Using Machine Learning Techniques on Digitised Histopathological Images to Distinguish Lipoma from Atypical Lipomatous Tumours

Binghao Chai

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London

Supervisors: Dr. Kevin Bryson and Dr. Nischalan Pillay

Department of Computer Science

University College London

May 18, 2023

I, Binghao Chai, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Lipomas are benign neoplasms of fat and are amongst the most common tumours with an estimated incidence rate of 1 per 1000 people per year (Rydholm & Berg 1983, Weiss et al. 2007). They need to be distinguished from atypical lipomatous tumours (ALT) which are malignant and rare and present both a clinical and histological challenge. The distinction is made by assessing nuclear and subtle architectural features which require reviewing multiple sections and the use of ancillary genetic testing. Benign fatty tumours, therefore, present a considerable workload in the general pathology setting and the distinction from malignancy often require specialist review. This problem could be addressed by automated whole slide image (WSI) analysis, however, there is a lack of existing tools for this task. The paucicellular nature of the fatty tissues also presents a computational challenge which is addressed in this study.

A tile-based deep learning workflow is developed to quantitatively analyse fatty tumour WSIs. A total of 206 internal and 402 external slides were employed for the training, validation and testing of the convolutional neural networks. A semi-automatic nucleus annotating workflow is also proposed, and a nucleus-level lipoma and ALT dataset is developed for the tasks of fatty tumour nucleus classification and segmentation. A lipoma and ALT nucleus detector is trained on top of this dataset. The tile-based workflow and the nucleus detector are then integrated for the slide-level classification of lipoma and ALT.

The pipeline achieved 70.31% overall accuracy with an AUC of 0.777 for slide-level diagnostic classification using the real-world dataset. The results suggest that WSI analysis of fatty tumours is feasible but presents a unique set of challenges.

Impact Statement

With the increasing routine implementation of whole slide scanning and storing technologies in clinical practice, massive archives of digitised tissue slides are being created. This new source of medical imaging data opens up exciting opportunities for computer scientists interested in digital pathology research. The workflows and application of existing methods that I have developed in this study enable the classification of benign and malignant fatty tumours at the digital slide level. To my knowledge, this is one of the first couple of such implementations that have been developed for tumours of fat. These solutions require further validation and could be integrated into the routine clinical pathology pathways to support and assist pathologists in making more rapid and accurate diagnoses.

There is currently a worldwide shortage of diagnostic histopathologists. The pathologist-to-inhabitant ratio in developed countries such as the USA and most European countries is about or below 1:30,000 (Märkl et al. 2021). However, this ratio is around 1:70,000 in Poland or China and is greater than 1:1,000,000 in the Sub-Saharan countries. Therefore, tools such as artificial intelligence models used by pathologists could aid in alleviating some of the more routine aspects of work, thereby freeing up time for pathologists to devote to more complex tasks of their jobs. Workflows such as those proposed in my study add value for both routine and bone and soft tissue specialist pathologists and complement other such methods being developed for more common cancers. By presenting the regions of concern detected by the model as well as proposing a putative diagnostic label of a digitised slide based on an aggregation of features, computational approaches could help pathologists triage cases appropriately for ancillary tests such as immunohisto-

chemistry or molecular testing and thereby make their decisions in a more efficient and timely way.

Acknowledgements

I would like to thank my supervisors Dr Kevin Bryson and Dr Nischalan Pillay, for their invaluable support, guidance and persistent patience during my PhD journey. I would like to extend my thanks to Dr Florian Puls, Dr Hongtao Ye, Elena Miranda and Dahmane Oukrif for their help in collecting sample data and digitising pathological slides. I would also like to thank Dr Constantino Carlos Reyes-Aldasoro and Professor Gary Zhang for their suggestions for my research project. And I would like to thank Dr Povilas Norvaisas, Dr Xianze Hong and Dr Christopher Steele for their advice when I was a junior PhD student in my first year.

I would like to thank UCL Computer Science department, UCL Cancer Institute and Royal National Orthopaedic Hospital for all the research and administrative support during my studies.

In addition to the above, I would also like to thank my parents and family, they have supported me, believed in me and encouraged me. Thanks for your infinite patience and understanding.

And finally, I would like to show my thanks to my friends Bingquan Wang, Xingyi Li, Jiayun Wang, Huaqiu Chen, Yuan Li, Jingkai Mao, Hanxi Qiao, Beini Chen, Keer Pan, Fangxin Xie, Yihao Zou, Kailun Yuan, Yang Zheng, Tao Liu, Jiayin Shi and Qianyu Liu for your support and inspiration over the years.

Contents

1 Introduction **23**

1.1 Chapter Overview 23

1.2 General Background 23

1.3 Whole Slide Imaging 25

1.4 Lipoma, Liposarcoma and Atypical Lipomatous Tumour 26

1.4.1 Lipoma and Liposarcoma 26

1.4.2 Atypical Lipomatous Tumour versus Well-differentiated Liposarcoma 27

1.4.3 Lipoma and Atypical Lipomatous Tumour 27

1.5 Deep Convolutional Neural Network 28

1.6 Challenges in this Study 30

1.6.1 Large Whole Slide Image Loading 32

1.6.2 Whole Slide Images Scanning Quality Control 32

1.6.3 Colour Variation, Deconvolution and Normalisation 33

2 The Data **39**

2.1 Chapter Overview 39

2.2 Internal Whole Slide Images 39

2.2.1 Quality Issue and Quality Control 40

2.2.2 Image Storing 41

2.2.3 Internal Slides Overview 41

2.2.4 Tile the Whole Slide Images 45

2.2.5 Tile-Level Data Augmentation 47

2.3 External Whole Slide Images 48

3 A Tile-Based Lipoma and Atypical Lipomatous Tumour Classifier 51

3.1 Chapter Overview 51

3.2 Introduction and Related Work 51

3.2.1 Tile-Based Method for the Benefits of Diagnosis 51

3.2.2 Tile-Aggregation Methods 54

3.2.3 Computational Tools in Digital Pathology 54

3.3 Methodology 56

3.3.1 The Tile-Based Workflow: an Overview 56

3.3.2 Tiles Generation and Annotation Strategy 58

3.3.3 The Tile-Level CNNs: the Evolution 62

3.4 Optimisation of Methods 67

3.4.1 The Tile-Level CNN Classifier 67

3.4.2 The Slide-Level Visualisation 70

3.4.3 Colour Normalisation 75

3.4.4 Slide-Level Classification Using Tile Aggregation Methods 76

3.4.5 Methodology Summary 87

3.5 Results and Discussion 88

3.6 Conclusions 93

4 A Novel Lipoma and Atypical Lipomatous Tumour Nucleus Dataset 95

4.1 Chapter Overview 95

4.2 Introduction and Related Work 96

4.2.1 Existing Nucleus-Level Datasets 96

4.2.2 Traditional Segmentation Methods 96

4.2.3 Nuclei Classification, Detection and Segmentation Using Deep Learning Methods 98

4.2.4 Image Registration 101

4.3 Methodology: A Semi-Automatic Nucleus Annotating Workflow . . 102

4.3.1 *HoverNet* Pretrained with *PanNuke* Dataset 102

4.3.2 The Semi-Automatic Lipoma and ALT Nucleus Annotating Work ow: An Overview 105
4.3.3 Nuclei Tile Selection 107
4.3.4 Label Correction 108
4.3.5 Large Aberrant Nuclei Augmentation Strategy 109
4.3.6 Colour Normalisation 113
4.4 Results and Discussion 114
4.4.1 Applying the Standard HoverNet-PanNuke on Sarcoma Slides 114
4.4.2 The Novel Lipoma and ALT Nucleus Dataset 118
4.5 Conclusions 120

5 Slide-Level Classi cation between Lipoma and ALT Using Handcrafted Nucleus Features 122
5.1 Chapter Overview 122
5.2 Related Work 122
5.3 Methodology 123
5.3.1 Nucleus and Their Underlying Masks 123
5.3.2 The Nucleus-Level Handcrafted Features 124
5.4 Results and Discussion 127
5.4.1 Initial Case Study of One Lipoma and One ALT Sample . . 128
5.4.2 Handcrafted Feature Analysis 135
5.4.3 Slide-Level Classi cation Using Handcrafted Features . . . 139
5.5 Conclusions 144

6 Slide-Level Classi cation between Lipoma and ALT Using HoverNet 146
6.1 Chapter Overview 146
6.2 Introduction 146
6.3 Methodology 147
6.3.1 Retrain the HoverNet 147
6.3.2 Evaluation Metrics for the Retraining Process 147
6.3.3 Preliminary Analysis: Consideration of Colour Normalisation 149

6.3.4 Slide-Level Classi cation Using HoverNet-Sarcoma-2 . . . 154
6.4 Results and Discussion 155
6.4.1 The HoverNet-Sarcoma-2 155
6.4.2 The Evaluation of the Slide-level Classi er Using Internal and External Testing Slides 163
6.5 An Integration of the Tile-Based Classi er and the Nucleus Detector 167
6.5.1 Introduction 167
6.5.2 Methodology 167
6.5.3 Results and Discussion 169
6.6 Conclusions 171

7 Conclusions and Future Studies 174
7.1 Conclusions 174
7.2 Suggestions for Future Studies 178

Appendices 181

A List of Internal Whole Slide Images 181

B The Mapping Table for Clinical and Internal IDs 184

C List of External Whole Slide Images 187

D QuPath Annotation Export Script 190

E High Performance Computing Jobscript Example 192

F HPC Deployment Technical Note 194

G Tools Installation Instructions 195

H The Slide-Level Handcrafted Nucleus Features 196

I Raw Results for the Nucleus Level Colour Normalisation Experiment 199

List of Figures

1.1 The cumulative numbers of digital pathology papers since 2015. . . 24

1.2 Examples of a lipoma and an atypical lipomatous tumour whole slide image digitised using Hamamatsu Nanozoomer S360 slide scanner. Due to the nature of fatty tissues, paucicellular and whitespaces in the WSIs present a computational challenge for image analysis. 26

1.3 An example of septum (black arrow) in an atypical lipomatous tumour whole slide image. 28

1.4 A sample output of skimage.color.rgb2hed() 36

2.1 Binarised whole slide image sample (internal sample ID: 021) showing scanning stripes. 40

2.2 HistoQC visualisation and hierarchical analysis on the internal dataset for quality control purposes (Hamamatsu Nanozoomer S210). 42

2.3 HistoQC visualisation and hierarchical analysis on the internal dataset for quality control purposes (Hamamatsu Nanozoomer S360). 43

2.4 An example of an atypical lipomatous tumour digital slide 206. . . . 44

2.5 An example of slide before and after tiling. 47

3.1 The proposed tile-based work ow for the prediction of lipoma and atypical lipomatous tumour. 57

3.2 An example of the annotations of sample 035. Grey annotations stand for non-informative areas, red annotations stand for atypical lipomatous tumour areas, and blue annotations stand for muscle areas. 59

3.3 Some examples of the informative tiles. 60
3.4 Some examples of the non-informative tiles. 60
3.5 A comparison of x5 and x10 magni cations tiles of lipoma, atypical lipomatous tumour and fat necrosis. 61
3.6 The tissue detection pipeline proposed by Bizzego et al. (2018). The identi cation of the tissue bounding box is performed on the WSI thumbnail in three steps: (1) binarisation of the grayscale image using Otsu's algorithm, (2) binary dilation and lling the holes, and (3) biggest connected region detection. 62
3.7 A success and a failed example for applying the normal tissue detection pipeline on our own fatty tissue dataset. 64
3.8 The tile-level ROC for the Alex-based CNNs for the malignant class of atypical lipomatous tumours. 65
3.9 WSI-level visualisation using class prediction map for Alex-CNNs on lipoma sample 137. In the class prediction maps, the green colour stands for lipoma and red colour stands for atypical lipomatous tumour, grey colour stands for non-informative tiles including background and all other colours stand for minor histology variations (details in Section 3.4.2). 66
3.10 Tile-level confusion matrix for Inception-7 and Inception-7-x5 models. 68
3.11 Tile-level ROC plots for Inception-7 and Inception-7-x5 models. 69
3.12 Inception-7-x5 (ne-tuned) evaluation. 69
3.13 A example class-prediction map of an atypical lipomatous tumour case 102 before and after background smoothing. 71
3.14 Two examples (001 and 206) of the class-prediction map with one conventional lipoma case and one conventional atypical lipomatous tumour case. 71
3.15 Another two class-prediction map examples (035 and 097) for atypical lipomatous tumour cases with some histology variation areas. . 72

3.16 Comparison of the tissue detection task between traditional threshold-based pipeline and our tile-based CNN methods for sample 051. 73
3.17 Two examples (001 and 206) of the heat map with one conventional lipoma case and one conventional atypical lipomatous tumour case. 73
3.18 Another two heat map examples (035 and 097) for atypical lipomatous tumour cases with some histology variation areas. 74
3.19 Some samples for lipoma, atypical lipomatous tumour, muscle, calci cation, in ammation and fat necrosis tiles before and after colour normalisation using different algorithms. 75
3.20 Tile-level malignant score distribution (all tiles included) samples. . 78
3.21 Tile-level distribution for all the putative lipoma and ALT tiles for a lipoma and an atypical lipomatous sample. 78
3.22 Tile-level malignancy distribution for all training slides with only informative (i.e., putative benign tiles and putative malignant) tiles. . 80
3.23 Contour map example for atypical lipomatous tumour sample 206. . 81
3.24 Scatter plot based on putative benign (contour value from 0 to 0.378 shown in blue) and putative malignant (contour value from 0.747 to 1 shown in orange and red) areas of contour maps for all the training slides. 83
3.25 The data structure of all mentioned putative features for the internal training slides visualised using PCA as well as t-SNE. 84
3.26 Hyperparameter search of the putative benign and putative malignant contour lines using the validation accuracy. 86
3.27 Slide-level ROC curves for the tile-based work ow on the internal testing dataset. 89

4.1 An example of PanNuke dataset. 103
4.2 PanNuke dataset nuclei type statistics (Gamper et al. 2019, 2020, PanNuke: An Open Pan-Cancer Histology Dataset for Nuclei Instance Segmentation and Classi cation 2019). 104

4.3 The proposed semi-automatic nuclei labelling work ow. 106

4.4 The large ALT nuclei label correction sub- ow. 109

4.5 The proposed large ALT nuclei label correction work ow: an example region from case 206, note: only the large ALT nuclei are shown in this gure as a highlight, and all the small nuclei are hidden. 110

4.6 The large ALT nuclei augmentation sub- ow. 112

4.7 Two regional HoverNet inference result examples for lipoma case 001. In the gure, red annotations stand for neoplastic predictions, green annotations stand for in ammatory predictions, blue annotations stand for connective predictions, yellow annotations stand for dead nuclei, and the very minor light brown annotations stand for epithelial predictions. 115

4.8 A stacked bar chart for the slide-level numbers of predicted nuclei for all the 59 internal training slides using HoverNet-PanNuke. In this gure, samples from 206 to 4 are ALTs, and samples from 157 to 1 are lipomas. 116

4.9 An example of the failure of large ALT nuclei prediction on an ALT slide 206. Only the predicted large nuclei are shown in this gure, and all the small nuclei of the predicted connective nuclei and in- ammatory nuclei are hidden in this gure. 117

4.10 Some examples of the ALT 256×256 tiles containing large aberrant nuclei. In this gure, blue annotations stand for the small ALT nuclei, red annotations stand for the aberrant lipoblast, and the green annotations stand for the in ammatory nuclei. 118

5.1 Two nuclei examples and their underlying masks. Some tiles may contain more than one nuclei (Fig 5.1a), but only the nucleus right in the tile centre is considered for those speci c tiles. 124

5.2 An example of the convex area and convex perimeter of an object. . 126

5.3 A concave nucleus example and its underlying mask from slide 001. 126

5.4 An example showing the stain differences for the nuclei in the same slide. 127

5.5 The slide-level Area, Convex Area, Perimeter and Convex Perimeter distribution for the lipoma sample 001 and the ALT sample 206. . . 129

5.6 The slide-level Major and Minor Axis Lengths distribution for the lipoma sample 001 and the ALT sample 206. 130

5.7 The slide-level Compactness and Solidity distribution for the lipoma sample 001 and the ALT sample 206. 130

5.8 The slide-level H-value and IOD distribution for the lipoma sample 001 and the ALT sample 206. 131

5.9 The clustering result using k-means (k=5) for the lipoma slide 001. T-SNE was used as the dimensional reduction method for visualisation. 133

5.10 The clustering result using K-means (k=5) for the ALT slide 206. T-SNE was used as the dimensional reduction method for visualisation. 134

5.11 The diagnostic label based scatter plot and the k-means with handcrafted features plot for the mix cases of 001 and 206. 135

5.12 Correlation analysis for the handcrafted nucleus features. 137

5.13 The box plots for the handcrafted features of convex area, compactness, h-value and IOD for lipoma (n=26) and ALT (n=33) groups. . 138

6.1 The illustration of Intersection over Union (IoU) and Dice score (Tiu 2019). 148

6.2 The slide-level proportions of different nuclei types for the training slides using HoverNet-Sarcoma-1. In this figure, samples from 206 to 4 are ALTs, and samples from 157 to 1 are lipomas. 150

6.3 The slide-level proportions of different nuclei types for the internal testing slides using HoverNet-Sarcoma-1. In this figure, samples from 205 to 23 are ALTs, and samples from 196 to 21 are lipomas. . 151

6.4 The slide-level proportions of different nuclei types for the external testing slides using HoverNet-Sarcoma-1. In this figure, samples named starting with ALT are ALTs, and samples named with L are lipomas. 152

6.5 Nucleus inference examples of HoverNet-Sarcoma-2 on randomly selected training slides, the yellow annotations are the putative inflammatory nuclei, the green annotations are the putative lipoma nuclei, the red annotations are the putative small ALT nuclei and the violet annotations are the putative large ALT nuclei. 156

6.6 The slide-level proportions of different nuclei types for the training slides using HoverNet-Sarcoma-2. In this figure, samples from 206 to 4 are ALTs, and samples from 157 to 1 are lipomas. 157

6.7 The slide-level proportions of different nuclei types for the internal testing slides using HoverNet-Sarcoma-2. In this figure, samples from 205 to 23 are ALTs, and samples from 196 to 21 are lipomas. . 158

6.8 The slide-level proportions of different nuclei types for the external testing slides using HoverNet-Sarcoma-2. In this figure, samples named starting with ALT are ALTs, and samples named with L are lipomas. 159

6.9 The box plots for the putative lipoma nuclei, the putative small ALT nuclei and the putative large ALT nuclei between the training lipoma and ALT groups using HoverNet-Sarcoma-2 to demonstrate their proportion distributions. 161

6.10 The box plots for the putative lipoma nuclei, the putative small ALT nuclei and the putative large ALT nuclei between the testing lipoma and ALT groups using HoverNet-Sarcoma-2 to demonstrate their proportion distributions. 162

7.1 The precision, recall, F1-score and AUC for different deep learning models. LR stands for the logistic regress model, RF stands for the random forest model. 176

7.2 A proposed atypical lipomatous tumour nuclei ground truth generation work ow using IHC stained images as the reference. 180

F.1 A technical note of the OpenSlide deployment method onto GPU cluster. 194

G.1 The dependence graph and installation process of StainTool 195

List of Tables

1.1 A roadmap for deep learning technologies relevant to this study. . . 25

1.2 A summary of open databases in digital histopathology. 29

2.1 An examples of the metadata of whole slide images. The samples were anonymised, and so the clinical ID was removed in this table. . 44

2.2 The internal slides count and slide selection process. 46

2.3 Internal training set statistics by diagnostic labels. 46

2.4 Internal testing set statistics by diagnostic labels. 47

2.5 External dataset: the patient list. 48

2.6 External dataset: the top scanning quality slides for each of the patients. 49

2.7 External dataset: slide-level batch 1 (53 slides in total). 49

2.8 External dataset: slide-level batch 2 (51 slides in total). 49

2.9 External dataset: slide-level batch 3 (49 slides in total). 50

3.1 Selected tile-based studies on digital pathology (chronological order). 53

3.2 Development environment and package requirements in terms of Python. 55

3.3 Training set tile-level statistics. 61

3.4 A rough tile-level evaluation for the Alex-CNNs. 65

3.5 The evolution of tile-based lipoma and atypical lipomatous tumour classi er. 66

3.6 Tile-level accuracy for Inception-7 and Inception-7-x5 models. 68

3.7 Tile-level accuracy for Inception-b-7-x5 (training from scratch) and Inception-b-7-x5 (ne-tuning). 68
3.8 Colour list in the class-prediction maps. 72
3.9 De nitions of proportions of benign and malignant tiles in the slides. 77
3.10 The baseline results for the tile-based work ow on the internal testing dataset. 77
3.11 Percentage of variance explained by each of the selected components using PCA. 85
3.12 Principal axes in feature space representing the directions of maximum variance in the data. 85
3.13 Feature ranking generated by random forest method. 85
3.14 Tile aggregation feature list in the tile-based work ow. 87
3.15 Slide-level results for the tile-based work ow on the internal training dataset. 88
3.16 Aggregation feature analysis based on logistic regression and random forest. 89
3.17 Slide-level results for the tile-based work ow on the internal testing dataset. 89
3.18 Incorrect predictions for the tile-based work ow on the internal testing dataset. 90
3.19 The comparison of the slide-level ground truth, the work ow prediction and the pathological review for the tile-based incorrectly predicted slides. 91
3.20 The slide-level results of the tile-based classi er for the internal and the external testing slides. 92

4.1 Open datasets for nucleus detection and classi cation task. 97
4.2 Input and output of HoverNet in each of the inference mode. 103
4.3 The classes of the proposed novel nucleus-level lipoma and atypical lipomatous tumour dataset. 107
4.4 The tile-level statistics for the proposed novel nucleus dataset. . . . 108

4.5 Nucleus class merge and relabel protocol. 108
4.6 The nuclei statistics for the proposed novel dataset before augmentation. 111
4.7 The nuclei statistics for the proposed novel dataset after augmentation (full version). 119
4.8 A proposed nucleus-level train-validation statistics for the HoverNet training purpose. 120

5.1 Handcrafted features on the nucleus-level. 125
5.2 Handcrafted feature importance generated using random forest (ranked by feature group and feature importance). 137
5.3 Statistical results for the features of convex area, compactness, H-value and iod. 138
5.4 Slide-level classi cation results for the internal training dataset using handcrafted nucleus features. 139
5.5 Handcrafted feature analysis: feature coef cients and feature importance for the slide-level logistic regression and random forest classi ers. 140
5.6 The incorrect prediction list (the lipoma group of internal training slides) on the slide-level together with their raw feature values with an ascending sorting of the feature convex area 141
5.7 The incorrect prediction list (the ALT group of internal training slides) on the slide-level together with their raw feature values with an ascending sorting of the feature convex area 142
5.8 Slide-level classi cation results for the testing dataset using handcrafted nucleus features and a comparison between the handcrafted-based classi er and the tile-based classi er. 143
5.9 Incorrect predictions using handcrafted features on the internal testing dataset. 144

6.1 HoverNet versions and their description. 149

6.2 Slide-level aggregation feature list in the nucleus-based work ow. . 154

6.3 The validation Dice Score in the retraining process of HoverNet-Sarcoma-2 155

6.4 The slide-level classi cation results for the training set using HoverNet 163

6.5 The slide-level classi cation results for the internal testing set using HoverNet 164

6.6 The slide-level logistic regression feature coef cients and the random forest feature importance using HoverNet. In the table, LR stands for the logistic regression model and RF stands for the random forest model. 164

6.7 The slide-level classi cation results for the external testing set using HoverNet 165

6.8 The benchmark results for the slide-level lipoma and ALT classi er using HoverNet-Sarcoma-2 166

6.9 The feature list of the integrated slide-level classi er: a combination of the tile-based and the nucleus-based classi ers. 168

6.10 The slide-level logistic regression feature coef cients and the random forest feature importance of the integrated classi er. In this table, LR stands for the logistic regression model, and RF stands for the random forest model. 169

6.11 The slide-level results of the integrated classi er for the training slides. 170

6.12 The slide-level results of the integrated classi er for the internal and the external testing slides. 170

6.13 A comparison of the slide-level classi cation results on the internal and external testing slides of the tile-based, the nucleus-based and the integrated work ows. 171

7.1 The precision, recall, F1-score and AUC for different deep learning models. LR stands for the logistic regression model, RF stands for the random forest model. 177

A.1 Metadata of Internal Training Slides. 182
A.2 Metadata of Internal Testing Slides. 183

B.1 The Mapping Table for Clinical and Internal IDs (1). 184
B.2 The Mapping Table for Clinical and Internal IDs (2). 185
B.3 The Mapping Table for Clinical and Internal IDs (3). 186

C.1 Metadata of All External Slides. 187

H.1 The mean values of the slide-level handcrafted nucleus features. . . 197

I.1 Raw nuclei counting colour normalisation results for selected internal tiles using HoverNet-Sarcoma-1 and HoverNet-Sarcoma-2. . . 199
I.2 Nuclei counting for selected external tiles using HoverNet-Sarcoma-1 and HoverNet-Sarcoma-2 202

Chapter 1

Introduction

1.1 Chapter Overview

In this chapter, the general background and research trends of histopathology and whole slide imaging are given, followed by an introduction to the benign lipoma and the malignant liposarcoma and atypical lipomatous tumour. The current situation of distinguishing benign from malignant fatty tumours is also described. Recent literature indicates that computational methods can be employed to address this question. The challenges of this study are also highlighted at the end of this chapter.

1.2 General Background

Histopathology is a medical speciality encompassing the study of mechanisms of disease and tissue reactions to disease processes through the microscopic examination of tissue sections and serves as the “gold standard” for the diagnosis of many diseases, including cancer. Tissue samples are chemically xed using buffered formalin and then processed such that thin sections can be generated. The tissue is embedded in paraf n, commonly referred to as formalin- xed paraf n-embedded (FFPE) tissue. The next step is sectioning, which is cutting the thin slices of tissue and mounting them on glass slides. Each of the tissue sections is 3 to 5 µm thick and is nearly transparent under a light microscope. Staining is important at this point to create contrast. Slides for routine histopathological assessment are stained with one of the principal staining methods, the Hematoxylin and Eosin stain. Hematoxylin and Eosin stain, also known as H&E stain or HE stain, is one of the most widely

used stains for both clinical and research histology. Hematoxylin stains nucleic acids resulting in a blue-purple colour in cell nuclei, while eosin stains cytoplasm and appears pink-red depending on their constituents.

The philosophy of digital pathology is applying computational methods including machine learning and deep learning techniques to quantitatively solve the problems in the eld of pathology for the bene t of diagnosis and prognosis. The use of deep learning in digital pathology was a relative niche topic when I started my PhD in early 2018, but was developing very quickly over the past few years. So a number of challenges I initially faced in processing and analysing my data have now been overcome with the rapid pace of research in this area. Fig 1.1 shows the cumulative numbers of digital pathology papers since 2015, and the numbers in this gure come from Google Scholar when using *digital pathology* and *deep learning* as the keywords. Table 1.1 shows a roadmap for deep learning technologies relevant to this study.

Figure 1.1: The cumulative numbers of digital pathology papers since 2015.

Table 1.1: A roadmap for deep learning technologies relevant to this study.

Year	Milestones and Events
2001	First colour normalisation method in digital pathology based on colour deconvolution was proposed (Ruifrok et al. 2001)
2009	Colour normalisation method optimised by Macenko et al. (2009)
2013	Deep learning was rst applied in digital pathology for the mitosis detection task (Cireşan et al. 2013)
2015	U-net was developed for biomedical imaging segmentation tasks, including cell segmentation (Ronneberger et al. 2015)
2016	A comprehensive tutorial of deep learning for digital pathology was proposed by Janowczyk & Madabhushi (2016), and this is a practical paper with various digital pathology use cases
2018	Different tile-based deep neural networks were widely started to be employed in digital pathology for breast, lung and colon cancers (Table 3.1)
2019	HoverNet was developed for nuclei classi cation and segmentation (Graham, Vu, Raza, Azam, Tsang, Kwak & Rajpoot 2019)
2019	HistoQC was developed for whole slide imaging quality control (Janowczyk et al. 2019)
2019	Debates started to come out about whether colour normalisation should be applied to deep learning in digital pathology (Tellez et al. 2019, Hameed et al. 2020, Bianconi et al. 2020)
2019	A clinical-grade deep learning application in digital pathology was developed (Campanella et al. 2019)
2021	The rst computational application in fatty tumour digital pathology was developed (Foersch et al. 2021)

1.3 Whole Slide Imaging

Whole slide imaging is a signi cant advance in the eld of digital pathology whose processes are scanning, digitising and then storing images of histopathological glass slides taken from the biopsy or tissue resection. Whole slide images (WSIs), which are created on slide scanners, are the foundation of clinicopathological computer-aided diagnosis (CAD) systems. Most existing released CAD systems take advantage of traditional image processing approaches for digital pathology tasks, such as colour deconvolution, nuclei detection, quanti cation, segmentation, and tumour classi cation. However, there are massive variations across the pipeline of producing WSIs such as staining variation and operational bias. Van Eycke et al. (2017) indicated traditional approaches are not that intelligent in dealing with these vari-

ances, and always require extra unique preprocessing and normalisation before the execution of specific tasks. Figure 1.2 shows one example of digitised lipoma whole slide image and one for the ALT.

(a) An example of lipoma whole slide image.

(b) An example of atypical lipomatous tumour whole slide image.

Figure 1.2: Examples of a lipoma and an atypical lipomatous tumour whole slide image digitised using Hamamatsu Nanozoomer S360 slide scanner. Due to the nature of fatty tissues, paucicellular and whitespaces in the WSIs present a computational challenge for image analysis.

1.4 Lipoma, Liposarcoma and Atypical Lipomatous Tumour

1.4.1 Lipoma and Liposarcoma

Lipomas are benign neoplasms that originate from fatty tissues and are amongst the most common tumours with an estimated incidence (number of patients consulting a doctor for a lipoma, even if not histologically verified) of 1 per 1000 people, and they can be located in almost any organ of the human body (Rydholm & Berg 1983, Weiss et al. 2007). Kransdorf & Murphey (2006) indicated that most lipomas are small and usually with a diameter of less than 5 cm. A lipoma is categorised to be a giant lipoma when it reaches a size of 10 cm or weight of 1000 grams (Sanchez et al. 1993). Liposarcoma, accounting for approximately 20%-25% of all mesenchymal malignancies in a large European database (Casecabase: European sarcoma database and tumour bank, 2023, Dei Tos 2000, Coindre et al. 2010), represents one of the most common soft tissue sarcomas. Liposarcoma can be categorised into three main subgroups, based on its clinicopathological and molecular

genetic characteristics (Fletcher et al. 2002, Jones et al. 2005, Weiss et al. 2007, Gardner 2020, Hisaoka 2014):

1. atypical lipomatous tumour (ALT)/well-differentiated liposarcoma (WDLPS) and dedifferentiated liposarcoma,
2. myxoid/round-cell liposarcoma,
3. pleomorphic liposarcoma.

1.4.2 Atypical Lipomatous Tumour versus Well-differentiated Liposarcoma

The use of the terminology atypical lipomatous tumour (ALT) and well-differentiated liposarcoma (WDLPS) is based on the tumour's location in the body. WDLPS usually refers to deep-seated tumours with a high risk of local recurrence, differentiation or death. And ALT usually refers to tumours located in the periphery that have a low risk of metastasis and can be treated by resection.

1.4.3 Lipoma and Atypical Lipomatous Tumour

Lipomas need to be distinguished from their malignant mimics atypical lipomatous tumours (ALT), which are rare and present both a clinical and histological challenge. Laurino et al. (2001), Coindre et al. (2010) indicated that ALTs show a genomic pro le characterised by ampli cation of chromosome 12q14-15 region involving theMDM2 gene. ALT makes up 40%-45% of liposarcoma and is the largest subgroup of adipocytic malignancies (Weiss et al. 2007).

The distinction between the lipoma and the atypical lipomatous tumour is rst made by assessing nuclear and sometimes what can be subtle architectural features. Due to tumour heterogeneity and variation in features, this process can require reviewing multiple sections to get a more representative view of the tumour. However, many malignant fatty tumours could be mistaken for a lipoma based only on gross and microscopic features, except for their extremely large size, the tendency to have more brous bands (also known as a septum, see example in Figure 1.3), or gelatinous zones said by Weiss et al. (2007). Coindre et al. (2010) found that the detection

of the MDM2 gene amplification is presented in 100% of cases. Fluorescence in Situ Hybridisation (FISH) test, an ancillary genetic test of the detecting MDM2 amplification plays a role for the diagnosis of ALT. However, the genetic test can be time-consuming and can be costly especially if the test needs to be performed on every large lipoma specimen. Benign fatty tumours, therefore, present a considerable workload in the general pathology setting and the distinction from malignancy often require specialist review. This problem could be addressed by automated whole slide image analysis; however, there is a lack of existing tools for this task.

Figure 1.3: An example of septum (black arrow) in an atypical lipomatous tumour whole slide image.

1.5 Deep Convolutional Neural Network

The deep convolutional neural network (CNN) technique is widely employed in the classification, detection and segmentation tasks of images. CNNs are usually structured with the combination of convolutional layers (Conv), rectified linear unit layers (ReLU) and pooling layers (Pool), and will sometimes contain fully-connected layer, softmax layer or other adjustment layers. Convolutional processing is

in fact a process of abstracting activation map and feature information from the original pixel information using feature identi ers (i.e. kernel), while ReLU brings in non-linearity and improves discriminative performance (Nair & Hinton 2010). Pooling (also known as down-sampling) decreases the data size and the computing time and can reduce over- tting to some degree.

There are many successful CNN architectures such as AlexNet, VGG and Inception and so on. LeCun et al. (1995) proposed LeNet as the rst CNN for the application of handwritten digit recognition. AlexNet, VGG, Inception, ResNet, MobileNet, U-Net are widely employed in digital pathology and other medical imaging areas. These CNNs have been used for tumour subtype classi cation, pattern detection, cell and nuclei segmentation, grading and prognosis prediction tasks with promising results in pathological imaging datasets of the breast, lung, colon, glomeruli and so on. Literature lists of digital pathology datasets and their deep convolutional neural network applications can be found in Table 1.2 and Table 3.1

Table 1.2: A summary of open databases in digital histopathology.

Dataset & Year	Cancer Type	Tasks
MITOS, Tripathi et al. (2013)	Breast	Mitosis detection
MITOS-14, 22nd International Conference on Pattern Recognition (2014)	Breast	Mitosis detection
TUPAC, Veta et al. (2019)	Breast	Mitosis counting and detection
HER2, Qaiser et al. (2018)	Breast	HER2 scoring
BreakHis, Spanhol et al. (2015)	Breast	Breast cancer detection
CAMELYON-16, Bejnordi et al. (2017)	Breast	Breast cancer metastasis detection
CAMELYON-17, Litjens et al. (2018)	Breast	Breast cancer metastasis detection and pN-stage prediction
PCam, Veeling et al. (2018)	Breast	Breast Cancer metastasis detection
BACH ICIAR, Aresta et al. (2019)	Breast	Breast cancer classi cation
BreastPathQ, Peikari et al. (2017)	Breast	Tumour cellularity scoring
Post-NAT-BRCA, Assessment of residual breast cancer cellularity after neoadjuvant chemotherapy using digital pathology (Post-NAT-BRCA) (2017)	Breast	Tumour cellularity scoring

Table 1.2 (continued)

Dataset & Year	Cancer Type	Tasks
ER+ Breast Cancer, Janowczyk & Madabhushi (2016)	Breast	Detection, segmentation and classification
CCMCT, Bertram et al. (2019)	Breast	Mitosis detection
GlAS, Sirinukunwattana et al. (2015, 2017)	Colon	Gland segmentation
CoNSep, Graham, Vu, Raza, Azam, Tsang, Kwak & Rajpoot (2019)	Colon	Nuclei segmentation, classi cation
NCT-CRC-HE-100k, Kather et al. (2018)	Colon	Tissue classi cation
CRAG, Awan et al. (2017), Graham, Chen, Gamper, Dou, Heng, Snead, Tsang & Rajpoot (2019)	Colon	Gland segmentation
CRCHistoPhenotypes, Sirinukunwattana et al. (2016)	Colon	Nuclei detection and classi cation
Colorectal Cancer Grading Dataset, Awan et al. (2017), Shaban et al. (2020)	Colon	Colorectal Cancer Grading
Lizard, Graham et al. (2021)	Colon	Colonic Nuclear Instance Segmentation and Classi cation
PanNuke, Gamper et al. (2019, 2020)	Pan Cancer	Nuclei segmentation and classi cation
MoNuSeg, Kumar et al. (2017)	Pan Cancer	Nuclei segmentation
LYON, Swiderska-Chadaj et al. (2019)	Pan Cancer	Lymphocyte detection
ACDC-LungHP, Reddy & Rana (2021)	Lung Cancer	Detection and classi cation of lung cancer subtypes
MSI & MSS, Kather, Pearson, Halama, Jäger, Krause, Loosen, Marx, Boor, Tacke, Neumann et al. (2019)	Gastric Cancer	Classi cation of MSI and MSS gastrointestinal cancer subtypes

1.6 Challenges in this Study

Though there are many deep learning studies using digital pathology imaging that have predominantly been conducted for a wide range of carcinomas, there is very limited research on sarcoma digital pathology nor public sarcoma imaging datasets.

This study is mainly focusing on lipoma and liposarcoma histopathological image analysis using computational methods, including machine learning and deep learning techniques, and traditional image processing algorithms. Two pipelines are proposed to detect the tile-level features and the aberrant nuclei features.

Traditional computational diagnosis prediction models rely on manual feature extraction, which needs pathological expertise. This situation is changing in recent years with the development of deep learning techniques (LeCun et al. 2015, Campanella et al. 2019). Litjens et al. (2017) indicated although deep learning has been widely applied in image analysis and diagnostic tasks for computational pathology, there are additional challenges for digital pathology due to the nature of pathological data compared to other elds.

One main challenge is the lack of extensively annotated images as the ground truth. This is mainly because of the high cost of the digitisation of the physical glass slides and the time-consuming process of manual annotation for ground truth generation. Despite the challenge of manual annotation, there are some open-sourced datasets in digital pathology, and Table 1.2 displays a list of public digital pathology datasets for different computational tasks. Another feature of WSI is the large size of a single image: there are about 10 billion pixels within a single digital pathology image scanned at x40 magni cation (0.25 µm pixel[1]). The entire ImageNet dataset of 14 million images contains approximately nine trillion pixels (Deng et al. 2009), which roughly equals only 930 whole slide images. Moreover, due to the nature of fatty tissues, paucicellular and whitespaces in the images also present a computational challenge which we sought to address in this study.

The Aachen protocol for deep learning in histopathology is a recent manual with recommendations for hands-on pre-processing of whole slide images (Muti et al. 2020). This protocol describes all steps from processing the clinical data of the patient through to a set of image tiles that are ready to use for downstream deep learning analyses, including metadata preparation, image preparation, annotating the ROIs, tile generation, training and inference. During the image preparation process, in general, whole slide images exhibit similar types of artefacts

and noise as other digital microscopy images which unless removed with appropriate pre-processing can affect downstream analysis and interpretation. Davidson & Abramowitz (2014) showed a comprehensive overview of noise and artefacts in digital microscopy and suggested methods to correct them.

1.6.1 Large Whole Slide Image Loading

The nature and structure of whole slide images are complex. There are four sequential steps for whole slide image digitisation, they are image scanning, storage, editing, and display of images. The scanners acquire images on line-scanning or tile-scanning bases, the multiple tiles of lines are then digitally assembled to generate a digital image of the entire slide (Zarella et al. 2019). Histopathological slides are usually scanned using professional equipment under x40 magni cation for digitisation. The le size of a single digital whole slide image is usually very large and over 2GB, and therefore, the large image loading issue becomes a challenge for most of the programming languages. A computational package OpenSlide, developed by Goode et al. (2013), is widely used for the quick loading of large pathological slides, including ndpi, svs and BigTIFF les generated respectively by Hammamatsu, Leica, Phillips (BigTIFF) machines. OpenSlide is compatible with Java and Python and is the current mainstream technique in large slide loading. Tools such as QuPath (Bankhead et al. 2017) and CellProfiler™ (Lamprecht et al. 2007) take advantage of the exibility of OpenSlide in handling such large and complex whole slide images. I have used the OpenSlide package for all the raw data loading throughout my work.

1.6.2 Whole Slide Images Scanning Quality Control

Slide quality issue is another challenge in digital pathology. Artefacts and batch effects can unintentionally be introduced during both slide preparation as well as the digitisation processes. Quality issues include variations in colour, tissue folding, dust and bubbles on the slide, out-of-focus and blurriness and other scanning issues such as variations in contrast and hue and stripes produced by poorly calibrated optics.

Manual review of the digital slides is time-consuming and laborious and may not be consistently performed. Automated quality control software has been developed for this purpose. For example, HistoQC, developed by Janowczyk et al. (2019, 2020), is a tool for rapidly performing quality control to not only identify and delineate artefacts but also discover cohort level outliers. A combination of image metrics, such as colour histograms, brightness, and contrast; features such as edge detectors; and supervised classi ers such as pen detection are integrated into this tool. These regions and metrics are presented interactively to the end-users for the digital slide's quality control purpose.

1.6.3 Colour Variation, Deconvolution and Normalisation

Back et al. (2016) stated that there are four main sources of colour variation in pathological slides: experimental factors, chemical factors, machine factors and timing factors. Experimental variability encompasses the variability among people and operational processes producing the physical slides that can affect the staining colour, this is also known as operational factors. Chemical variability is the variability of the staining reagents, for example, reagents produced from different companies could sometimes lead to staining variation. Additionally, different scanning machines could produce colour variations in the digitised whole slide image, and nally, timing variability represents the colour fading or bleaching over prolonged periods of storage, particularly in suboptimal conditions such as exposure to sunlight. Van Eycke et al. (2017) indicated that colour variation among samples may affect some of the analysis tasks in digital pathology, and colour normalisation methods should be employed to reduce colour variation.

There exists a myriad of colour normalisation approaches speci cally designed for H&E stained images, and the solutions can be categorised into three main groups: histogram matching, colour transfer and spectral matching (Li & Plataniotis 2015). Histogram matching and colour transfer methods rely on histogram or values analysis for each of the red, blue and green colour channels (or hue, saturation and lightness channels in HSV/HSL colour space) (Kothari et al. 2011, Tabesh et al. 2007, Reinhard et al. 2001). The main limitation of these two groups of meth-

ods is they cannot differentiate the origin of colour variation nor normalise the stain variation properly. Spectral matching methods outperform other groups, as the aim of spectral matching is to remove the stain variation which is the main concern of colour variation in digital pathological image analysis tasks. Ruifrok et al. (2001) proposed the rst spectral matching algorithm. Based on Lambert-Beers law below, Ruifrok's algorithm takes advantage of colour deconvolution to estimate the contribution values for different stains respectively to red, green and blue channels.

$$I = I_0 \quad 10^{\ A_c} \tag{1.1}$$

$$OD = \quad \log_{10}(I = I_0) = A_c \tag{1.2}$$

with I_0 is the intensity of light entering the specimen, I is the intensity of colour that is detected, and A stands for the amount of stains where c is the absorption factors matrix of the stains. The intensities I_R, I_G, and I_B are obtained by the camera for each pixel in the RGB colour space. However the intensities depend on the concentration of stain in a non-linear way, therefore the intensity values of the image cannot directly be used for measurement of the stains. Alternatively, the optical density (OD) for each channel is linear, and can therefore be used to measure the contribution of multiple stains in a speci c colour space. To be speci c, the OD gives an idea of how much dye is within the sample, whereas the actual pixel intensity value doesn't give anything physical since it's a logarithm relationship, for example, if an OD is twice the quantity from one sample to the other, then we can say that the sample has twice the amount of dye from the other, as the ODs are additive when it comes to adding amounts of dye together. As the original data is a three-channel RGB image, therefore for digital histopathological slides, the components of stains can be deconvoluted into hematoxylin, eosin, and a third channel. For H&E images, the third channel is regarded as the complement channel which can be the residual of colour. For H&E-DAB images, the third channel can be the diaminobenzidine (DAB) colour. As a result, the colour deconvolution of the H&E-stained image can be represented as:

$$
\begin{aligned}
OD_{R;G;B} = {} & OD_{Hematoxylin}(R;G;B) \quad a_{Hematoxylin} + \\
& OD_{Eosin}(R;G;B) \quad a_{Eosin} + \\
& OD_{Complement}(R;G;B) \quad a_{Complement}
\end{aligned}
\tag{1.3}
$$

$a_{Hematoxylin}$, a_{Eosin} and $a_{Complement}$ here are regarded as the optical density contribution rate for hematoxylin, eosin and the complement to original H&E image in RGB colour space. The staining depth for hematoxylin and eosin can be measured by computing $a_{Hematoxylin}$ and a_{Eosin}, which quanti es the staining variation within inter-batch samples or the colour variation from slide to slide. Colour deconvolution enables us to map the physical quantities (the RGB values) onto the real-world quantities (the staining depth). For example, we can operate on the hematoxylin channel extracted using colour deconvolution technique for those nucleus-related computational tasks as nuclei are mainly stained by hematoxylin in H&E images according to domain knowledge.

In Python, skimage.color.rgb2hed() and skimage.color.hed2rgb() can be used for RGB and Haematoxylin-Eosin-DAB (HED) colour space conversion (van der Walt 2021). Fig 1.4 shows a sample output of applying skimage.color.rgb2hed() to pathological tile.

Colour deconvolution is a method to obtain stain concentration values when the stain matrix is given, and the stain matrix describes how the colour is affected by the stain concentration. This could sometimes be dif cult for fatty tissues due to their paucicellular nature that only a very small amount of hematoxylin and eosin are stained on the slides. Another generic limitation of Ruifork's algorithm is this algorithm is based on a mathematical calculation which may sometimes result in negative contribution values. Negative values lead to imperfect normalisation, as a negative amount of light cannot be absorbed and there does not exist a negative amount of stain. Rabinovich et al. (2004) presented an improved algorithm based on non-negative matrix factorisation (NMF) to force the staining contributions to be non-negative to solve the negative value problem. Macenko et al. (2009) proposed a fully automatic process which can be executed before colour deconvolution by

Figure 1.4: A sample output of `skimage.color.rgb2hed()`.

calculating the stain vectors plane and adjusting the stain vectors from the estimated ones to overcome this kind of limitation. Colour normalisation methods are still imperfect at this stage as all the above-mentioned methods cannot be used under the condition of three or more stains within one slide. Alternatively, Kothari et al. (2011) sampled and extracted the unique colours in an image and created a colour map for the normalisation instead of counting all pixels within the whole image, and this pipeline is applicable for multi-stains.

Unlike all other colour normalisation methods before 2015, Li & Plataniotis (2015) used a combination of illuminant and spectral normalisation. Li's algorithm outperformed other methods such as (Rabinovich et al. 2004, Macenko et al.

2009, Kothari et al. 2011). Li & Plataniotis (2015) also suggested this model can be useful for histopathology image analysis tasks such as colour-based histological component detection and segmentation. Bejnordi et al. (2016) developed a whole-slide image colour standardiser (WSICS) method which utilises colour and spatial information to classify the image pixels into different stain components under hue-saturation-density (HSD) colour space. This approach is designed for H&E images but can also be adopted in other kinds of histological staining images such as immunohistochemistry (IHC). Bejnordi et al., using three quantitative empirical evaluation methods, claimed their WSICS pipeline is the best in digital slide colour normalisation compared with all published algorithms at that time. Anand et al. (2019) proposed a fast GPU-enabled colour normalisation work ow based on Vahadane et al. (2015) to make this structure preserving colour normalisation (SPCN) method more robust as well as being able to apply on gigapixel whole slide images. Ramakrishnan et al. (2010) also published their codebase for this task.

There are some good implementations for a combination of different colour normalisation methods written in Python. Package StainTool, developed by By eld (2018), is one of them, and is also applied in this study (Fig G.1 in Appendix G). There also exist some debates in recent studies on whether colour normalisation is actually needed in digital pathology image analysis tasks. Bianconi et al. (2020) made a statement that colour pre-processing resulted in a noticeable reduction of the accuracy in most cases, especially when coupled with image descriptors that rely heavily on the colour of the image. This statement is also aligned with studies of (Gadermayr et al. 2017, Hameed et al. 2020, Tellez et al. 2019) where all the overall top performances in their tasks were achieved without the use of colour normalisation techniques.

I employed all three colour normalisation algorithms in our own dataset for the tile-based and the nucleus-based tasks. I decided not to apply colour normalisation on the tile-based work ow, and to apply this technique on the nucleus-based task based on the validation results of each of the tasks. More details including why to apply or not to apply the colour normalisation technique in each of the tasks will be

discussed in Section 3.6 and Section 6.3.3.

Chapter 2

The Data

2.1 Chapter Overview

In this chapter, I mainly describe data acquisition, image digitisation and data curation. Two independent cohorts of whole slide images of fatty tumours were collected individually from Royal National Orthopaedic Hospital (RNOH), United Kingdom and Sahlgrenska University Hospital (SUH), Sweden. A total of 206 slides from RNOH are the internal ones for prototyping, training, validation and evaluation. Another 402 slides from SUH are external and for generalisability and scalability, but only 153 top-quality slides out of 402 are selected for the external evaluation. In total, 359 slides are employed in this study for both the tile-based and the nucleus-based tasks. The le size for each of the slides is large, approximately 2-3GB, and there could be 8-10 billion pixels within a single slide scanned at x40 magni cation (with 0.25 µm pixel^{1}).

2.2 Internal Whole Slide Images

206 cases collected from RNOH were scanned and digitised at UCL Cancer Institute pathology research department using three different scanners from the same manufacturer, namely the Hamamatsu Nanozoomer C9600 series, Hamamatsu Nanozoomer S210 and Hamamatsu Nanozoomer S360. As there exist scanning quality issues for the rst two machines (stripes for C9600 and out-of-focus for the paucicellular fatty tumours for S210, which I will describe in the following subsection), the S360 machine was nally employed for the digitisation of slides.

2.2.1 Quality Issue and Quality Control

Figure 2.1: Binarised whole slide image sample (internal sample ID: 021) showing scanning stripes.

Fig 2.1 is an example of a whole slide image scanned and digitised using Hamamatsu Nanozoomer C9600 series and it can be seen that there is an issue with stripes within this image after binarisation. The stripe issue is common among slides scanned with the C9600 machine. In order to investigate where the stripes come from, I digitised the same samples with other scanning devices. I found the stripes disappeared when scanning with Hamamatsu Nanozoomer S210 or Hamamatsu Nanozoomer S360 machines. This scanning quality issue was caused by the scanner.

I applied HistoQC to the Hamamatsu Nanozoomer S210 and S360 scanned slides and observed the S210 scannings consistently contain focal out-of-focus quality issues. The quality control results of the S360 machine show consistency. Fig 2.2 and 2.3 display the quality control results respectively for these two machines, from which we can observe the three main clusters of all the slides: the conventional cases, the cases with pen marker, and the cases with various histology variations.

Finally, I decided to use the Hamamatsu Nanozoomer S360 scanning copies in our studies after manual examination of the original slides as well as their underlying HistoQC quality control visualisation. A glass slide surface cleaning process was executed to remove all the pen markers before the re-scanning process with the S360 machine. HistoQC results shown in Fig 2.3 indicate the quality consistency of all the internal slides, with two main clusters of conventional slides and the slides with histology variation. Furthermore, I applied hierarchical clustering to explore the structure of all the internal slides, and the hierarchical result aligns with the slide clusters based on the manual labelled on-slide features. Fig 2.4 is an example of internal digital slide.

2.2.2 Image Storing

All the whole slide images are stored under ndpi format, and are anonymised using AnonymizeSlide tool developed by Gilbert (2015). Apart from the digital images and their underlying diagnostic labels, we hold other clinical information for each of the samples, including (1) age at diagnosis, (2) patient gender, (3) whether the tumour arose super cially or deeply, (4) the size of the tumour, (5) whether the uorescence in situ hybridisation (FISH) genetic test has been carried out for MDM2 gene, and if yes the FISH result, (6) the resection status, (7) whether there are other histology features shown on the slides. All these details are stored in a spreadsheet as metadata. Table 2.1 is an anonymised example of the metadata of internal slides, a full metadata is presented in Appendix A, and Appendix B shows the mapping table of internal and clinical IDs.

2.2.3 Internal Slides Overview

163 slides out of 206 slides were selected to rescan with the S360 machine. Some slides, based on the review of images and metadata, were removed from this study because they are not conventional and contain special histology variations which make them out-of-scope for this study. Slides that have unclear ground truth in terms of the whole slide level diagnostic label were also removed from this study. The details of these slides are listed below:

(a) The conventional cluster.

(b) The pen-marker cluster.

(c) The histology variation cluster.

Figure 2.2: HistoQC visualisation and hierarchical analysis on the internal dataset for quality control purposes (Hamamatsu Nanozoomer S210).

(a) HistoQC visualisation on our internal lipoma and atypical lipomatous tumour cases.

(b) The hierarchical clustering of slides based on HistoQC features.

Figure 2.3: HistoQC visualisation and hierarchical analysis on the internal dataset for quality control purposes (Hamamatsu Nanozoomer S360).

Figure 2.4: An example of an atypical lipomatous tumour digital slide 206.

Table 2.1: An examples of the metadata of whole slide images. The samples were anonymised, and so the clinical ID was removed in this table.

Internal ID	Diagnosis	FISH	Site	Super cial/Deep	Max Size
001	Lipoma	No	Pelvis	Deep	80mm
002	Lipoma	No	Abdomen	Super cial	50mm
003	Lipoma	No	Abdomen	Deep	80mm
004	ALT	No	Thigh	Deep	180mm
005	ALT	No	Thigh	Deep	190mm
006	ALT	No	Thigh	Deep	280mm
007	ALT	No	Thigh	Deep	250mm
008	ALT	No	Thigh	Deep	70mm
009	ALT	No	Arm	Deep	70mm
010	ALT	No	Arm	Deep	70mm

1. cases 152, 167, 172, 183 and 184 contain only the necrosis or are dedifferenciate liposarcoma,

2. cases 22, 24 and 192 are spindle cell lipomas,

3. case 150 has unclear ground truth from a FISH point of view: if we are using the clinical diagnostic criteria such as counting the number of cells, then this case would score as negative; but when these cells have an unusual pattern that whilst they may not be obviously MDM2 amplified, they do have gains of MDM2 not quite yet at the level of amplification.

First, the MDM2 FISH test result is used as the ground truth diagnostic label at the whole slide level. I then split the internal slides into the training set of 87 slides and the testing set of 76 slides. 59 conventional slides out of 87 of the training set were further selected in this study for the training of the lipoma and ALT areas on the tile-level, and these 59 slides were also used for the training of slide-level aggregation. The rest of the training slides were employed for the training of histology variation areas, including muscle, fat necrosis, calcification and inflammation, on the tile-level. 64 out of 76 of the testing set were further selected for testing, evaluation and for the distinction of the lipomas and the ALTs on the slide-level. The testing set was fully unseen during the entire training process. Table 2.2 is a summary of the number of internal slides as well as a description of the slide selection process.

Table 2.3 and Table 2.4 respectively show the internal training and testing set statistics by diagnostic labels.

2.2.4 Tile the Whole Slide Images

The first step of any tile-based whole slide image analysis task is to tile the slides up into small patches. Fig 2.5 shows an example of a slide before and after tiling. There exists a tool, PyHIST, that can be used for this tiling process (Muñoz-Aguirre et al. 2020), but I have developed my own scripts to allow flexibility since we also need to be able to tile regions of interest (ROIs) from digital slides (Section 3.3.2).

Table 2.2: The internal slides count and slide selection process.

Slides Count	Description	Issue Identified and Selection Process
206	A mix of Hamamatsu C9600, Hamamatsu S210, and Hamamatsu S360 scanned slides.	Scanning quality issues are identified for different scanning machines (stripes and out-of-focus for Hamamatsu C9600, out-of-focus for Hamamatsu S210). This initial dataset also consists of out-of-scope slides with labels of neither lipoma nor liposarcoma, and these slides should be removed.
163	After the slides with labels of neither lipoma nor liposarcoma are removed from the batch, all other slides are rescanned with Hamamatsu S360 (Section 2.2.1, Table 2.3) with high quality.	After reviewing the slides, some of the slides are identified as out-of-scope with special variants (e.g., spindle-cell lipoma) or histology variation, though they have the diagnostic label of either lipoma or liposarcoma. The scope of this project is to distinguish the conventional lipomas and atypical lipomatous tumours, and so slides with special variants need to be removed.
151	87 training WSIs and 64 testing WSIs scanned by Hamamatsu S360 are the final in-house training and testing dataset. Within the training set, 59 out of 87 are used for both the slide-level and tile-level training, the rest 28 slides are only used for the tile-level training of histology variations.	

Table 2.3: Internal training set statistics by diagnostic labels.

Diagnostic Label	Number of Slides	Slide Internal IDs
Lipoma	26	001, 002, 003, 018, 020, 042, 043, 044, 047, 051, 053, 055, 063, 064, 108, 109, 111, 112, 113, 115, 116, 117, 118, 119, 120, 157
ALT	33	004, 005, 006, 008, 009, 010, 019, 025, 028, 032, 034, 035, 041, 069, 070, 077, 079, 083, 085, 086, 090, 092, 093, 096, 097, 098, 099, 100, 194, 197, 199, 200, 206

Table 2.4: Internal testing set statistics by diagnostic labels.

Diagnostic Label	Number of Slides	Slide Internal IDs
Lipoma	39	021, 031, 033, 048, 049, 056, 103, 114, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 136, 137, 138, 139, 151, 159, 173, 175, 181, 182, 185, 186, 187, 188, 193, 195, 196
ALT	25	023, 036, 038, 040, 072, 074, 076, 080, 084, 088, 094, 101, 106, 107, 161, 179, 180, 189, 191, 198, 201, 202, 203, 204, 205

(a) An example of whole slide image.

(b) An example of whole slide image after tiling.

Figure 2.5: An example of slide before and after tiling.

2.2.5 Tile-Level Data Augmentation

The philosophy of training a machine learning model is tuning its parameters such that it can map a particular input to some output. State-of-the-art neural networks typically have hundreds of millions of parameters, and if there are a lot of parameters in a model, we would need to train the model with a proportional amount of input data in the training loop. Krizhevsky et al. (2012) indicated that there are lots of popular methods for data augmentation in image-based computational tasks, including basic transformations of ip (horizontally or vertically), rotation, crop and translations. Gaussian noise is also widely used to remove those high-frequency features and patterns in the images that lead to over- tting. In digital pathology, stain colour augmentation is also used as one of the data augmentation methods (Janowczyk 2018, Tellez et al. 2019).

2.3 External Whole Slide Images

Similar to internal slides, a mix of 402 lipoma and atypical lipomatous tumour digital slides were acquired from Sahlgrenska University Hospital, Sweden. These external slides were employed mainly to evaluate the generalisability and scalability of our proposed work ows. There are 80 lipoma cases from 20 patients and 322 atypical lipomatous tumour cases from 33 patients in the external dataset, Table 2.5 shows the patient list for the external dataset and Appendix C shows the full list of the external dataset. HistoQC is also applied to all the external slides, and Table 2.6 shows the top scanning quality slides for each patient.

Table 2.5: External dataset: the patient list.

Diagnostic Label	Number of Patients	Patient ID
ALT	33	1, 2, 3, 4, 5, 6, 7, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 35, 36
Lipoma	20	1, 2, 3, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36

In order to generate a fair slide-level external testing set, I used only one slide per patient of the external dataset. Three batches of slide-level testing sets were selected based on Table 2.6 as the external and independent sets for the evaluation of machine learning models. Table 2.7, Table 2.8 and Table 2.9 show more details.

Table 2.6: External dataset: the top scanning quality slides for each of the patients.

Patient ID	Number of Slides	Top Quality Slides	Patient ID	Number of Slides	Top Quality Slides
ALT.1	8	1, 2, 4	ALT.30	15	3, 7, 8
ALT.2	12	3, 4, 5	ALT.31	8	3, 4, 7
ALT.3	12	4, 9, 10	ALT.32	12	1, 2, 5
ALT.4	16	1, 2, 5	ALT.33	6	2, 3, 4
ALT.5	10	1, 2, 8	ALT.35	10	2, 5, 10
ALT.6	8	2, 5, 8	ALT.36	11	11, 15, 19
ALT.7	6	2, 5, 6	L.1	5	1, 3, 4
ALT.10	10	1, 2, 5	L.2	6	2, 4, 6
ALT.11	5	2, 3, 4	L.3	2	1, 2
ALT.12	3	1, 2, 3	L.20	3	1, 2, 3
ALT.13	9	3, 4, 6	L.21	3	1, 2, 3
ALT.14	8	1, 3, 6	L.22	3	1, 2, 3
ALT.15	6	2, 3, 6	L.23	4	1, 3, 4
ALT.16	8	1, 6, 7	L.24	4	1, 3, 4
ALT.17	9	2, 6, 8	L.25	5	1, 4, 5
ALT.18	14	3, 7, 8	L.26	5	1, 2, 4
ALT.19	8	1, 2, 5	L.27	2	1, 2
ALT.20	10	6, 8, 9	L.28	1	1
ALT.21	12	3, 4, 7	L.29	8	2, 3, 8
ALT.22	11	1, 2, 10	L.30	7	3, 5, 7
ALT.23	3	1, 2, 3	L.31	5	1, 2, 5
ALT.24	8	2, 7, 8	L.32	4	2, 3, 4
ALT.25	19	9, 10, 18	L.33	5	1, 2, 4
ALT.26	20	6, 17, 18	L.34	3	1, 2, 3
ALT.27	8	3, 4, 6	L.35	1	1
ALT.28	10	6, 7, 8	L.36	4	1, 2, 3
ALT.29	7	2, 6, 7			

Table 2.7: External dataset: slide-level batch 1 (53 slides in total).

Diagnostic Label	Number of Slides	WSI ID
ALT	33	1.1, 2.3, 3.4, 4.1, 5.1, 6.2, 7.2, 10.1, 11.2, 12.1, 13.3, 14.1, 15.2, 16.1, 17.2, 18.3, 19.1, 20.6, 21.3, 22.1, 23.1, 24.2, 25.9, 26.6, 27.3, 28.6, 29.2, 30.3, 31.3, 32.1, 33.2, 35.2, 36.11
Lipoma	20	1.1, 2.2, 3.1, 20.1, 21.1, 22.1, 23.1, 24.1, 25.1, 26.1, 27.1, 28.1, 29.2, 30.3, 31.1, 32.2, 33.1, 34.1, 35.1, 36.1

Table 2.8: External dataset: slide-level batch 2 (51 slides in total).

Diagnostic Label	Number of Slides	WSI ID
ALT	33	1.2, 2.4, 3.9, 4.2, 5.2, 6.5, 7.5, 10.2, 11.3, 12.2, 13.4, 14.3, 15.3, 16.6, 17.6, 18.7, 19.2, 20.8, 21.4, 22.2, 23.2, 24.7, 25.10, 26.17, 27.4, 28.7, 29.6, 30.7, 31.4, 32.2, 33.3, 35.5, 36.15
Lipoma	18	1.3, 2.4, 3.2, 20.2, 21.2, 22.2, 23.3, 24.4, 25.4, 26.2, 27.2, 29.3, 30.5, 31.2, 32.3, 33.2, 34.2, 36.2

Table 2.9: External dataset: slide-level batch 3 (49 slides in total).

Diagnostic Label	Number of Slides	WSI ID
ALT	33	1.4, 2.5, 3.10, 4.5, 5.8, 6.8, 7.6, 10.5, 11.4, 12.3, 13.6, 14.6, 15.6, 16.7, 17.8, 18.8, 19.5, 20.9, 21.7, 22.10, 23.3, 24.8, 25.18, 26.18, 27.6, 28.8, 29.7, 30.8, 31.7, 32.5, 33.4, 35.10, 36.19
Lipoma	16	1.4, 2.6, 20.3, 21.3, 22.3, 23.4, 24.4, 25.5, 26.4, 29.8, 30.7, 31.5, 32.4, 33.4, 34.3, 36.3

Chapter 3

A Tile-Based Lipoma and Atypical Lipomatous Tumour Classi er

3.1 Chapter Overview

In this chapter, I rst examine state-of-the-art of the deep learning application in digital pathology for diagnosis bene ts. I then introduce and display the tile-based work ow that is proposed in this study for the distinction between benign lipomas and malignant atypical lipomatous tumours, including the methodology, experiments and results.

3.2 Introduction and Related Work

3.2.1 Tile-Based Method for the Bene ts of Diagnosis

The development of whole slide scanners in late 1990 made the digitisation of histopathological images easier. The scanners can scan slides at the microscopic resolution which opens up opportunities for computer scientists to develop image-based applications for the bene t of diagnosis and prognosis.

Yuan et al. (2012) indicated the features extracted from histopathological slides can be complementary to the genomic data. In the same year, Krizhevsky et al. (2012) showed that convolutional neural networks (CNNs) can be employed as an image classi er for 1000 classes, winning the ImageNet competition that year. In the same year, CNNs were proved to outperform other methods in mitotic cell de-

tection in histopathology images (Ciresan et al. 2012). CNN-based deep learning methods have consistently had a better performance compared to other traditional image-based methods in digital pathology since 2013. On the one hand, CNNs are able to learn features directly from the raw data without feature engineering assisted by pathologists. But on the other hand, CNNs are like black-boxes, and pathologists still play an important role in the initial annotation and manual validation processes.

The rst challenge for whole slide image analysis is that images are very large. As explained in Section 2, a single whole slide image scanned under x40 magni ca-tion can be approximately 2-3GB with in total of 8-10 billion pixels. As a result, the images need to be broken down into thousands of tiles so that the CNN can process them. There are three major types of CNN models that are commonly used in digital pathology image analysis and are based on the nature of their computational tasks namely classi cation, object detection and segmentation. The rst type of model does tile-level classi cation. The second type of model focuses on predicting the position of objects, an object in digital pathology can be a cell or nucleus. And the third type is to generate pixel-level semantic segmentation results either for an area or an object. This chapter focuses on the rst category of tasks, which is applying the tile-based method for slide-level classi cation. Table 3.1 shows some highly cited tile-based studies on digital pathology in chronological order.

There are only two published studies on the use of digital image analysis on soft tissue sarcoma. The rst study using the computational method based on digi-tised pathology slides is (Foersch et al. 2021). A tile-based DenseNet121 CNN was trained in this study to classify ve subtypes of sarcoma, they are: dedifferentiated liposarcoma, leiomyosarcoma, myxo brosarcoma, synovial sarcoma and undifferentiated pleomorphic sarcoma. Their work ow achieved an accuracy of 79.9% in the classi cation of these ve subtypes of soft tissue sarcomas. Separately, the au-thors trained another deep neural network which showed the ability to predict the prognosis of leiomyosarcoma. The other study indicated that tile-based deep learn-ing models outperformed the pathologists with a signi cant improvement of accu-racy up to 97% compared to average pathologists of 69.7% for the classi cation of

Table 3.1: Selected tile-based studies on digital pathology (chronological order).

Authours	Cancer Type	Application
Cruz-Roa et al. (2014)	Breast	Invasive breat cancer detection
Sirinukunwattana et al. (2016)	Colon	Nuclei classi cation
Wang et al. (2016)	Breast	Breast cancer detection
Litjens et al. (2016)	Prostate, Breast	Prostate and breast cancer detection
Sharma et al. (2017)	Gastritis	Necrosis detection
Vandenberghe et al. (2017)	Breast	HER2 status scoring
Mobadersany Mobadersany et al. (2018)	Gliomas	Survival rate prediction
Arvaniti et al. (2018)	Prostate	Prostate grading
Campanella et al. (2019)	Prostate, basal cell carcinoma, breast	Slide-level classi cation
Kather, Krisam, Charoentong, Luedde, Herpel, Weis, Gaiser, Marx, Valous, Ferber et al. (2019)	Colon	Survival prediction
Wei et al. (2019)	Lung	Subtyping of lung cancer
Nagpal et al. (2019)	Prostate	Gleason score prediction
Zhang et al. (2019)	Bladder	Bladder cancer diagnosis
Sudharshan et al. (2019)	Breast	Breast cancer classi cation
Xie et al. (2019)	Breast	Breast cancer diagnosis
Aresta et al. (2019)	Breast	Breast cancer classi cation
Vu et al. (2019)	Lung, squamous cell carcinoma, glioblastoma, glioma	Tumour detection, nucleus segmentation and classi cation
Wang et al. (2019)	Lung	Lung cancer classi cation
Iizuka et al. (2020)	Gastritis, Colon	Gastritis and colonic tumour classi cation
Fu, Jung, Torne, Gonzalez, Vöhringer, Shmatko, Yates, Jimenez-Linan, Moore & Gerstung (2020)	Pan-cancer	Tumour subtyping and grading
Kather et al. (2020)	Pan-cancer	Molecular change prediction
Echle et al. (2020)	Colon	Colorectal cancer detection
Lu et al. (2021)	Renal, Lung	Tumour classi cation

ve main myxoid soft tissues (Yeung & Cheng 2022). Through an exhaustive literature search, I have not identi ed a study on the classi cation of lipoma and ALT using deep learning on digital pathology images. However, a recent study (Yang et al. 2022) used deep learning for the classi cation of lipomas and ALTs based on CT and MR images.

3.2.2 Tile-Aggregation Methods

Counting-based slide-level classi cation using the tile-level CNN results is not robust, as misclassi cation of the tile-level CNN could result in a different class prediction on the slide level. One way to correct this issue is to train a slide aggregation model on top of the tile-based CNN outputs. There are lots of existing slide-level tile aggregation approaches. Hou et al. (2016) trained a logistic regression model based on the number of tiles on the class basis as the slide-level classi er. Wang et al. (2016) extracted geometrical features from the slide-level visualisation result of the heap map based on the tile-CNN and then trained a random forest to classify the slides. Campanella et al. (2019) used the output from the feature extractor stage of their primary model and select the N most suspicious (i.e., hottest) tiles, and then they pass all these tiles sequentially into a recurrent neural network (RNN) for the slide-level classi cation.

3.2.3 Computational Tools in Digital Pathology

3.2.3.1 Python scienti c tools

Although strictly not only used for digital pathology, I made extensive use of python scienti c tools such as *scikit-image* for image processing, *scikit-learn* for machine learning, *Seaborn* for data visualisation, and *PyTorch* for training deep neural networks. A large number of bespoke python scripts were written to carry out this research. Table 3.2 shows the list of packages that were employed for this study. The well-documented GitHub repository for this study can be found at https://gitfront.io/r/user-3345834/arKL97mKRVQz/liposarcoma/

Table 3.2: Development environment and package requirements in terms of Python.

Package	Version Number
openslide-python	1.1.1
scikit-image	0.17.2
scikit-learn	0.23.1
torch	1.7.1 (cuda 10.1)
torchvision	0.8.2 (cuda 10.1)
seaborn	0.11.2
large-image	0.3.0
libtiff	0.4.2
matplotlib	3.3.0
numpy	1.19.1
pandas	1.1.0
scipy	1.5.2
tiff le	2020.9.3
tqdm	4.48.0

3.2.3.2 QuPath

There are some existing open-sourced digital pathology tools for slide preprocessing and analytics for the purpose of research such as QuPath. QuPath, developed by Bankhead et al. (2017, 2022), is an open-source software for bioimage analysis and is widely used in digital pathology. QuPathis integrated with basic image analysis features such as viewing and annotating whole slide images, colour deconvolution and analysis, and work ows for bright eld or uorescence image analysis and so on. QuPathis also been developed iteratively with new algorithms for common and computational tasks, including nuclei segmentation using the watershed method, tissue microarray de-arraying, and some interactive machine learning for both object and pixel classi cation and customisation. QuPathis also integrated with ImageJ In this study,QuPathis employed for annotation and regions of interest (ROIs) generation from the original slides, and for daily cross-platform slide viewing.

3.2.3.3 HistoQC

HistoQC, a quality control tool for digital pathology slides, consists of a pipeline of modules sequentially applied to an image according to (Janowczyk et al. 2019,

2020). HistoQC integrates traditional image processing methods together with some machine learning techniques for the detection of digitisation (i.e., scanning) quality issues, including dark, bright or fuzzy areas within slides, artefact (such as bubbles, pen marker and so on), and for the generation of tissue mask using Otsu's thresholding methods and so on. In this study, HistoQC is employed as the automatic quality screening tool as a supplementary to manual veri cation of the internal and external slides.

3.2.3.4 Other useful tools

ImageJ is a public domain Java image processing program inspired by NIH Image for the Macintosh, which can load, edit, analyse, process images and conduct some basic calculation and statistical tasks such as density analysis, sharpening, smoothing, edge detection and so on according to (Abramoff et al. 2004, Schneider et al. 2012, Abramoff et al. n.d., Baecker 2012, Della Mea et al. 2017). PyHIST is designed for tissue area detection using thresholding methods and tile generation for further computational tasks. PyHIST failed in this study because of the paucicellular nature of fatty tissue. The details will be discussed in Section 3.4.2, and Fig 3.7 and 3.16 show an example and a comparison.

3.3 Methodology

3.3.1 The Tile-Based Work ow: an Overview

I trained a tile-based CNN to automatically distinguish the malignant atypical lipomatous tumour from the benign lipoma, the entire process is summarised in Fig 3.1. The rst step of the work ow is to have all the original slides annotated on a regions-of-interest (ROI) level for classes of (1) non-informative areas, this includes background and artefact, (2) conventional lipoma areas, (3) conventional atypical lipomatous tumour areas, (4) muscle areas, (5) calci cation areas, (6) in ammation areas, (7) fat necrosis areas. I tiled ROIs up into 224 224 images and used them to train an Inception v3 deep CNN network to create a tile-level classi er. I then sent all the training slides into this classi er for the calculation of tile-level prediction scores. I then wrote bespoke python scripts to visualise the results by mapping

Figure 3.1: The proposed tile-based work ow for the prediction of lipoma and atypical lipomatous tumour.

the tile-level classi cation results onto the whole slide image space, plots including class-prediction map, heat map and contour map were employed at this stage. I employed Gaussian smoothing as a post-processing operation before statistically analysing the training results on the slide-level. Finally, a series of the slide-level features were extracted for tile aggregation on the slide-level as well as to train a slide-level diagnostic label predictor.

Pathologists played an important role in this whole work ow. At the beginning, the ROIs were generated by taking advantage of the pathologist's (Dr Nischalan Pillay) annotations as the input. And although the slide-level features were extracted and selected using computational methods at the tile aggregation stage, the pathologist's viewpoints and comments on those selected features provided some veri cations in agreement. And nally, during the inference steps, the pathologist's review of slide prediction results also opened up some work ow optimisation ideas and further discussion.

3.3.2 Tiles Generation and Annotation Strategy

The original tiles were generated from annotated regions of interest in the training process in this study. The annotations were made through QuPath version 0.2.0, and Fig 3.2 shows an example of the annotations of an atypical lipomatous tumour case. As QuPath annotations can only be exported into their dedicated format of *.qpdata, a groovy script (Appendix D) is needed for annotation conversion from *.qpdata into *.jpg image and *.png mask pairs. This script was modi ed according to (Bankhead 2021), and can only be used for QuPath version greater than 0.2.0. Fig 3.3 shows some examples of the informative tiles generated. And Fig 3.4 shows some non-informative tile examples including background, pen-marker, dust, glue and bubbles.

The ROIs were tiled up under x5 and x10 magni cation, and I nally decided to train the tile-based deep convolutional neural network under x5 magni cation (discussed in the methodology optimisation Section 3.4.1, Table 3.6, Fig 3.11, Fig 3.10). x5 tiles contain more structural information compared to x10s within each of the tile windows for the paucicellular fatty tissues. Fig 3.5 shows a compar-

Figure 3.2: An example of the annotations of sample 035. Grey annotations stand for non-informative areas, red annotations stand for atypical lipomatous tumour areas, and blue annotations stand for muscle areas.

Figure 3.3: Some examples of the informative tiles.

Figure 3.4: Some examples of the non-informative tiles.

ison of x5 and x10 tiles.

Figure 3.5: A comparison of x5 and x10 magni cations tiles of lipoma, atypical lipomatous tumour and fat necrosis.

Table 3.3: Training set tile-level statistics.

Training Class	Slides Count	ROIs Count	Tiles (x10) Count	Tiles (x5) Count
Lipoma	27	289	26,312	6,335
ALT	39	521	27,730	6,697
Background	87	347	52,511	12,603
Muscle	9	46	1,456	358
Calci cation	5	13	174	42
In ammation	8	40	112	23
Fat necrosis	16	66	11,160	2,711
Total	87	1,322	119,455	28,769

One slide can contain multiple classes of training ROIs and annotations.

Tiles were generated from ROIs annotated from training slides. 27 conventional lipoma slides and 39 atypical lipomatous tumour slides were annotated for the dominant classes of lipoma and ALT, and another 21 slides were annotated for the main histology variations in fatty tumours, including muscle, calci cation, in-

Figure 3.6: The tissue detection pipeline proposed by Bizzego et al. (2018). The identi - cation of the tissue bounding box is performed on the WSI thumbnail in three steps: (1) binarisation of the grayscale image using Otsu's algorithm, (2) binary dilation and lling the holes, and (3) biggest connected region detection.

ammation and fat necrosis. To be mentioned, there exist some histology variation annotations within some of the conventional lipoma and atypical lipomatous tumour slides, and all of the 87 training slides contain non-informative (i.e., the background class) ROIs. In summary, 7 classes were annotated for the tile-based part of work in this study: (1) non-informative areas, which include background and artefact, (2) conventional lipoma areas, (3) conventional atypical lipomatous tumour areas, (4) muscle areas, (5) calci cation areas, (6) in ammation areas, (7) fat necrosis areas, and in total, there were 28,769 tiles generated from 1,322 ROIs annotated from 87 whole slide images. Table 3.3 is the tile-level statistics for the training slides. The training tiles were then split into a real training set and a validation set by classes for the CNN training, with a splitting factor of 0.8.

3.3.3 The Tile-Level CNNs: the Evolution

A series of CNN models were trained in this study along with our exploration and understanding of the dataset. At the very beginning, I did not have any of the annotations of ROIs in slides, so I generated the tiles directly from the whole slide images and assumed the label of each of the tiles is the diagnostic label of its slide. I then trained a simplest binary classi er (Alex-a-binary) using AlexNet (Krizhevsky et al. 2012) structure with these rough training tiles. The rst observed issue was the background (i.e., non-tissue areas in the slides), that I should really not assign a tumour type class as the label of a background tile in that slide. I then tried to identify the tissue area by taking advantage of a tissue detection pipeline shown in Fig 3.6 (Bizzego et al. 2018), and to train a 3-class classi er (Alex-a-3) of lipoma, atypical lipomatous tumour and the background. Another observed issue was about

histology variations, which the CNN confused when the muscle was presented in a slide. The same as the background, I should not assign the diagnostic label to areas of histological variation. So I manually and roughly annotated some of the muscle areas and trained an initial 4-class classifier (Alex-a-4) of lipoma, ALT, background and muscle. The same logic was then applied to a broader range of histology variations of inflammation, calcification and fat necrosis, and at this stage, a 7-class classifier (Alex-a-7) was trained. All these experiments were explorational.

As discussed in Section 1.6.2 and 2.2, slides were re-scanned considering the scanning quality issues, such as stripes and focal out-of-focus, with the new Hamamatsu NanoZoomer S360 machine. Regions of interest of the 7 mentioned classes were generated from this batch of digital slides in a more systematic way using a pathologist's manual annotation plus verification before tiles are generated. See more details for the ROI and tile generation strategies in Section 3.3.2. One thing to be stressed is I annotated the background ROIs rather than applying the thresholding-based pipeline proposed by Bizzego et al. (2018) and shown in Fig 3.6, as this pipeline is not robust in our scenario of fatty tissue slides. Lipoma and atypical lipomatous tumour slides are paucicellular and contain lots of whitespaces (examples shown in Fig 1.2) which may sometimes result in low contrast between the tissue regions and the non-tissue background, especially for those conventional cases without much histology variation. Fig 3.7 shows a successful and a failed example for applying the normal tissue detection pipeline on our own dataset, which well illustrates the failure of a typical tissue detection pipeline on conventional fatty tissues.

So following the data cleaning and filtering processes, I then trained another 4-class (Alex-b-4) and 7-class (Alex-b-7) classifiers using this final version of tile-level training set. I also prepared an independent tile-level dataset together with a pathologist to roughly evaluate the two dominant classes of lipoma and atypical lipomatous tumour. This tile-level dataset contains 10,392 lipoma tiles and 11,793 atypical lipomatous tumour tiles (22,185 tiles in total), and was applied to all Alex-4, Alex-a-7, Alex-b-4, and Alex-b-7. As explained before, the training slides were

(a) Slide sample 171.

(b) Slide sample 171 after tissue detection pipeline.

(c) Slide sample 051.

(d) Slide sample 051 after tissue detection pipeline.

Figure 3.7: A success and a failed example for applying the normal tissue detection pipeline on our own fatty tissue dataset.

not re ned at the start, so there are a huge amount of non-informative tiles that were mislabeled to be either lipoma or ALT for the training of Alex-4 and Alex-7. I would expect them to have much lower accuracy when comparing the tile-level performance with other models. Table 3.4 shows a tile-level evaluation accuracy for all the Alex-CNNs, and Fig 3.8 shows the tile-level ROC for the Alex-CNNs for the malignant class of atypical lipomatous tumours. I also visualised the tile prediction onto WSIs to generate the class-prediction maps, and Fig 3.9 is an example, from which we can nd the classi cation of non-informative areas including background and artefact is poor in the Alex-family due to the mislabelling issue in the training tiles.

I also employed Inception V3 (Szegedy et al. 2016) to train our bespoke sarcoma dataset (Inception-7) as the second last of our tile-based CNN evolution. And nally, Inception V3 trained on x5 magni cation of tiles (x5 dataset is also

Table 3.4: A rough tile-level evaluation for the Alex-CNNs.

CNN ID	Lipoma Accuracy	ALT Accuracy	Overall Accuracy
Alex-a-4	33.63%	47.88%	41.20%
Alex-a-7	32.37%	43.94%	38.52%
Alex-b-4	87.11%	74.96%	80.65%
Alex-b-7	82.92%	66.11%	73.98%

Figure 3.8: The tile-level ROC for the Alex-based CNNs for the malignant class of atypical lipomatous tumours.

described in Section 3.3.2) was the nal version of the tile-based part of classi er (Inception-b-7-x5).

Table 3.5 shows the evolution of our tile-based lipoma and atypical lipomatous tumour classi er. At the beginning, no annotation of regions of interest (ROIs) was introduced to the slides and only binary classes of benign and malignant were considered based on their diagnostic labels. This was then expanded to 7 classes, including two dominant classes, a background class and four additional histology variation classes, after preliminary experiments. The training of 7 classes (versions in Table 3.5) was performed with the pathologist's annotation on an ROI basis.

(a) Sample 137.

(b) Class prediction map for Alex-a-4 on sample 137.

(c) Class prediction map for Alex-a-7 on sample 137.

(d) Class prediction map for Alex-b-7 on sample 137.

Figure 3.9: WSI-level visualisation using class prediction map for Alex-CNNs on lipoma sample 137. In the class prediction maps, the green colour stands for lipoma and red colour stands for atypical lipomatous tumour, grey colour stands for non-informative tiles including background and all other colours stand for minor histology variations (details in Section 3.4.2).

Table 3.5: The evolution of tile-based lipoma and atypical lipomatous tumour classi er.

CNN ID	CNN Structure	Class Count	Data Cleaned?	Magni cation
Alex-a-binary	AlexNet	2	N	x10
Alex-a-3	AlexNet	3	N	x10
Alex-a-4	AlexNet	4	N	x10
Alex-a-7	AlexNet	7	N	x10
Alex-b-4	AlexNet	4	Y	x10
Alex-b-7	AlexNet	7	Y	x10
Inception-b-7	Inception V3	7	Y	x10
Inception-b-7-x5	Inception V3	7	Y	x5

3.4 Optimisation of Methods

This section shows the optimisation of methods discussed before. Preliminary analysis and results will be presented and discussed to support the method optimisation process, and the final methodology of the tile-based workflow will be summarised at the end of this section.

3.4.1 The Tile-Level CNN Classifier

As mentioned in Section 3.3.1, I trained the Inception V3 model with our dataset of lipoma and atypical lipomatous tumour scanned using Hamamatsu NanoZoomer S360 machine. I prepared the dataset to be under x5 magnification as described in Section 2.2 and 3.3.2, and I also took some basic transformations such as flip and rotation, with a triggering probability of 0.5, to augment the training tiles. I used Adam optimiser (Zhang 2018) together with a learning rate scheduler as suggested by Loshchilov & Hutter (2017) to improve the training performance. I split the training tiles by class with a train-validation factor of 0.8 (the statistics shown in Table 3.3), and I employed the validation dataset to conduct a tile-level evaluation for our final version models of Inception-b-7-x5.

I first calculated the accuracy by class (Table 3.6). Furthermore, I employed confusion matrix to have a better understanding and comparison between the actual target labels and those predicted labels for each of the classes (Fig 3.10). And finally, ROC curves were plotted for both the Inception-b-7 and Inception-b-7-x5 models (Fig 3.11). According to these tile-level evaluation, the Inception-b-7-x5 model achieved a better performance for the dominant classes of lipoma, atypical lipomatous tumour and fat necrosis compared to Inception-b-7 and to all other Alex-family CNNs (Table 3.4) on the tile-level. Inception-b-7-x5 also outperformed slightly for background class from other models, and this shows Inception-b-7-x5 has a better ability to differentiate the tissue areas from the non-informative areas on the slide-level. This will be addressed in next slide-level visualisation section.

I also trained a fine-tuning version of Inception-b-7-x5 using ImageNet pre-trained model. Please refer to Fig 3.12 and Table 3.7 for more details of the per-

Table 3.6: Tile-level accuracy for Inception-b-7 and Inception-b-7-x5 models.

Class	Inception-b-7	Inception-b-7-x5
Lipoma	80.09%	88.95%
Atypical Lipomatous Tumour	69.58%	73.36%
Background	99.80%	99.96%
Muscle	93.83%	94.44%
Calci cation	88.89%	100.00%
In ammation	83.33%	100.00%
Fat Necrosis	88.26%	93.37%
Overall	87.26%	90.65%

(a) Tile-level confusion matrix (in percentage) for Inception-b-7.

(b) Tile-level confusion matrix (in percentage) for Inception-b-7-x5 (the nal tile-based CNN).

Figure 3.10: Tile-level confusion matrix for Inception-b-7 and Inception-b-7-x5 models.

Table 3.7: Tile-level accuracy for Inception-b-7-x5 (training from scratch) and Inception-b-7-x5 (ne-tuning).

Class	Training from scratch	Fine-tuning
Lipoma	88.95%	87.69%
Atypical Lipomatous Tumour	73.36%	73.06%
Background	99.96%	99.92%
Muscle	94.44%	93.06%
Calci cation	100.00%	80.00%
In ammation	100.00%	100.00%
Fat Necrosis	93.37%	96.32%
Overall	90.65%	90.52%

(a) Tile-level ROC curve for Inception-v3.

(b) Tile-level ROC curve for Inception-v3-x5 (the final tile-based CNN).

Figure 3.11: Tile-level ROC plots for Inception-v3 and Inception-v3-x5 models.

(a) Tile-level Inception-v3 (fine-tuned) confusion matrix.

(b) Tile-level ROC plot for Inception-v3-x5 (fine-tuned).

Figure 3.12: Inception-v3-x5 (fine-tuned) evaluation.

formance of the fine-tuned Inception-7-x5. These tile-level comparisons showed that the train-from-scratch CNN has a slightly better performance compared to the fine-tuned CNN in this computation task.

The conclusion at the tile-level stage is Inception-7-x5 (train from scratch) achieves the best performance over other CNNs, and is also slightly better compared to the fine-tuned Inception-7-x5. Inception-7-x5 (train from scratch) should be employed in the tile-based workflow for the slide-level diagnosis prediction.

3.4.2 The Slide-Level Visualisation

Although the previous section has showed the tile-level evaluation for different CNNs, pathologists will never make their diagnosis based on a single tile. Pathologists make their decision on the slide-level, so the first thing afterward is to visualise the CNN outputs and map the tile-level results on the slide-level. I first plotted a class-prediction map to spatially map the CNN prediction for each of the tiles within a whole slide image. There are 7 outputs for the 7-class Inception-7-x5 representing a score for each class, I applied SoftMax function to those 7 scores to calculate the probabilities of each tile being each in each class. I took the label of the highest probability for each tile as its prediction, and applied this strategy to all the tiles to plot the class-prediction map. Seven colours are used to represent seven CNN output classes, and Table 3.8 is the colour list for the class-prediction map. Fig 3.13 is an example where Fig 3.13a is the thumbnail of the original slides, Fig 3.13b is its underlying class-prediction map, and Fig 3.13c is the optimised version of class-prediction map with a background smoothing post-processing step to remove noise in the non-tissue area.

Fig 3.14 shows two examples of the class-prediction map (background optimised) with one conventional lipoma case and one conventional atypical lipomatous tumour case, and Fig 3.15 shows another two atypical lipomatous tumour cases with some histology variation areas.

One thing to be stressed is this tile-based CNN is able to differentiate the tissue area from the non-informative area as we took the background and artefact as an output class of the CNN. This CNN-based tissue detection method also outperforms

(a) Thumbnail image for atypical lipomatous tumour sample 102.

(b) Original class-prediction map for atypical lipomatous tumour sample 102.

(c) Background optimised class-prediction map for atypical lipomatous tumour sample 102.

Figure 3.13: A example class-prediction map of an atypical lipomatous tumour case 102 before and after background smoothing.

(a) Thumbnail image for lipoma sample 001.

(b) Class-prediction map for lipoma sample 001.

(c) Thumbnail image for atypical lipomatous tumour sample 206.

(d) Class-prediction map for atypical lipomatous tumour sample 206.

Figure 3.14: Two examples (001 and 206) of the class-prediction map with one conventional lipoma case and one conventional atypical lipomatous tumour case.

Table 3.8: Colour list in the class-prediction maps.

Class	Colour	Colour RGB
Lipoma	Green	[0.00, 1.00, 0.00]
Atypical Lipomatous Tumour	Red	[1.00, 0.00, 0.00]
Background	Grey	[0.70, 0.70, 0.70]
Muscle	Blue	[0.00, 0.00, 1.00]
Calci cation	Black	[0.00, 0.00, 0.00]
In ammation	Violet	[0.75, 0.50, 1.00]
Fat Necrosis	Pink	[1.00, 0.70, 0.85]

(a) Thumbnail image for atypical lipomatous tumour sample 035, there are some large muscle areas in this sample.

(b) Class-prediction map for atypical lipomatous tumour sample 035, and the muscle areas are corrected predicted by the CNN with blue colour in the map.

(c) Thumbnail image for atypical lipomatous tumour sample 097, there are some fat necrosis areas in this sample.

(d) Class-prediction map for atypical lipomatous tumour sample 097, and the fat necrosis areas are corrected predicted by the CNN with pink colour in the map.

Figure 3.15: Another two class-prediction map examples (035 and 097) for atypical lipomatous tumour cases with some histology variation areas.

(a) Sample 051 thumbnail.

(b) Sample 051 after traditional tissue detection pipeline.

(c) Sample 051 class-prediction map showing tissue and non-tissue area.

Figure 3.16: Comparison of the tissue detection task between traditional threshold-based pipeline and our tile-based CNN methods for sample 051.

the traditional threshold-based pipeline shown in Fig 3.6 for those paucicellular fatty tissue slides. I used the same failed sample of 051 shown in Fig 3.7c and 3.7d as an example to evaluate and compare the traditional pipeline and the tile-based CNN for the tissue detection task. In this example, the CNN performs signi cantly better at background detection than the traditional pipeline (Fig 3.16).

(a) Thumbnail image for lipoma sample 001.

(b) Heat map for lipoma sample 001.

(c) Thumbnail image for atypical lipomatous tumour sample 206.

(d) Heat map for atypical lipomatous tumour sample 206.

Figure 3.17: Two examples (001 and 206) of the heat map with one conventional lipoma case and one conventional atypical lipomatous tumour case.

(a) Thumbnail image for atypical lipomatous tumour sample 035, there are some large muscle areas in this sample.

(b) Heat map for atypical lipomatous tumour sample 035.

(c) Thumbnail image for atypical lipomatous tumour sample 097, there are some fat necrosis areas in this sample.

(d) Heat map for atypical lipomatous tumour sample 097.

Figure 3.18: Another two heat map examples (035 and 097) for atypical lipomatous tumour cases with some histology variation areas.

Though the class-prediction plot can map the tile-level prediction spatially and visually onto the slides, it will be good to have a heat map to highlight the dominant output classes of lipoma and atypical lipomatous tumour. I took the softmax-ed output value for the atypical lipomatous tumour class as the malignant score on the tile-level, and then plotted a heat map for each of the whole slide images using this tile-level malignant score. I used the same lipoma (sample ID: 001) and atypical lipomatous tumour (sample ID 206, 035 and 097) cases as examples to show the malignant heat map, see Fig 3.17 and Fig 3.18 for more details.

From the class-prediction map and the heat map we can nd there are some inconsistencies or ”noise” in some prediction areas within these plots, for example, we can nd some minor and isolated predicted histology variation tiles in the malignant area in Fig 3.14d. Another example is we can nd some minor and isolated

predicted benign or malignant tiles in the muscle area in Fig 3.15b. From a technical perspective of view, the prediction output for each of the tiles is independent, but from a biological perspective, isolated tiles that are predicted to be a different class from the surrounding ones are likely to be incorrectly predicted. One way to resolve this minor inconsistency and to optimise the plots is applying a Gaussian smoothing. Finally, I decide to keep the original class-prediction map and heat map to show how the original CNN outputs mapping onto the slide-level, and to use a contour map with Gaussian smoothing as a post-processing step to build the consistency on the basis of predicted areas. And these will be discussed in the next section.

The visualisation plots for all the training and testing slides are shared in this link https://gitfront.io/r/user-3345834/x4kyAGo621eb/ binghao-chai-phd-thesis-supplementary-material/ .

3.4.3 Colour Normalisation

Figure 3.19: Some samples for lipoma, atypical lipomatous tumour, muscle, calci cation, in ammation and fat necrosis tiles before and after colour normalisation using different algorithms.

As discussed in Section 1.6.3, there are debates on whether to apply colour normalisation in the tile-based work ow. For some of the computational tasks, the

colour normalisation process makes the performance better while in some other tasks not. I applied Reinhard's, Macenko's and Vahadane's colour normalisation methods through StainTool implementation By eld (2018) to some lipoma sample tiles, and Fig 3.19 shows some comparison samples of different classes before and after normalisation. It can be seen that in some cases, applying colour normalisation can result in artefacts such as the blue regions for some of these images. This blue-ish fake colour in eosin-stained muscle may be caused by the matrix inversion to solve the least square with a low staining coef cient that yields a negative value. This issue may be addressed by using non-negative matrix factorisation or using stain vectors from within the target image that is less sensitive. A comprehensive investigation in colour normalisation on fatty tissue could be conducted, but this is not really the scope of this study.

I designed an experiment to support the decision of whether colour normalisation should be employed in this speci c task. I took a subset of the training dataset, and trained two CNNs respectively with and without the colour normalisation as a preprocessing step. The slide-level accuracy of training with colour normalisation is 68.33% which is lower than without colour normalisation of 73.44%. No colour normalisation was applied in this speci c task according to these results.

3.4.4 Slide-Level Classi cation Using Tile Aggregation Methods

3.4.4.1 Aggregation features exploration

The next step after mapping and visualising CNN results onto WSIs is to nd a way to analyse and aggregate tile-level outputs into a diagnostic label prediction, and we call this tile-aggregation for the prediction of WSI-level diagnostic labels. A straightforward way to aggregate the tile-level benign and malignant information to WSI-level is to count the predicted-benign and predicted-malignant tiles within each of the whole slide images, and then calculate the proportions of predicted benign and malignant tiles in each of the slides, as stated in Table 3.9. This shows some information, such as whether our CNN model infers slides to be putative benign as the lipoma or putative malignant as the atypical lipomatous tumour. A baseline result based on this shows in Table 3.10. However, according to domain knowl-

edge, aggregating tile information this way is a sort of super cial and is not always correct, and we cannot simply use the ratio of the number of malignant tiles and the number of benign tiles as the “malignant score” of a slide. An extreme scenario for this statement is if we have a malignant case which is at a very early malignant stage, where most areas are benign apart from a very limited number of malignant tiles, we cannot make the diagnostic prediction in such cases by simply calculate the predicted benign and malignant proportions. Instead, further and deeper statistical analysis needs to be applied to nd some more proper aggregation features.

Table 3.9: De nitions of proportions of benign and malignant tiles in the slides.

Aggregation Features	De nition
Benign Proportion	$\frac{\text{predicted lipoma tiles count}}{\text{total informative tile count in the slide}}$
Malignant Proportion	$\frac{\text{predicted ALT tiles count}}{\text{total informative tile count in the slide}}$

Table 3.10: The baseline results for the tile-based work ow on the internal testing dataset.

Label	Correct	Wrong	Total	Accuracy (%)
Lipoma	28	11	39	71.79%
ALT	18	7	25	72.00%
Overall	46	18	59	71.88%

I then plotted the tile-based CNN output (after SoftMax) distribution for the malignant ALT class for all the tiles in each of the slides, and some examples are displayed in Fig 3.20. The distribution for case 097 shows on the left-hand side, and on the right-hand side are for cases of 007, 039, 121, 056, 088, 127, 125 and 134. This gure shows the distribution of the probabilities for all the tiles to be malignant within one WSI, where high score means high malignancy for that tile. Low scores in this plot do not mean benign, as low score can be background or other histology variation tiles as well.

To better visualise the tile-level malignant score distribution, I made the plot with only the informative tiles (i.e., the putative malignant tiles and the putative benign tiles) rather than all the tiles within each of the slides. Fig 3.21 shows a

(a) Tile-level malignant score distribution for ALT sample 097.

(b) Tile-level malignant score distribution for 8 cases as another example.

Figure 3.20: Tile-level malignant score distribution (all tiles included) samples.

lipoma and an atypical lipomatous tumour example for the distribution excluding all background areas as well as the histology variation areas.

(a) Tile-level distribution for all informative tiles for lipoma sample 111.

(b) Tile-level distribution for all informative tiles for atypical lipomatous tumour sample 206.

Figure 3.21: Tile-level distribution for all the putative lipoma and ALT tiles for a lipoma and an atypical lipomatous sample.

Furthermore, Fig 3.22 shows the tile-level malignant score distribution for all the informative tiles (i.e., only putative benign tiles and putative malignant tiles are included, all other tiles such as background and histology variation are excluded)

for all of the whole slide images, where Fig 3.22a shows the distribution for all training slides inferred by Inception-7 under magni cation of x10, and Fig 3.22b shows the distribution for all the training slides inferred by Inception-7-x5 under magni cation of x5. Statistic metrics, such as mean or median, of the distribution for each of the slides could be used as one of the aggregation features.

To better visualise the malignant area in each of the slides, I also employed contour map as a supplementary to the heat map. I took the median of the median malignant scores for all the informative tiles for all the ALT slides (0.747), and the median of the median malignant scores for all the informative tiles for all the lipoma slides (0.378) shown in the distribution of Fig 3.22 as the initial contour lines, and 0.9 was also provisionally taken as a supplementary contour lines to show the extreme heat areas over the slides. Fig 3.23b shows an initial example of the contour map for a malignant case of 206. The contour map can also be optimised by applying Gaussian smoothing as well as removing the background areas (i.e., non-tissue areas), as shown in Fig 3.23c and Fig 3.23d.

A supplementary instructions on how the initial contour lines of 0.378 and 0.747 are calculated:

1. Calculate the median of tile-level malignant scores for all the informative tiles within each WSI.

2. Take the above median values as the WSI-level malignant scores.

3. Calculate the median of WSI-level malignant scores for all the ALT slides which equals to 0.747.

4. Calculate the median of WSI-level malignant scores for all the LP slides which equals to 0.378.

There are ve areas in the nal contour map for each of the slides: (1) the white area is the non-tissue area (i.e., the background), (2) contour value from 0 to 0.378 is the putative benign area shown in blue in the contour map, (3) contour value from 0.378 to 0.747 is the uncertain area shown in green in the contour map, (4)

(a) Tile-level distribution for all training slides inferred by Inceptionb-7 under magni cation of x10.

(b) Tile-level distribution for all training slides inferred by Inceptionb-7-x5 under magni cation of x5.

Figure 3.22: Tile-level malignancy distribution for all training slides with only informative (i.e., putative benign tiles and putative malignant) tiles.

(a) Thumbnail image for sample 206.

(b) Contour map (initial) for sample 206.

(c) Contour map (after Gaussian smoothing) for sample 206.

(d) Contour map (after Gaussian smoothing and non-tissue area removed) for sample 206.

Figure 3.23: Contour map example for atypical lipomatous tumour sample 206.

contour value from 0.747 to 0.9 is the putative malignant area shown in orange in the contour map, and (5) contour value from 0.9 to 1.0 is also the putative malignant area, but this range of area is the hottest malignant area in the slides, shown in red in the contour map. As I would like to capture the benign and malignant information, making a scatter plot that shows each WSI as a single point with the putative benign areas (contour value from 0 to 0.378 shown in blue) against the putative malignant areas (contour value from 0.747 to 1.0 shown in orange and red) would be useful. So I plotted the putative benign area (contour values from 0 to 0.378) against the putative malignant area (values from 0.747 to 1.0) as a scatter plot (Fig 3.24). This gure shows roughly two regions, where the malignant cases are on the left-hand side and benign cases are on the right-hand side. And these align with the x-axis

and y-axis. Case 035 is a unique one as most of the area in slide 035 is histology variation of muscle, but as I only take the putative benign and putative malignant area into account in the scatter plot, case 035 is on the correct malignant cluster in the plot. Case 005 and 006 are two dif cult cases as most of the areas in their contour maps are uncertain or benign, but they are in fact both malignant samples, and these two cases are wrongly shown in the benign clusters in the scatter plot. All other samples seems to have consistency with their ground-truth diagnostic label.

At this stage, I extracted six putative features on top of the tile-level malignancy and the slide-level contour map, they are:

1. Median value of the tile-level malignant score for all the informative tiles within each of the whole slide images.

2. Median value of the tile-level benign score for all the informative tiles within each of the whole slide images.

3. The proportion of areas of contour values range from 0.9 to 1.0 (i.e., the hottest putative malignant red areas) over all the informative areas in the contour map.

4. The proportion of areas of contour values range from 0.747 to 0.9 (i.e., the putative malignant orange areas) over all the informative areas in the contour map.

5. The proportion of areas of contour values range from 0.378 to 0.747 (i.e., the uncertain green areas) over all the informative areas in the contour map.

6. The proportion of areas of contour values range from 0 to 0.378 (i.e., the putative benign blue areas) over all the informative areas in the contour map.

Fig 3.25 displays the data structure of all mentioned features (before re ne-ment) for the internal training slides visualised using PCA as well as t-SNE, and both sub- gures show consistency.

Figure 3.24: Scatter plot based on putative benign (contour value from 0 to 0.378 shown in blue) and putative malignant (contour value from 0.747 to 1 shown in orange and red) areas of contour maps for all the training slides.

(a) PCA results for all 6 coarse features for all the internal training slides.

(b) t-SNE results for all 6 coarse features for all the internal training slides.

Figure 3.25: The data structure of all mentioned putative features for the internal training slides visualised using PCA as well as t-SNE.

3.4.4.2 Aggregation features selection and re nement

In order to re ne the putative features, some further analysis was applied. Table 3.11 shows the percentage of variance explained by each of the selected components using PCA from which we can nd that PC-1 is the main axis. Table 3.12 shows the principal axes in feature space, representing the directions of maximum variance in the data, and we can nd putative feature 2 (i.e., the median of tile-level benign score) as well as putative feature 6 (i.e., the proportion of putative benign area on the contour map) both have reasonable positive values showing that they positively predict lipoma which is what we would expect. On the other hand, putative feature 1 (i.e., the median of tile-level malignant score) as well as the putative feature 4 (i.e., the proportion of the dominant putative malignant area on the contour map) both have reasonable negative values showing their relationship to ALT. Putative feature 3 (i.e., the proportion of the minor hottest putative malignant area on the contour map) and putative feature 5 (i.e., the proportion of the uncertain area on the contour map) make the least contribution to distinguishing of the benign lipoma and the malignant ALT. A random forest classi er was also applied to the internal training and testing set, the feature ranking of all 6 features shows a similar result

of the feature importance, see Table 3.13 for details.

Table 3.11: Percentage of variance explained by each of the selected components using PCA.

Principal Axes	Percentage of variance
PC-1	0.7535
PC-2	0.2255

Table 3.12: Principal axes in feature space representing the directions of maximum variance in the data.

PC Axes	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6
PC-1	-0.4581	0.4703	-0.0775	-0.4109	-0.1264	0.6149
PC-2	-0.0897	0.0664	-0.1493	-0.4372	0.8419	-0.2555

Table 3.13: Feature ranking generated by random forest method.

Feature ID	Importance
Feature 1	0.2658
Feature 2	0.2956
Feature 3	0.0588
Feature 4	0.1531
Feature 5	0.0196
Feature 6	0.2072

According to both the putative feature analysis above and the pathological knowledge, I rstly suggest the feature 5 to be removed from the slide-level tile aggregation feature list as this feature might be non-informative to some degree. Regarding feature 3, a potential reason for this feature to be the least signi cant is not all the ALT case contains the red areas on the contour map. As a result, I then suggest that feature 3 (i.e., the proportion of areas of contour values range from 0.9 to 1.0) to be merged into feature 4 (i.e., the proportion of areas of contour values range from 0.747 to 0.9), and make them to be a single feature of the proportion of the putative malignant areas (i.e., the proportion of areas of contour values range from 0.747 to 1.0) in the contour map visualisation. So the putative features can then be re ned to be the following four:

1. Median value of the tile-level malignant score for all the informative tiles within each of the whole slide images.

2. Median value of the tile-level benign score for all the informative tiles within each of the whole slide images.

3. The proportion of areas of contour values range from 0.747 to 1.0 (i.e., the putative malignant orange and red areas) over all the informative areas in the contour map.

4. The proportion of areas of contour values range from 0 to 0.378 (i.e., the putative benign blue areas) over all the informative areas in the contour map.

Figure 3.26: Hyperparameter search of the putative benign and putative malignant contour lines using the validation accuracy.

To re ne the contour lines, I also conducted some hyperparameter search with the validation set. Fig 3.26 is the result showing the hyperparameter search process, from which we can nd a proper putative benign contour line is 0.25, and the putative malignant contour line is 0.75.

To conclude, two contour map based features were employed together with two tile-level statistical metrics for the tile aggregation process, the aggregation features

are listed in Table 3.14. And a slide-level diagnostic label can be predicted using this tile aggregation method, the aggregation results will be discussed in Section 3.5 and 3.6.

3.4.5 Methodology Summary

In this study, 59 lipoma and ALT slides were employed in the training process of a lipoma-ALT classi er. I rst generated ROIs based on the pathologist's annotation for seven training classes, they are: (1) background and artefact, (2) conventional lipoma, (3) conventional ALT, (4) muscle, (5) calci cation, (6) in ammation, (7) fat necrosis. Smaller tiles were then generated from these ROIs and used to train an Inception-V3 model to create a tile-based CNN classi er. I visualised the tile-level outputs and mapped them onto the slide-level. A class prediction map, a heat map and a contour map were generated for each of the slides. I extracted and re ned four tile-aggregation features from the statistics and the visualisation (Table 3.14), and then trained a logistic regression and a random forest classi er for the slide-level diagnostic prediction of the lipomas and the ALTs. Please refer to Fig 3.1 as a graphical abstract for this full tile-based study.

Table 3.14: Tile aggregation feature list in the tile-based work ow.

ID	Feature	Description
1	median_benign	Median value of the tile-level benign score for all the informative tiles within each of the whole slide images.
2	median_malignant	Median value of the tile-level malignant score for all the informative tiles within each of the whole slide images.
3	area_benign	The proportion of areas of contour values range from 0 to 0.25 (i.e., the putative benign blue areas) over all the informative areas in the contour map.
4	area_malignant	The proportion of areas of contour values range from 0.75 to 1.0 (i.e., the putative malignant orange and red areas) over all the informative areas in the contour map.

3.5 Results and Discussion

A logistic regression and a random forest model were employed for the slide-level classi cation of WSIs. The slide-level results for the entire tile-based work ow (i.e., the tile-based CNN plus the tile-aggregation) for internal training slides are shown in Table 3.15. For logistic regression, the F-score for lipoma is 0.943, and for ALT is 0.954, and for the random forest method, it is 0.963 for lipoma and for 0.969 for ALT.

Table 3.15: Slide-level results for the tile-based work ow on the internal training dataset.

Method	Label	Correct	Wrong	Total	Accuracy (%)	AUC
Logistic Regression	Lipoma	25	1	26	96.15%	0.987
	ALT	31	2	33	93.94%	
	Overall	56	3	59	94.92%	
Random Forest	Lipoma	26	0	26	100%	0.998
	ALT	31	2	33	93.94%	
	Overall	57	2	59	96.61%	

Feature analysis is also conducted based on the logistic regression and random forest, the logistic regression coef cients and random forest feature importance are listed in Table 3.16. These feature values make a lot of sense: the putative lipoma proportion on the contour map as well as the median of benign scores both have negative values for logistic regression showing that they negatively predict atypical lipomatous tumour, and on the other hand, the putative malignant proportions on the contour map and the median of the malignant score have a positive value for logistic regressions showing that the system correctly uses this information. And in random forest, most of the features make a reasonable contribution to the differentiation between lipoma and ALT as well.

The tile-based work ow was then applied to the internal testing slides (described in Section 2.2), and similarly, Table 3.17 and Fig 3.27 show the general results. It's very dif cult to say whether random forest or logistic regression model is generally better than the other one, as the testing dataset size is small and both are doing much better than random. The list of wrong predictions is shown in Table 3.18 for the internal testing set.

Table 3.16: Aggregation feature analysis based on logistic regression and random forest.

Feature ID	Feature	Logistic Regression	Random Forest
1	medianbenign	-1.755	0.362
2	medianmalignant	1.720	0.362
3	areabenign	-0.763	0.070
4	areamalignant	2.243	0.205

Table 3.17: Slide-level results for the tile-based work ow on the internal testing dataset.

Method	Label	Correct	Wrong	Total	Accuracy (%)	AUC
Logistic Regression	Lipoma	29	10	39	74.36%	0.776
	ALT	18	7	25	72.00%	
	Overall	47	17	64	73.44%	
Random Forest	Lipoma	27	12	39	69.23%	0.789
	ALT	18	7	25	72.00%	
	Overall	45	19	64	70.31%	

(a) Slide-level ROC curve for the tile-based work ow on the internal testing dataset using logistic regression.

(b) Slide-level ROC curve for the tile-based work ow on the internal testing dataset using random forest.

Figure 3.27: Slide-level ROC curves for the tile-based work ow on the internal testing dataset.

Table 3.18: Incorrect predictions for the tile-based work ow on the internal testing dataset.

Diagnostic Label	Incorrectly Predicted Slides (Logistic Regression)	Incorrectly Predicted Slides (Random Forest)
Lipoma	021, 033, 128, 131, 138, 139, 151, 181, 182, 185	021, 033, 128, 131, 137, 138, 139, 151, 173, 181, 182, 185
ALT	023, 038, 072, 074, 076, 106, 189	023, 038, 072, 074, 076, 106, 189

Though a tile-aggregation method was performed, there exist false predictions. All the incorrectly predicted slides were sent to Sahlgrenska University Hospital for an independent pathological review. Table 3.19 compares the slide-level ground truth, the work ow prediction and the pathological review for the incorrectly predicted slides. The incorrectly predicted lipoma cases are dif cult cases with either variation presenting on the slide or focal out-of-focus areas. All the incorrectly predicted ALT cases were diagnosed to be lipoma by pathologists with only the histopathological images to be used. Pathologists indicated these cases to be challenging, and suggested a further MDM2 FISH test would consider to make a more accurate diagnosis.

The tile-based classi er was also evaluated on the selected external testing slides (Section 2.3). Table 3.20 shows a slide-level comparison results of the tile-based classi er for the internal and external testing sets. There are two observations based on this table. First, the tile-based classi er tends to predict WSIs to be lipoma for the external testing slides, that 95% of the lipoma cases are correctly predicted. Then, more than half of the external ALT cases are incorrectly predicted to be lipoma, and so the accuracy of the ALT also drops signi cantly in the external testing slides. This is an issue of sensitivity to the ALT class, and the generalisability of the classi er will be discussed in the next section.

Table 3.19: The comparison of the slide-level ground truth, the work ow prediction and the pathological review for the tile-based incorrectly predicted slides.

Slide ID	Original Pathological Diagnosis	Tile-Based Prediction	Pathological Review	Pathological Comments
021	Lipoma	ALT	Lipoma	
033	Lipoma	ALT	Uncertain but more like Lipoma	with hibernomatous change
128	Lipoma	ALT	Lipoma	
131	Lipoma	ALT	Lipoma	Small area not in focus
138	Lipoma	ALT	Lipoma	Small area not in focus
139	Lipoma	ALT	Lipoma	Intramuscular
151	Lipoma	ALT	Lipoma	Growth in fascia, spindle cells, would consider MDM2 FISH
181	Lipoma	ALT	Lipoma	
182	Lipoma	ALT	Lipoma	
185	Lipoma	ALT	Lipoma	
023	ALT	Lipoma	Lipoma	Dif cult but there are some atypical perivascular cells
038	ALT	Lipoma	Lipoma	Intramuscular, small area not in focus
072	ALT	Lipoma	Lipoma	Intramuscular
074	ALT	Lipoma	Lipoma	Small area not in focus
076	ALT	Lipoma	Lipoma	Small area not in focus
106	ALT	Lipoma	Lipoma	
189	ALT	Lipoma	Lipoma	

Table 3.20: The slide-level results of the tile-based classi er for the internal and the external testing slides.

Method	Label	Internal Testing Set Accuracy (%)	F1	AUC	External Testing Set Accuracy (%)	F1	AUC
Logistic Regression	Lipoma	74.36%	0.773	0.776	95.00%	0.655	0.695
	ALT	72.00%	0.679		42.42%	0.583	
	Overall	73.44%	–		62.26%	–	
Random Forest	Lipoma	69.23%	0.740	0.789	95.00%	0.667	0.684
	ALT	72.00%	0.655		45.46%	0.612	
	Overall	70.31%	–		64.15%	–	

3.6 Conclusions

In this study, I proposed a CNN-based work ow for the classi cation of H&E stained lipoma and atypical lipomatous tumour histopathological images. On the tile-level, the tile features were learned by the deep convolutional neural networks. And furthermore, tile-aggregation methods using logistic regression and random forest were also performed for the slide-level prediction of the diagnostic labels. The work ow was trained on an annotated and re ned dataset, and then validated and tested on a separate dataset. The combination of the CNN plus the tile-aggregation classi ers achieves comparable results showing the success of the distinction of lipoma and ALT slides. Finally, an external and more challenging dataset was used as the external validation and testing, and this is to explore the generalisability of the work ow.

Pathologists played an important role in the development of this tile-based work ow. Unlike computational tasks of classi cation and detection of general objects in daily life, pathological images need to be annotated carefully by domain experts. Though manual annotating is time-consuming, the quality of the re ned training dataset depends on the quality of annotation, and ROI annotation by pathologists is an essential step in the dataset preparation process. Pathologists also gave their reviews in the tile-level evaluation and slide-level aggregation processes which accelerate the evolution of the work ow.

In most cases, the results of tile-based classi er cannot be used directly for the diagnostic classi cation of the slides, and there are two reasons for this. First, the CNN cannot always predict all the tiles correctly, and on the other hand, some tumours arise from a tumour centre and then grow larger amongst benign tissue. An extreme example is if for an early tumour slide, there are 99% of the whole informative area showing benign while only 1% of areas in the slides showing malignant, assuming the tile-based CNN have an 100% accuracy, and in this case, the slide should be classi ed as malignant rather than benign. Slide-level tile aggregation is a key step on top of the tile-based classi er when dealing with real-world data. In this work ow, a tile-aggregation method is employed for the slide-level diagnostic

label prediction. I took a combination of the statistical features and the contour map, and then employ a logistic regression and a random forest to make the classi cation. The contour map itself might potentially be used as the input of another CNN to capture the geometrical or morphological features on the slide-level and train a diagnostic classi er if suf cient numbers of samples and contour maps can be generated. However, in this study, I only have 59 training points and so I can only aggregate tiles by those more fundamental features.

This tile-based work ow shows some abilities to differentiate lipoma from ALT on the independent external dataset to some degree, however, the classi er is not as good as inference with the in-house dataset. Generalisability is a common problem in the digital pathology eld, and is one of the reasons why only a few work ows are employed in clinical usage. The generalisability issue may be addressed by training with external datasets. Future work for other researchers includes the exploration of generalisability and the aggregation of a patient-level classi er if they have suf cient data.

Chapter 4

A Novel Lipoma and Atypical Lipomatous Tumour Nucleus Dataset

4.1 Chapter Overview

The tile-based methods that are so widely used in deep learning models differ substantially from how pathologists view lipoma slides i.e. pathologists use low-power magni cation images to have a survey of the architecture of the tumour and then go down to higher magni cations to better determine speci c characteristics of cells. In particular, the identi cation of lipoblasts or atypical hyperchromatic cells are key diagnostic ndings that could prompt a diagnosis of malignancy. Given the "black box" nature of deep learning models, it is not known if the tile-based methods capture these important cellular or nuclear features. Therefore using an entirely tile-based work ow to distinguish ALT from lipoma could be seen as a limitation. As introduced in Section 1.4, pathologists will make use of the aberrant nuclei and the lipoblast for the distinction of lipomas and ALTs. Our hypothesis is that the distinction between the lipomas and the ALTs can achieve a better performance if we mimic the pathologist's routine process of assessing slides by focusing on the nuclei of the slides. This kind of "nucleus-based" work ow could be evaluated by taking its standalone performance or integrating it into the tile-based one. In order to develop such a deep learning work ow that is essentially based on what pathologists do when analysing lipomas and ALTs, a bespoke nucleus-level lipoma and

ALT dataset would be needed for the training process.

4.2 Introduction and Related Work

4.2.1 Existing Nucleus-Level Datasets

There are several open-source datasets of object classi cation, detection and segmentation for different morphological features from digitised slides. The computational tasks are mainly mitosis detection, lymphocyte detection and nuclei classi- cation, detection and segmentation. Table 4.1 shows more details of the nucleus-level datasets for different computational tasks. A thorough review of the literature and accessible data repositories revealed no open-source lipoma and ALT nucleus datasets. The rst step to develop such a nucleus-based work ow and to evaluate the proposed assumption is to develop a lipoma and ALT nucleus dataset. Techniques such as object detection and segmentation could be used for the pixel-level nucleus dataset generation.

4.2.2 Traditional Segmentation Methods

There exist many well-known conventional segmentation algorithms (i.e., machine learning or deep learning methods are not used), which can be mainly categorised into thresholding segmentation, edge-based algorithm and region-based segmentation (Irshad et al. 2014).

Thresholding segmentation is a straightforward way to segment the selected object from an image: simply choose a threshold based on the intensity histogram of single channel values. The threshold can be applied on either single-channel gray scale values, or multi-channel of red, green and blue values for segmentation. Unfortunately, thresholding algorithms might miss the signi cant parts of the identi ed object, or merge parts of the background with the identi ed object and would then give an inappropriate outline of the objects.

Regarding edge-based segmentation, the main idea is to delineate the contours of the speci c object using related algorithms like Canny edge-detector or Sobel edge-detector, thereafter, these contours are then lled using mathematical morphology. Although this method performs better compared to thresholding, edge-

Table 4.1: Open datasets for nucleus detection and classi cation task.

Dataset	Annotation	Images
MITOS (Tripathi et al. 2013)	Pixel-level annotation of mitotic cells	50 images of 2048*2048px, 2250*2250px, 2767*2767px
MITOS-14 (22nd International Conference on Pattern Recognition 2014)	Centroid pixel of mitotic cells	2112 images of 1663*1485px
TUPAC (Veta et al. 2019)	ROI of mitotic cells	821 images from 73 cases
CCMCT (Bertram et al. 2019)	44,880 mitotic annotations	44,880 mitotic annotations
CoNSep (Graham, Vu, Raza, Azam, Tsang, Kwak & Rajpoot 2019)	Pixel-level nuclei annotation with associated class labels	24,319 exhaustively annotated nuclei with associated class labels
CRCHistoPhenotypes Sirinukunwattana et al. (2016)	29,756 nuclei centres out of which 22,444 with associated class labels	29,756 nuclei from 100 WSIs
PanNuke (Gamper et al. 2019, 2020)	Pixel-level nuclei annotation with associated class labels	205,343 labeled nuclei from 481 visual elds
MoNuSeg (Kumar et al. 2017)	Pixel-level nuclei annotation	29,000 nuclei from 44 images
LYON (Swiderska-Chadaj et al. 2019)	171,166 lymphocytes in 932 ROIs were annotated	83 WSIs
Lizard (Graham et al. 2021)	Instance-level annotation for colonic nuclei	About half a million nuclei

based segmentation is still not robust, as contours that are not perfectly closed will not be filled correctly.

Region-based segmentation outperforms either thresholding or edge-based methods in many scenarios, and an existing representative algorithm for region-based segmentation is watershed segmentation. The first step of watershed segmentation is to find an elevation map of an image using Sobel gradient method. Then markers of the background and the to-be-identified objects need to be defined based on the extreme parts of the density histogram of the selected channel. And finally, the watershed transform can be perfectly applied to fill regions of the elevation map starting from the determined markers.

One suggestion when applying traditional image processing approaches to segment nuclei in the H&E images is to execute the colour deconvolution before any of the segmentation operations (more details are given in Section 1.6.3). As nuclei are mainly bound with hematoxylin, applying algorithms such as watershed segmentation on the hematoxylin-channel in H&E-Complement colour space should perform better than applying the same algorithms on the original RGB colour space. Fortunately, existing tools are available for a fast implementation of nuclei identification and quantification with traditional image segmentation approaches. A straightforward workflow could be like this: first of all, OpenSlide (Goode et al. 2013) provides the interfaces to load and handle the large-size pathological images without any pre-processing, and secondly, skimage.color could be employed for colour deconvolution. Subpackages skimage.feature, skimage.morphology and skimage.filters enables conventional segmentations to be implemented. And in the end, label function in scipy.ndimage package provides quantification features.

4.2.3 Nuclei Classification, Detection and Segmentation Using Deep Learning Methods

There are three main types of computational tasks at nucleus-level in digital pathology which are classification, detection and segmentation. Classification summarises an image into a few words (i.e., classes), for example, a classification result tells us the types of nuclei in a tile containing different nuclei. More specific details, such

as location of the nucleus, is also important and can be useful information to have in some scenarios. Object detection is a computer vision technique that uses bounding boxes to localise objects on top of the classi cation, and in digital pathology, the object can be a nucleus for instance. Segmentation provides pixel-level details which outline the exact shape of each of the objects within an image. On the other hand, segmentation can reveal pixels from overlapping objects.

Nuclei classi cation, detection and segmentation are important in digital pathology and can be used for tumour diagnosis and grading (Srinidhi et al. 2021). There exist many nuclei detection, segmentation and quanti cation methods for H&E images, and these approaches vary in principle and process. All existing published algorithms can be divided into two groups: detection using traditional image processing methods, and detection based on machine learning or deep learning techniques.

Some of the traditional methods rst segment clumps of nuclei areas from the rest of the tissue and then proceed to deal with the overlapping nuclei to separate those into individual ones, whereas others act reversely. The earliest work for nuclei detection dates back to 1996, Thiran & Macq (1996) proposed a computational method using traditional image process techniques for nuclei segmentation based on their underlying morphological features. Later in 2000, Automated Cellular Imaging System (ACIS) developed by ChromaVision Medical Systems Inc. is the rst hardware system designed for nuclei detection. And after that, Bauer et al. (2000), Wang et al. (2001) tested the reliability and sensitivity of ACIS. They stated that ACIS is a colour-based analysis system, and can only be able to analyse physical slides with speci c staining protocol. Qi et al. (2012) rst presented a modern nuclei detection algorithm for digital H&E slides by following the direction of the image gradient to infer the geometric centre of each nucleus. However, the limitation of this pipeline is its rst step of inferring the initial markers as the location of nuclei, and this led to low performance and uncertainty. Veta et al. (2013) proposed a watershed-based nuclei segmentation pipeline for H&E-stained breast cancer images that achieved a positive predictive value of 0.886 on their validation set.

To my knowledge, the first published machine learning techniques used for nuclei detection were in 2010. A supervised segmentation method using Bayesian classification was developed for the segmentation of overlapping nuclei, and this work reduced the over- and under-segmentation to some degree (Jung et al. 2010). Then in 2013, the first deep learning based method for nuclei detection and segmentation was developed where a pixel-based CNN classifier was developed for the mitosis detection task using the ICPR2012 dataset (Cireşan et al. 2013). Wang et al. (2014) integrated handcrafted features on top of the deep learning method using the same ICPR2012 dataset for the same computational task. Xing et al. (2015) developed a combination workflow of the CNN and the selection-based sparse shape model for the nuclei segmentation of the brain, pancreatic and breast cancer images. Thereafter, Sirinukunwattana et al. (2016) proposed a spatially constrained convolutional neural network (SC-CNN) architecture which regresses the likelihood of a pixel being the centre of nuclei, where high probability values are spatially constrained to locate in the vicinity of the centres of nuclei. And in the same year, Xu et al. (2016) published a stacked sparse auto-encoder (SSAE) workflow to learn the higher-level morphological features for the ER+ breast cancer nuclei detection task. Naylor et al. (2018) developed a CNN-based regression model for the colon nuclei segmentation using TNBC and MoNuSeg datasets.

One of the latest and most advanced CNN architectures for nuclei classification and segmentation is HoverNet (Graham, Vu, Raza, Azam, Tsang, Kwak & Rajpoot 2019). HoverNet is a multiple-branch network and was applied on CoNSeP, Kumar, MoNuSeg and PanNuke datasets for nuclei classification and segmentation. The pre-activated residual network with 50 layers (Pre-act-ResNet50) was first employed in HoverNet for the feature extraction. Three distinct up-sampling branches were then integrated onto the Pre-act-ResNet50 for nucleus segmentation and classification. These three branches are:

1. nuclear pixel (NP) branch: first predicts on the pixel-level whether a pixel belongs to nucleus,

2. the hover branch: predicts the horizontal and vertical distances of nucleus

pixels to their centres of mass in order to separating touching nuclei,

3. the nucleus classi cation (NC) branch: predicts the type of nucleus by aggregating the pixel-level types within each nucleus instance.

HoverNetdemonstrated state-of-the-art performance compared to another 14 methods and tools, including the mainstream of U-Net and Mask-RCNN, on nucleus classi cation and segmentation on multiple independent multi-tissue pathological datasets.HoverNetachieved a dice score of 0.853 on its main dataset (CoNSeP).

4.2.4 Image Registration

Most published image registration studies related to medical imaging are focused on computerised tomography (CT) and magnetic resonance (MR) images. Image registration is a key component prior to 3D reconstruction. Registration methodologies are commonly classi ed using feature space image information. This information could be the intensity of the raw pixels, the intensity gradient or statistical information related to the intensity, or structures extracted from the images to be registered, such as sets of points, edges, contours, surfaces and volumes and ROIs according to (Song et al. 2017, Lundervold & Lundervold 2019, Fu, Lei, Wang, Curran, Liu & Yang 2020, Oliveira & Tavares 2014).

Moles Lopez et al. (2015) proposed an ROI-matching work ow for the registration of a series of digital slides in digital pathology. Tellez et al. (2018) aimed to detect mitosis in the pathological slides, and they would like to use another stain apart from H&E as their reference and ground truth. In this study, Tellez et al. (2018) designed a two-step pipeline for pathological images co-registration for the same tissue stained with H&E and H-DAB, with the rst step a global and coarse registration to minimise the vertical and horizontal shift between image pairs, and then to repeat this procedure multiple times for each of the image pairs before averaging across all the trials, and the second step is to register each mitotic gure individually following a similar procedure as before. Bulten et al. (2019) mentioned in their studies that there are additional nonlinear deformations caused by

the chemical treatment during the re-staining process and these sometimes result in the unmatching of spatially corresponding structures. They proposed to use the normalised gradient elds (NGF) distance that measures the alignment of image gradients, to account for the multi-modality of the registration problem.

The registration of H&E and IHC image pair could be used in nuclei instance level dataset generation, this is addressed in Section 7.2 (Fig 7.2).

4.3 Methodology: A Semi-Automatic Nucleus Annotating Work ow

There are potentially two ways for developing a nucleus-level dataset. On the one hand, we can build the dataset using the pathologist's annotation on the nucleus level. This is impracticable, as the approximate average number of nuclei in a lipoma slide is 20,000, and this number in an ALT slide ranges from 30,000 to 200,000. Manual annotation on the nucleus level is time-consuming and is not cost-ef cient. On the other hand, we can attempt to annotate the lipoma and ALT nuclei using a semi-automatic or automatic work ow and build the dataset.

4.3.1 HoverNet Pretrained with PanNuke Dataset

I rst took advantage of HoverNet as well as the pan-cancer dataset PanNuke. There are two modes for HoverNet, the Image Mode and the WSI Mode, and Table 4.2 shows the input and output of HoverNet in either mode. According to (HoverNet Github Site 2019), both the Image Mode and the WSI Mode output a .json le with keys of:

1. bbox: bounding box coordinates for each nucleus
2. centroid: centroid coordinates for each nucleus
3. contour: contour coordinates for each nucleus
4. type_prob: per class probabilities for each nucleus (default con guration doesn't output this)
5. type: prediction of category for each nucleus

Table 4.2: Input and output of HoverNet in each of the inference mode.

Mode	Input	Output
WSI	Whole slide images supported by OpenSlide (Goode et al. 2013), such as .svs, .tif, .ndpi and .mrxs	A .json le showing the detected nuclei instances
Image	Standard images les, including .png, .jpg and .tiff	A .json le showing the detected nuclei instances, an instance segmentation map in mat format, and an overlaid results on image

The inference process for HoverNet takes a massive amount of GPU hours for either mode. An estimation of inference of our 59 training slides using HoverNet will be over 2,000 running hours on a local machine (eight-processor Xeon(R) W-2123 CPU and Nvidia Quadro P1000 GPU). It will take even more time for the inference of 64 internal testing slides and 402 external testing slides, or for the re-train of a bespoke HoverNet. Multiple high-performance GPU nodes on the UCL-HPC machine, Myriad, were employed to accelerate the inference process of HoverNet. Technical notes of Appendix E and Appendix F show a sample script and the development method respectively.

(a) A PanNuke dataset example: the original RGB image.

(b) A PanNuke dataset example: the instance segmentation mask.

(c) A PanNuke dataset example: the classi cation type mask.

(d) A PanNuke dataset example: the 5-channel stack.

Figure 4.1: An example of PanNuke dataset.

PanNuke is a semi-automatically generated pan-cancer nuclei dataset for the computational tasks of segmentation and classi cation (Gamper et al. 2019, 2020). The nuclei are selected from 481 regions of interest, of which 312 are randomly sampled from more than 20K whole slide images. There are 205,343 well-annotated

Figure 4.2: PanNuke dataset nuclei type statistics (Gamper et al. 2019, PanNuke: An Open Pan-Cancer Histology Dataset for Nuclei Instance Segmentation and Classi cation 2019).

nuclei in the PanNuke dataset, where each of them has an instance segmentation mask and its underlying classi cation type map. Fig 4.1 shows an example of the PanNuke dataset, where Fig 4.1a is the original RGB tile image, Fig 4.1b is the instance segmentation mask that each nucleus has a unique ID, Fig 4.1c is the classi cation type mask that each nucleus type has a unique ID, and Fig 4.1d shows how those channels are stacked together into a 5-channel image. The PanNuke dataset consists of ve annotated classes and a miscellaneous class for all other nuclei. The ve dominant classes are: in ammatory nuclei, neoplastic nuclei, connective nuclei, dead nuclei and epithelial nuclei. Fig 4.2 shows the nuclei type statistics of PanNuke dataset from which we can nd most of the nuclei are selected from different carcinoma.

As there is no existing open source lipoma and ALT nucleus dataset, I attempted a HoverNet model pretrained with a well-built pan-cancer dataset (the Pan-

Nuke) at this stage. As a result, inference of all the 59 internal training slides using the PanNuke trained HoverNet is the starting point for our nucleus-based study.

4.3.2 The Semi-Automatic Lipoma and ALT Nucleus Annotating Work ow: An Overview

As will be presented in the result section (Section 4.4.1), the standard HoverNet-PanNuke failed on the lipoma and ALT slides in this study. We can still make use of the original classi cation and segmentation of the predicted in ammatory nuclei as well as the segmentation of the predicted neoplastic nuclei and connective nuclei. These results were taken as the initial labels. A label correction process together with a large ALT nuclei segmentation correction step can be applied on top of the initial segmentation and classi cation annotation. On this basis, a semi-automatic work ow was developed for the nucleus-level data labelling (i.e., ground truth generation) for the in-house lipoma and atypical lipomatous tumour slides.

The rst step of the work ow is to generate large tiles with the size of 1024 1024 pixel from the pathologist's regions-of-interest annotation from the whole slide images. And then the HoverNet-PanNuke was applied to all these large tiles for an initial nucleus-level classi cation and segmentation result. RGB-Segmentation-Classi cation 5-channel images are reconstructed using the source RGB images together with the .json outputs generated from HoverNet. Immediately thereafter, a label correction process is applied to the 5-channel images according to the ROI-level pathological annotations as well as the slide-level diagnostic labels. And if any of the ROIs are selected from an ALT slide, an additional label correction for the large nuclei will be performed, and this will be discussed in detail in Section 4.3.4. Lastly, a nuclei augmentation process is applied to the large ALT nuclei to balance the classes of the dataset, and this will be discussed in Section 4.3.5. Once all these steps are nished, the 1024 1024 large tiles will be further tiled up into the size of 256 256 which aligns with the training input of HoverNet. Table 4.3 shows the classes of the proposed novel nucleus-level dataset, and a HoverNet-based lipoma-ALT nuclei detector can be trained with this novel dataset using a transfer learning technique. The full work ow is summarised in

Figure 4.3: The proposed semi-automatic nuclei labelling work ow.

Fig 4.3.

Table 4.3: The classes of the proposed novel nucleus-level lipoma and atypical lipomatous tumour dataset.

Class ID	Class Name
1	In ammatory nuclei
2	Lipoma nuclei
3	Atypical lipomatous tumour nuclei (conventional)
4	Atypical lipomatous tumour nuclei (large)
5	Miscellaneous (i.e., other)

4.3.3 Nuclei Tile Selection

The nucleus-level lipoma and ALT dataset was developed using images of size 1024 1024, therefore, HoverNet-PanNuke was applied to 1024 1024 large tiles generated from the annotated ROIs from the source slides. 1024 1024 pixel region is a small window compared to most of the annotated ROIs, so some selection processes need to be taken place to ensure the high quality of 1024 1024 tiles. First, we need to make sure there are a reasonable amount of nuclei in each of the tiles as the nuclei density in lipoma and some of the ALT cases are low, and those tiles that contain few or no nuclei need to be removed. And then, tiles that contain too much histology variation need to be removed as well. Third, the 1024 1024 windows with out-of-focus areas or artefacts need to be removed. In conclusion, the 1024 1024 tiles were selected under the following criteria:

- the 1024 1024 tiles should contain clear nuclei,
- the nuclei density within each of the tiles should not be low,
- the tiles should not contain histology variation,
- the tiles should not have scanning quality issues such as out-of-focus nor containing artefact.

As the lymphocytes only appear in the histology variation of in ammation in the lipoma and ALT slides, the number of lymphocytes is relatively small compared

to the dominant classes of lipoma and ALT nuclei. I added more lymphocytes to balance the number of each class of nuclei by including a lymph node slide in the proposed dataset. Table 4.4 shows the tile-level statistics, and the statistics with the number of nuclei in each class will be introduced later in Table 4.6.

Table 4.4: The tile-level statistics for the proposed novel nucleus dataset.

Nuclei Type(s)	Number of Tiles	Slides IDs
In ammatory nuclei	32	A lymph node slide
Lipoma nuclei	3,205	1, 2, 3, 18, 20, 42, 43, 44, 51, 52, 53, 55, 57, 63, 64, 108, 109, 111, 113, 115, 117, 120 (in total 22 WSIs)
Small and large ALT nuclei	722	4, 5, 6, 8, 9, 10, 19, 25, 26, 28, 32, 34, 41, 69, 70, 71, 73, 75, 77, 78, 79, 83, 85, 86, 90, 92, 93, 96, 99, 100, 104, 105, 197, 199, 206 (in total 35 WSIs)

4.3.4 Label Correction

HoverNet-PanNuke was employed for the inference of all the selected 1024×1024 tiles to generate an initial nuclei segmentation and classi cation result. A combination of the ROI annotations and the slide-level diagnostic label ground truth are taken for the nucleus-level label correction, to be speci c, the relabel protocol is listed in Table 4.5.

Table 4.5: Nucleus class merge and relabel protocol.

Diagnostic Label of WSI	HoverNet-PanNuke Output	Relabelled Class
Lipoma	In ammatory	1. In ammatory
	Neoplastic Connective	2. Lipoma
	Dead Non-neoplastic epithelial	5. Miscellaneous (other)
ALT	In ammatory	1. In ammatory
	Neoplastic Connective	3. Small ALT
	n/a (additional process)	4. Large ALT
	Dead Non-neoplastic epithelial	5. Miscellaneous (other)

An additional process was executed for the label correction of the large ALT nuclei in the ALT tiles. A sub- ow is proposed for the large ALT nuclei label correction, and this is summarised in Fig 4.4.

Figure 4.4: The large ALT nuclei label correction sub- ow.

Fig 4.5a shows a 12532 8491 region sampled from case 206, where six smaller regions are randomly selected (Fig 4.5b), and Fig 4.5c to 4.5h shows the corrected large ALT nuclei using the work ow proposed in Fig 4.4 (Note: only the large ALT nuclei are shown in this gure as a highlight, and all the small nuclei are hidden). First, a colour deconvolution process (described in Section 1.6.3) is applied to the ROIs to extract the Hematoxylin-channel and the Eosin-channel, a watershed segmentation (described in Section 4.2.2) is then performed on the Hematoxylin-channel for a coarse nuclei segmentation. In the end, we set a threshold to remove the small segmented nuclei and only leave the large ones.

The 1024 1024 windows were further tiled up into smaller 256 256 tiles in this study so that HoverNet can be retrained using these tiles later. 256 256 tile was de ned as a valid tile if it contains nuclei. Table 4.6 displays the statistics after the label correction (including large ALT nuclei label correction) stage, from which we can nd the numbers of nuclei are balanced apart from the large ALT nuclei. A large ALT nuclei augmentation is carried out as the last step to solve this problem, and this will be discussed in the next section.

4.3.5 Large Aberrant Nuclei Augmentation Strategy

According to Table 4.6, the class of large ALT nuclei is imbalanced compared to all other types of nuclei, so a large ALT nuclei augmentation is performed. The large ALT nuclei come with the small ALT nuclei in each of the tiles, and we do

(a) A 12532 8491 region sampled from case 206.

(b) Six randomly selected regions from the sample.

(c) Large ALT nuclei label correction result (region 1).

(d) Large ALT nuclei label correction result (region 2).

(e) Large ALT nuclei label correction result (region 3).

(f) Large ALT nuclei label correction result (region 4).

(g) Large ALT nuclei label correction result (region 5).

(h) Large ALT nuclei label correction result (region 6).

Figure 4.5: The proposed large ALT nuclei label correction work ow: an example region from case 206, note: only the large ALT nuclei are shown in this gure as a highlight, and all the small nuclei are hidden.

Table 4.6: The nuclei statistics for the proposed novel dataset before augmentation.

Nuclei Type(s)	Number of 1024 1024 Tiles	Number of Valid 256 256 Tiles	Number of Nuclei
In ammatory nuclei	32	510	17,723
Lipoma nuclei	3,205	20,223	24,875
Small ALT nuclei	722	7,298	33,589
Large ALT nuclei		2,694	3,961
Total	3,959	30,752	80,148

not wish the small ALT nuclei to be augmented while we augment the large ALT nuclei. Therefore, another sub- ow is designed for this computational task. This sub- ow is summarised in Fig 4.6.

Figure 4.6: The large ALT nuclei augmentation sub- ow.

This sub- ow is described as: rst, all small nuclei are masked out, and only the large ALT nuclei are left in each of the ROIs, and then a cluster of augmentation strategies are applied onto the masked ROIs with a probability of triggering. In the preparation of this novel dataset, the following augmentation strategies were employed:

1. augmentation A: rotated 90 degrees counter-clockwise,
2. augmentation B: re ected upside down,
3. augmentation C: re ected diagonally from left-bottom to right-top (equivalent to augmentation A plus augmentation B),
4. augmentation D: rotated 180 degrees contour-clockwise,
5. augmentation E: rotated 270 degrees contour-clockwise,
6. augmentation F: re ected from the left-hand side to the right,
7. augmentation G: re ected diagonally from left-top to right-bottom (equivalent to augmentation E plus augmentation B).

The above augmentation strategies gave eight images (including the original one) for each of the tiles containing large ALT nuclei. We randomly selected about

40% proportion of all these augmented tiles to match the number of large ALT nuclei to other classes. This augmentation process enlarged the number of 256 tiles containing large ALT nuclei from 2,694 to 10,236, and this also resulted in a total of 20,607 large nuclei.

Data augmentation can be an extremely effective technique to improve generalisability by balancing the dataset. However, it is true that certain biases may be inadvertently introduced during this process, and here are some potential sources of bias:

1. Over-representation: certain transformations may over-represent speci c characteristics or patterns within the data, which can lead to a model that's overly tuned to those features. For instance, if we are doing excessive horizontal ipping, the model may overlearn certain symmetric properties that might not be prevalent in the actual population of pathology slides.

2. Unrealistic transformations: some forms of data augmentation may create unrealistic or rare instances that can mislead the model. For instance, excessive rotation of pathology images might generate orientations that would never occur in practice, leading the model to learn from non-existent or unlikely scenarios.

3. Ignoring domain-speci c features: pathological images have certain domain-speci c properties that should be maintained during augmentation. For example, structures within a cell should be in the correct orientation. Introducing augmentations that disregard these domain-speci c factors may lead to unrealistic data that could skew the model.

4.3.6 Colour Normalisation

As we are focusing on nucleus-level classi cation and segmentation, a colour normalisation method was applied before the execution of the above-proposed work- ow to reduce the intra-batch colour variation. The nal version of our proposed dataset was colour-normalised using Reinhard's method. And the full steps of developing the nal version of the dataset are:

1. apply Reinhard's colour normalisation method on the 1024 1024 tiles,

2. apply HoverNet-PanNuke to all the normalised tiles to generate their initial segmentation and classi cation,

3. apply our label correction and large ALT augmentation work ow to those initial segmentation and classi cation results,

4. tile up the 1024 1024 images into 256 256, integrate the segmentation instance mask as well as the classi cation type mask together into 256 256 5 images which align with the input format of HoverNet for training purpose.

The complete nal dataset will be described in the following result section.

4.4 Results and Discussion

4.4.1 Applying the Standard HoverNet-PanNuke on Sarcoma Slides

Fig 4.7 shows two zoomed-in areas of the lipoma sample 001 for the original HoverNet-PanNuke inference result. In both gures, red annotations stand for neoplastic predictions, green annotations stand for in ammatory predictions, blue annotations stand for connective predictions, yellow annotations stand for dead nuclei, and the very minor light brown annotations stand for epithelial predictions.

The HoverNet-PanNuke was then applied to all 59 internal training slides. And a WSI-level predicted nuclei count statistics (heatmap) is displayed in Fig 4.8. In this gure, samples from 206 to 4 are ALTs, and samples from 157 to 1 are lipomas. From this stacked bar chart, we can nd that HoverNet-PanNuke failed to make the correct prediction on our own lipoma and atypical lipomatous tumour dataset, as most of the nuclei in our own lipoma and atypical lipomatous tumour slides were predicted to be connective nuclei and were not as we expected to be the neoplastic or miscellaneous classes. A potential reason is that HoverNet-PanNuke is trained on various types of carcinoma originating from epithelial carcinoma, and the output classes of HoverNet-PanNuke are most epithelial-related.

(a)

(b)

Figure 4.7: Two regionalHoverNetinference result examples for lipoma case 001. In the gure, red annotations stand for neoplastic predictions, green annotations stand for in ammatory predictions, blue annotations stand for connective predictions, yellow annotations stand for dead nuclei, and the very minor light brown annotations stand for epithelial predictions.

Figure 4.8: A stacked bar chart for the slide-level numbers of predicted nuclei for all the 59 internal training slides using HoverNet-PanNuke. In this figure, samples from 206 to 4 are ALTs, and samples from 157 to 1 are lipomas.

We also conducted some validation of the Image Mode inference, and the results were showing that standard HoverNet-PanNuke does not work on our sarcoma samples again. When doing neoplastic nuclei prediction on the ALT slides, the large ALT nuclei (lipoblast) were badly predicted. These large ALT nuclei were usually predicted to be multiple small nuclei or only part of the large nuclei was detected. Fig 4.9 is an example of this failure on slide 206, only the predicted large nuclei are shown in this gure, and the small nuclei of the predicted connective nuclei and in ammatory nuclei are hidden. In this gure, we can nd that most of the large ALT nuclei in this area were missing under HoverNet-PanNuke

Figure 4.9: An example of the failure of large ALT nuclei prediction on an ALT slide 206. Only the predicted large nuclei are shown in this gure, and all the small nuclei of the predicted connective nuclei and in ammatory nuclei are hidden in this gure.

A conclusion for applying the standard HoverNet-PanNuke to our own lipoma and ALT slides are:

1. HoverNet failed to classify lipoma and ALT nuclei,
2. although HoverNet is able to segment the small lipoma and ALT nuclei to some degree, the segmentation of the large ALT nuclei failed,

A suggestion for distinguishing lipoma and ALT nuclei is to retrain the HoverNet with a bespoke nuclei-level lipoma and ALT nuclei dataset and to build a

lipoma-ALT nuclei detector. A novel nucleus-level lipoma and ALT dataset can be developed under a semi-automatic work ow as the rst step.

4.4.2 The Novel Lipoma and ALT Nucleus Dataset

(a) 256 256 sample tile 1.
(b) 256 256 sample tile 2.
(c) 256 256 sample tile 3.
(d) 256 256 sample tile 4.

(e) HoverNet result of 256 256 sample tile 1.
(f) HoverNet result of 256 256 sample tile 2.
(g) HoverNet result of 256 256 sample tile 3.
(h) HoverNet result of 256 256 sample tile 4.

(i) Large ALT correction result of 256 256 sample tile 1.
(j) Large ALT correction result of 256 256 sample tile 2.
(k) Large ALT correction result of 256 256 sample tile 3.
(l) Large ALT correction result of 256 256 sample tile 4.

Figure 4.10: Some examples of the ALT 256 256 tiles containing large aberrant nuclei. In this gure, blue annotations stand for the small ALT nuclei, red annotations stand for the aberrant lipoblast, and the green annotations stand for the in ammatory nuclei.

A semi-automatically generated nuclei instance dataset of lipoma and ALT is proposed for the computational tasks of segmentation and classi cation. The nuclei instance generation work ow includes a label correction protocol, a large ALT nuclei correction sub- ow and a large ALT nuclei augmentation sub- ow. This novel

Table 4.7: The nuclei statistics for the proposed novel dataset after augmentation (full version).

Nuclei Type(s)	Number of 1024 1024 Tiles	Number of Valid 256 256 Tiles	Number of Nuclei in the Lympha Node Slide	Number of Nuclei in Lipoma Slides	Number of Nuclei in ALT Slides	Number of Nuclei (Total)
In ammatory	32	510	14,772	1,231	1,720	17,723
Lipoma	3,205	20,223	0	24,875	0	24,875
Small ALT	722	7,298	0	0	33,589	33,589
Large ALT	3,401	10,236	0	0	24,568	24,568
Miscellaneous	n/a	n/a	0	1,366	534	1,900
Total	6,627	38,267	14,772	27,472	60,411	102,655

Table 4.8: A proposed nucleus-level train-validation statistics for the HoverNet training purpose.

Nucleus Type	Total Count	Train Count	Validation Count
In ammatory	17,723	12,721	5,002
Lipoma	24,875	17,480	7,395
Small ALT	33,589	23,214	10,375
Large ALT	24,568	17,284	7,284
Miscellaneous	1,900	1,321	579
Total	102,655	72,020	30,635

dataset consists of 6,627 visual elds (i.e., the 1024 1024 tiles) that were selected from 22 lipoma and 35 atypical lipomatous tumour whole slide images. Reinhard's colour normalisation method was applied before this semi-automatic work ow. In total, the dataset contains 102,655 labelled nuclei, each with an instance segmentation mask and a classi cation-type map. A tiled version consisting of 38,267 tiles of size 256 256 was created that aligns with the training input of HoverNet.

Table 4.7 is the full statistics of the nal version dataset. And as mentioned before, we have 2,679 more augmented 1024 1024 tiles for the large ALTs containing 20,607 more large ALT nuclei after the augmentation process to make each of the classes balanced. A train-validation split is also performed on the 256 256 image level with a split factor of 0.7, and Table 4.8 shows a proposed nucleus-level train-validation statistics for this novel dataset. This train-test split dataset will be used in the next Chapter to retrain the HoverNet.

A subset of this dataset was validated manually, and the semi-automatic generated annotations for each type of nuclei are generally good though there exist a few over- or under-segmentations. Some examples of the small 256 256 tiles (the size that is accepted by HoverNet training) containing large ALT nuclei are shown in Fig 4.10.

4.5 Conclusions

Analysis of nuclei in pathological slides is an important step towards automated differentiation of the diagnostic labels, and the nuclear features have been often used to

distinguish the lipomas and the ALTs in pathological routine work. However, visual examination of nuclei is time-consuming as there are often hundreds of thousands of nuclei within each of the whole slide images. Nucleus segmentation and classi- cation help subsequent analysis of the role that nucleus features play in diagnosis decision making, and existing work ows such as HoverNet generate reasonable re- sults for the nuclei segmentation but fail on the nuclei classi cation on lipoma and ALT cases. A nucleus-level dataset is the pre-condition for training a good nucleus detector for the segmentation and classi cation of lipoma and ALT.

In this study, I have proposed a semi-auto work ow, including a label cor- rection protocol, a large ALT nuclei correction sub- ow and a large ALT nuclei augmentation sub- ow, for the generation of a novel nucleus-level dataset of lipoma and atypical lipomatous tumour. Though there are lots of open nucleus-level dataset (described in Section 4.2.1) for other types of cancer, this proposed dataset will be the rst lipoma and liposarcoma dataset to date.

Colour normalisation was applied before the execution of the semi-automatic nucleus annotation work ow. And the nal dataset contains 6,627 images of 1024 1024, with a total of 102,665 labelled nuclei, including 17,723 in ammatory nuclei, 24,875 lipoma nuclei, 33,589 small ALT nuclei, 24,568 large ALT nuclei and 1,900 miscellaneous.

One limitation of this work is only the conventional lipoma and ALT nuclei were labelled. More variants of fatty tumours, such as spindle cell lipoma or ALT with histology variation, may be included as well to make the dataset more solid. From a larger perspective, nuclei from more types of sarcoma such as synovial sarcoma (SS), solitary brous tumour (SFT) and undifferentiated pleomorphic sar- coma (UPS) could be also included in one dataset for a wider range of sarcoma nuclei segmentation and classi cation tasks. The dataset can only be expanded if we have suf cient slides of these variants.

Chapter 5

Slide-Level Classi cation between Lipoma and ALT Using Handcrafted Nucleus Features

5.1 Chapter Overview

Before the utilisation of machine learning techniques, computational image analysis based on handcrafted features (Street et al. 1993) was used in many biological and medical studies. Some studies in H&E histopathological images suggested that handcrafted nucleus features can be employed for many computational tasks, including tumour grading, diagnostic label classi cation, prognostic prediction and predicting whole genome doubling, etc. In this chapter, I explore the capability of handcrafted nucleus features on distinguishing lipomas and atypical lipomatous tumours at the whole slide level.

5.2 Related Work

Axelrod et al. (2008) indicated the nucleus grade based on handcrafted features is able to contribute to the classi cation and prognosis of ductal carcinoma in situ (DCIS) of the breast. In this study, 39 handcrafted features were extracted and calculated for nucleus grade before statistical analysis. These features were categorised into three groups: morphological features (such as area and perimeter),