. Tachezy, R. Špíšek, *et al.*, "Distinct patterns of intratumoral immune cell infiltrates in patients with hpv-associated compared to non-virally induced head and neck squamous cell carcinoma," *Oncoimmunology*, vol. 4, no. 1, p. e965570, 2015.

[143] J. Fang, X. Li, D. Ma, X. Liu, Y. Chen, Y. Wang, V. W. Y. Lui, J. Xia, B. Cheng, and Z. Wang, "Prognostic significance of tumor infiltrating immune cells in oral squamous cell carcinoma," *BMC cancer*, vol. 17, no. 1, p. 375, 2017.

[144] T. J. Honkanen, T. Moilanen, P. Karihtala, S. Tiainen, P. Auvinen, J. P. Väyrynen, M. Mäkinen, and J. P. Koivunen, "Prognostic and predictive role of spatially positioned tumour infiltrating lymphocytes in metastatic HER2 positive breast cancer treated with trastuzumab," *Scientific reports*, vol. 7, no. 1, p. 18027, 2017.

[145] 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*, p. 1, 2018.

[146] S. Maman and I. P. Witz, "A history of exploring cancer in context," *Nature Reviews Drug Discovery*, vol. 17, no. 3, pp. 13–30, 2018.

[147] B. Lim, W. A. Woodward, X. Wang, J. M. Reuben, and N. T. Ueno, "Inflammatory breast cancer biology: the tumour microenvironment is key," *Nature reviews Cancer*, vol. 18, no. 8, pp. 485–499, 2018.

[148] M. Wang, J. Zhao, L. Zhang, F. Wei, Y. Lian, Y. Wu, Z. Gong, S. Zhang, J. Zhou, K. Cao, *et al.*, "Role of tumor microenvironment in tumorigenesis," *Journal of Cancer*, vol. 8, no. 5, p. 761, 2017.

[149] F. R. Balkwill, M. Capasso, and T. Hagemann, "The tumor microenvironment at a glance," 2012.

[150] A. Heindl, I. Sestak, K. Naidoo, J. Cuzick, M. Dowsett, and Y. Yuan, "Relevance of spatial heterogeneity of immune infiltration for predicting risk of recurrence after endocrine therapy of er+ breast cancer," *JNCI: Journal of the National Cancer Institute*, vol. 110, no. 2, pp. 166–175, 2018.

[151] L. Chan, M. S. Hosseini, C. Rowsell, K. N. Plataniotis, and S. Damaskinos, "Histosegnet: Semantic segmentation of histological tissue type in whole slide images," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 10662–10671, 2019.

[152] 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.

[153] S. Graham, M. Shaban, T. Qaiser, S. A. Khurram, and N. Rajpoot, "Classification of lung cancer histology images using patch-level summary statistics," in *Medical Imaging 2018: Digital Pathology*, vol. 10581, p. 1058119, International Society for Optics and Photonics, 2018.

[154] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

[155] C. Kramer, K. Vangangelt, G. van Pelt, T. Dekker, R. Tollenaar, and W. Mesker, "The prognostic value of tumour–stroma ratio in primary breast cancer with special attention to triple-negative tumours: a review," *Breast cancer research and treatment*, vol. 173, no. 1, pp. 55–64, 2019.

[156] G. van Pelt, S. Kjær-Frifeldt, J. van Krieken, R. Al Dieri, H. Morreau, R. Tollenaar, F. B. Sørensen, and W. Mesker, "Scoring the tumor-stroma ratio in colon cancer: procedure and recommendations," *Virchows Archiv*, vol. 473, no. 4, pp. 405–412, 2018.

[157] C. G. A. Network *et al.*, "Comprehensive genomic characterization of head and neck squamous cell carcinomas," *Nature*, vol. 517, no. 7536, p. 576, 2015.

[158] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Computer graphics and applications*, vol. 21, no. 5, pp. 34–41, 2001.

[159] V. Thorsson, D. L. Gibbs, S. D. Brown, D. Wolf, D. S. Bortone, T.-H. O. Yang, E. Porta-Pardo, G. F. Gao, C. L. Plaisier, J. A. Eddy, *et al.*, "The immune landscape of cancer," *Immunity*, vol. 48, no. 4, pp. 812–830, 2018.

[160] B. Peltanova, M. Raudenska, and M. Masarik, "Effect of tumor microenvironment on pathogenesis of the head and neck squamous cell carcinoma: a systematic review," *Molecular cancer*, vol. 18, no. 1, p. 63, 2019.

[161] M. E. Spector, E. Bellile, L. Amlani, K. Zarins, J. Smith, J. C. Brenner, L. Rozek, A. Nguyen, D. Thomas, J. B. McHugh, *et al.*, "Prognostic value of tumor-infiltrating lymphocytes in head and neck squamous cell carcinoma," *JAMA Otolaryngology–Head & Neck Surgery*, vol. 145, no. 11, pp. 1012–1019, 2019.

[162] E. J. de Ruiter, M. L. Ooft, L. A. Devriese, and S. M. Willems, "The prognostic role of tumor infiltrating t-lymphocytes in squamous cell carcinoma of the head and neck: A systematic review and meta-analysis," *Oncoimmunology*, vol. 6, no. 11, p. e1356148, 2017.

[163] Q. Xu, C. Wang, X. Yuan, Z. Feng, and Z. Han, "Prognostic value of tumor-infiltrating lymphocytes for patients with head and neck squamous cell carcinoma," *Translational oncology*, vol. 10, no. 1, pp. 10–16, 2017.

[164] S. Hendry, R. Salgado, T. Gevaert, P. A. Russell, T. John, B. Thapa, M. Christie, K. Van De Vijver, M. V. Estrada, P. I. Gonzalez-Ericsson, *et al.*, "Assessing tumor infiltrating lymphocytes in solid tumors: a practical review for pathologists and proposal for a standardized method from the international immuno-oncology biomarkers working group: Part 1: Assessing the host immune response, tils in invasive breast carcinoma and ductal carcinoma in situ, metastatic tumor deposits and areas for further research," *Advances in anatomic pathology*, vol. 24, no. 5, p. 235, 2017.

[165] S. Hendry, R. Salgado, T. Gevaert, P. A. Russell, T. John, B. Thapa, M. Christie, K. Van De Vijver, M. V. Estrada, P. I. Gonzalez-Ericsson, *et al.*, "Assessing tumor infiltrating lymphocytes in solid tumors: a practical review for pathologists and proposal for a standardized method from the international immuno-oncology biomarkers working group: Part 2: Tils in melanoma, gastrointestinal tract carcinomas, non-small cell lung carcinoma and mesothelioma, endometrial and ovarian carcinomas, squamous cell carcinoma of the head and neck, genitourinary carcinomas, and primary brain tumors," *Advances in anatomic pathology*, vol. 24, no. 6, p. 311, 2017.

[166] N. Kemi, M. Eskuri, A. Herva, J. Leppänen, H. Huhta, O. Helminen, J. Saarnio, T. J. Karttunen, and J. H. Kauppila, "Tumour-stroma ratio and prognosis in gastric adenocarcinoma," *British journal of cancer*, vol. 119, no. 4, pp. 435–439, 2018.

[167] O. J. Hénaff, A. Srinivas, J. De Fauw, A. Razavi, C. Doersch, S. Eslami, and A. v. d. Oord, "Data-efficient image recognition with contrastive predictive coding," *arXiv preprint arXiv:1905.09272*, 2019.

[168] R. J. Chen, M. Y. Lu, J. Wang, D. F. Williamson, S. J. Rodig, N. I. Lindeman, and F. Mahmood, "Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis," *arXiv preprint arXiv:1912.08937*, 2019.

[169] A. Cheerla and O. Gevaert, "Deep learning with multimodal representation for pancancer prognosis prediction," *Bioinformatics*, vol. 35, no. 14, pp. i446–i454, 2019.

[170] 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.